In Defense of the Systems Reply

by

Sascha Lecours

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Arts

in

Philosophy

Waterloo, Ontario, Canada, 2010

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.
I understand that my thesis may be made electronically available to the public

Abstract

In John Searle's *Minds, Brains, and Programs*, he argues against the possibility of a digital computer capable of understanding. In particular, Searle puts forward the Chinese room thought experiment, which appears to neatly dismiss the concept of a machine made to manipulate symbols thereby obtaining a mind. This thesis replies to *Minds, Brains, and Programs* by critically examining the Chinese room experiment and the conclusions Searle draws from it. This paper defends what Searle fterms the "systems reply" to the Chinese room, which proposes that in the Chinese room thought experiment, the total system of operator, instructions, and input-output may achieve understanding even if the operator alone does not. The defence of the systems reply requires rebutting Searle's direct objections to it as well as a more general account of the possibility of obtaining semantic content from what appear to be a purely syntactic manipulations.

Acknowledgements

I would like to begin by thanking Steven Weinstein for his support and guidance during the writing of this thesis. I also thank my readers Shannon Dea and Chris Eliasmith for their timely, thorough, and helpful comments.

I could not have written this without the debates and discussions provided by my fellow graduate students at the university of Waterloo, and in particular I would like to recognize Bradley Shubert and Paul Smith for their contributions to my argument.

Finally, I wish to thank my parents for their encouragement and support, and Erin Reznick for everything she has done to help me see this thesis to its completion.

Table of Contents

Introduction:

In *Minds, Brains, and Programs,* John Searle sets out the Chinese Room argument, in which he argues against the notion that a given computer running a program can be said to possess a mind, or to engage in the act of understanding, and do so purely in virtue of the program being run. This thesis aims to demonstrate that Searle's arguments against computer understanding are flawed, and that a computer instantiating the correct program could indeed be said to understand. I will achieve this by defending the Chinese Room counterargument known as the systems reply against the points Searle raises against it. I furthermore argue that the correct standard for detecting a digital computer's ability to "understand" is a modified version of the Turing Test.

The structure of this thesis is first to present the Chinese room thought experiment as well as the systems reply against it. Second, I will deal with Searle's direct reply to the systems reply, which is the memorization counter-argument. Third, I will address Searle's indirect arguments against the systems reply, such as the impossibility of deriving semantic meaning from syntactic manipulation. Having thus repelled the direct attacks against the possibility of the systems reply being correct, I will offer positive arguments that the ability to pass a properly-conceived behavioural test is the best possible measure of understanding and the possession of a mind. I intend to make the case for a purely behaviourist test of understanding along the lines proposed by Turing, responding to Searle's objections to this approach as appropriate. I will also respond briefly to the

objections of other thinkers such as Ned Block and Stevan Harnad against the suitability of the particular Turing test version that I advocate in this thesis.

Finally I will acknowledge those points of Searle's paper which I believe to be correct and think are worth noting even as one accepts my conclusions, in particular with respect to the limitations of digital computers and the proper standards that should be set with regard to what is considered "information processing" in a meaningful sense.

1 Overview of the Chinese Room

Searle's "Minds, Brains, and Programs" was intended originally as a response to a school of thought that he labels "Strong Artificial Intelligence" (Strong AI). According to this doctrine, "a computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states." (Searle, 1980, p 417).

The second facet of Strong AI is that a computer whose program enables it to understand serves as an explanatory tool for human cognition as well: "In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations."

In "Minds, Brains and Programs," Searle targets some conclusions made regarding the story-understanding program produced by Roger Schank and his colleagues at Yale (Schank & Abelson 1977). Searle chose this particular case as an example of a program that would be held up as being capable of understanding in order to illustrate that a program might produce adequate responses but still not understand in a meaningful sense.

The program in question was written with the purpose of imitating the human ability to understand a story. This ability was represented by giving the program the ability to first accept a given story as input, and then, when asked questions about that

story, respond correctly even about information not explicitly given in the text. For example, says Searle, suppose one were to provide the program the following story: "A man went into a restaurant and ordered a hamburger. When the hamburger arrived, it was burned to a crisp, and the man stormed out of the restaurant without paying for the burger or leaving a tip." Now if we ask the program "Did the man eat the hamburger?" it will presumably answer "No, he did not." In this way, it duplicates the human ability to glean information from a story beyond its actual text. (Searle, p. 417)

Suppose that the program is successful enough to reliably produce plausible answers to such questions just as a human might. The proponent of Strong AI, says Searle, will then make two claims about it:

1. That the machine can be literally said to *understand* the story and provide the answers to questions
2. That what the machine and its program do *explains* the human ability to understand the story and answer questions about it.

Searle says of these: "Both claims seem to be to be totally unsupported by Schank's work, as I will attempt to show in what follows." [1]

Searle then produces the Chinese Room, a thought experiment whose intention is to show that merely producing the same outputs as a human might under certain circumstances is not sufficient to demonstrate *understanding* of the topic at hand – or the possession of a mind more generally.

[1] Searle notes that Schank himself does not advance either of these claims of his own work.

1.1 The Original Chinese Room

The Chinese room thought experiment goes as follows: suppose that someone were to write a book of instructions, and that this book contains English instructions . The instructions read "When you see the following symbol on a slip of paper, then respond with the following two symbols, unless it is the case that *this* symbol precedes it, in which case respond with these symbols. If instead the symbol is immediately followed by *this* one, then respond with this single symbol," and so on and so forth. These instructions are in fact a set of rules which, if followed, allow the operator to receive as inputs slips of paper containing Mandarin Chinese characters and produce as outputs slips of paper on which are written the appropriate responses, also in Mandarin Chinese. The instructions are written such that it is not necessary for the operator to be able to read Chinese. In fact, the operator need not even recognize the symbols as Chinese characters, even as language. The instructions are thorough enough that the operator need only be able to visually distinguish one symbol from the other, without any deeper recognition.

Thus, equipped with these instructions, which we may imagine take the form of a book or filing cabinet, an operator could pass for a native Chinese speaker under the right circumstances. Specifically, suppose that we were to lock Searle, who neither speaks nor reads the slightest bit of Chinese, in a room with nothing but the instructions, a pen, and a few slips of paper. Through a slot in the door, a Chinese speaker slips pieces of paper on which questions are written in Chinese. Searle, receiving these, performs the appropriate

series of steps from his instruction manual, and produces a slip of paper with a string of Chinese characters on it which spell out an acceptable response to the question of the person outside. Recall that Searle has no fluency whatsoever in Chinese, and has no idea of the nature of what he is doing. When Searle returns this slip under the door, the Chinese speaker outside reads what Searle has written and may conclude that inside the room is a man who understands how to read and write Mandarin Chinese.

However, says Searle, it is obvious that the man inside does not understand Chinese in the least. The operator inside understands only English[2], and is no closer at all to understanding Chinese than before stepping into the room.

This process, says Searle, is just the sort of thing that happens when we input a story and questions to Schank's program: a series of characters arrives from outside, each character in turn is interacted with according to the complex rules laid out by the program, and a second string of characters is produced and sent back outside. Throughout this process, the computer has not the foggiest understanding of the meaning of its inputs or outputs. It merely performs its operations.

How can we be sure that the computer does not understand? The role of the computer in the story program is the same as Searle's role in the Chinese Room: receive characters, follow rules, and produce characters. If Searle does not understand the inputs and outputs in the Chinese room case, then the computer cannot understand in the case of the story, because the computer has "nothing more than I have in the case where I understand nothing." (Searle, p. 418)

[2] … or whatever language the instructions may be printed in.

Thus, the first claim of the Strong AI proponent is refuted. The computer does not understand the stories given it, because Searle does not understand the Chinese questions given to him, and the computer has no more ability to understand in a given situation than does Searle. As regards the second claim, that the program explains human understanding, Searle holds it as obvious that the program is insufficient by way of explanation since the interaction of the program and computer does not produce any understanding and therefore naturally ought not to be supposed to reflect human understanding.

The weight of this argument is not only that the same outputs can be produced in different ways, but also that no program will ever be sufficient to produce understanding. This is because no matter what instructions are provided to a machine, so long as they are defined in purely formal terms, as a program, they will be insufficient to produce understanding on the grounds that a human could also follow those instructions without thereby understanding anything.

As an illustration of this last point, take the case of existing programs intended to replicate the textual output of a human, such as ELIZA (Weizenbaum, 1965). ELIZA takes a given line of input from a human subject, such as “Hi Eliza, how are you?” and then processes this sentence (string of characters) according to the instructions in its program, and then produces a string of characters as output, such as “I am well. How are you feeling today?” According to Searle’s position, no matter how convincing ELIZA’s responses may prove to be, we should not assume that any understanding is taking place on her side of the conversation so long as the program’s responses are dictated by a set of formal instructions. We need only remind ourselves, says Searle, that any human could also be

producing Eliza's outputs by simply following those same instructions without thereby understanding either the inputs or outputs.

Suppose, for instance, that the Chinese analogue to John Searle, a Mandarin Chinese speaker who neither reads nor writes English, is given the Chinese instructions that are equivalent to the ELIZA program. Even if one were to have an extended conversation by slipping sheets of paper with English statements under the door to the locked room containing the Chinese person, and receiving appropriate responses on sheets slipped back under the door, it is evident that the Chinese person in the room does not thereby understand English in the slightest. Since the Chinese person did not understand English, and a digital computer executing the ELIZA program adds no capacity to understand that was not present with the human operator, then ELIZA does not understand. And so it goes for *any* formally defined program, claims Searle; none of them can understand, because any conceivable interaction they may have with humans can be described in a format analogous to the Chinese room.

1.2 The Systems Reply

I believe that the most logical place to make a stand against Searle's thesis in the Chinese room is to defend perhaps its most straightforward and popular objection, the systems reply.

The claim of the systems reply is that the ability to produce the correct responses shows that some system exists which does, in fact, understand. For instance, when Searle is in the Chinese room with his symbol-manipulating instructions, although we grant that

Searle himself does not understand Chinese, nevertheless the combination of Searle, the instructions and the inputs and outputs *do* understand Chinese. The argument is that we should not *expect* Searle to understand in the Chinese room case, because the role he is fulfilling in this computational system is merely that of the processor, and not the entire computer, which would also contain memory and other components. These components do not individually understand, but when considered as a total system, then they *do* understand. A favourite analogy of the systems reply advocate is to compare the Chinese room to a human brain: no single component of the human brain can be said to understand anything when taken in isolation. Nevertheless, the *system* of the brain as a whole clearly *does* understand. The operator in the Chinese room experiment is analogous to a part of the brain, while the entire room is analogous to the entire brain. It is therefore unsurprising that the operator does not come to understand Chinese. Rather, the total system of the room possesses whatever mental capacity is necessary to produce the correct responses. The combination of operator, program, and other components *do* understand.

1.3 Definition of Terms

Before we review Searle's attack on the systems reply, let us define the terms. This paper is chiefly concerned with the question of whether or not an appropriately programmed computer can possess *understanding*. Searle defines "understanding" as implying "the possession of mental (intentional) states and the truth (validity, success) of these states." He adds that for the purposes of his argument he is concerned only with the

possession of these states - as opposed, presumably, to the possession of such states as "consciousness" and the like (Searle, 1980.) He further clarifies intentionality as being that property of mental states whereby they are directed at or about objects and situations in the world. This means that beliefs and desires are intentional states while general depression or anxiety are not.[3] What does an opponent of Searle's position need to prove, then?

Certainly there is one important detail about what does *not* need to be proven by an advocate of the systems reply: consciousness. One can subscribe to the systems reply without needing to demonstrate that an appropriately programmed computer is *sentient, conscious, or self-aware.* These three terms are nebulous and, often enough, difficult enough to define as to be themselves the subject of a paper. Fortunately, it is not necessary to involve them in the defence of the systems reply. All we need to attribute to the correctly designed program are *mental states* of a certain type. Therefore I shall endeavour to show that a properly programmed computer is capable of understanding, but not that it is conscious.[4]

1.4 Against the Systems Reply

Searle attacks the systems reply in two ways: First, his direct attack on the systems reply, and second, his argument that no amount of syntactic complexity will ever produce semantic meaning - and therefore, he claims, no amount of mere programming will produce intentionality. This second argument, if accepted, also defeats the systems reply,

[3] (From endnotes 2 & 3 of Minds, Brains, Programs)
[4] Though I have no objection to the possibility of a conscious computer.

by entailing that any system based merely on the syntactic manipulation of symbols cannot attain intentionality. I will address each of these attacks separately. My response to the memorization reply is detailed immediately below. My response to the question of syntax and semantics is addressed in depth in chapter 2.

1.4.1 The Memorization Reply

Searle's direct response to the systems reply is the memorization counterexample. Regarding suggestion that the system of the instructions plus the person instantiating them constitutes a mind that understands Chinese, he answers: Suppose that our operator took his book of Chinese character-manipulation instructions and simply committed all of it to memory. The operator could then throw out the book (and leave the room), and simply be handed slips of paper with Chinese sentences and hand back slips of paper with Chinese responses – all the while having no idea that he is in fact receiving questions or giving answers, or even that it is occurring in the Chinese language. In this example, where the operator has memorized all of the instructions, it seems obvious that she does not understand any more than when she were following from a book, despite the fact that their mind now contains the entire "system". Therefore the claim made by the systems reply advocate, that the system as a whole must understand, cannot be so. Is this really the case? Are there any significant problems with the analogy of the person having memorized the complete instructions?

There are in fact several problems with Searle's response. First, Searle's reply is vulnerable to the suggestion that Searle does in fact understand Chinese both while in the

room following instructions and also after memorizing them. All that he lacks in this case is the *awareness* of understanding Chinese. Typically, any time that a human understands how to do something, we are also aware of our understanding – but it does not seem to be a *necessary* property of understanding even in the sense of the word put forward by Searle. Searle states that in order to understand, one needs to possess a *mind*, but he does not stipulate that we need to have any particular awareness of the understanding taking place (Searle, 1980, p. 424, Notes). Thus one can still hold that Searle understands Chinese both in the normal case and after memorizing the instructions, and simply happens to be unaware of it in both cases.

However, the reply that Searle does understand Chinese but is not aware of it is a problematic one, because it fails to respond to several of Searle's illustrations of the Chinese room's inability to convey understanding to the operator. For example, when we ask Searle a question in English and the same one in Chinese, even if he were to give the same response, the process occurring in his mind is very different in either case. When we ask Searle a question in English, he engages in thought, reflection, summons memories, weighs emotions, and produces a thoughtful reply. When we ask the question in Chinese characters, he merely performs a few feats of memory and careful mental operation and produces what are, to him, meaningless symbols. Even if one contends that in the latter case he has "understood" the Chinese questions asked of him without being aware of his doing so, it is still evident that there are important differences in the process taking place in Searle's mind besides merely his awareness of them. This is a line of argument examined in more detail by Hauser[5] and others. The response that I advocate here is

[5] Hauser, L., 2002, *'Nixin' Goes to China'* in Preston and Bishop (eds.)

slightly different: I believe that Searle's memorization reply already has several major flaws as a valid counterexample to the systems reply.

Recall that in order to demonstrate that, in the case of the Chinese room, no understanding on the part of any system external to the operator's mind is taking place, he proposes a revision. In his revised example, the operator of the Chinese room program is not following written instructions from a book, but has instead committed them to memory. Since every part of the equation is contained within the mind of the operator, says Searle, and the operator still does not understand, then the systems reply is defeated as no system exists that can be said to understand. If one did exist, it would surely exist within the mind of the operator – and therefore the operator would have to also understand. Perhaps initially convincing, this argument leads to some difficult problems when subjected to careful scrutiny. It relies on certain intuitions on the part of the reader – misleading ones that conceal an important disanalogy, as I will show.

First, it is possible to simply reject the assumption implicit in the memorization reply that it is self-evident that the memorizer's mind is the only one to consider. After all, the systems reply advocate proposes that a person running the Chinese room program creates a mind of sorts capable of understanding Chinese as a consequence of running it. Why would they abandon this attitude the moment the operator memorizes the instructions?

In order for the systems reply to be coherent, it must be claiming that when a computational operation of sufficient sophistication and power (for example, enough to produce convincing, native-speaker quality answers to questions in a human language)

takes place, then there is at least a virtual[6] mind in existence that is carrying out the action we call understanding. The underlying assumption of the systems reply in this case is that it is impossible to produce the outputs of a human speaker without thereby duplicating in some form the essential structure of human thinking required to understand. Furthermore, this essential structure that is duplicated cannot be reproduced without producing *genuine understanding*. How can this be reconciled with the case of a monolingual English speaker who has memorized the instructions of the Chinese room?

In order to answer this question, I must first present the following example of a person memorizing a program to illustrate how a virtual mind may exist inside of a human brain. Imagine that, in addition to being able to perform simple arithmetic, I take it upon myself to utterly memorize the circuit layout and mechanical functions of a pocket calculator. I memorize the structure of every wire, every capacitor and diode, power source, conductivity, and all computer chips, and am able to imagine the entire system clearly in my mind. Now, if someone says to me: "Suppose I pushed the following sequence of buttons on a pocket calculator: 'nine, zero, division, three, zero, equality symbol.' – what would be the final display?" I can imagine the sequence of events: pushing the nine button completes a circuit which sends a current through a certain wire, which has a certain effect on the components of the processing chip, and then so on for all the keys. In the end, purely by imagining the layout and capabilities of the machine, I can produce the answer "The screen will display the configuration of opaque areas corresponding to the number

[6] By "virtual mind", I mean a mind that is neither a human mind nor an animal one. This category could contain minds such as those produced by digital computers, but naturally also includes minds such as the one put forward by the systems reply advocate in the case of the Chinese room. That is to say, a mostly artificial mind which may happen to include a human mind as a sub-component.

'three'." And, indeed, such is the depth of my memorization that my ability to simulate the calculator's operations in my mind is never wrong. I always achieve the answer that the actual calculator would give. If I could achieve such a perfect simulation of the calculator, would it not be apt to say that, indeed, there exists a calculator of that specification in my mind – or at least, my brain? Let us remember that if we believe that mental states correlate to brain states, then it must be the case that in some sense, my brain now contains the effective structure of this calculator. My neurons and axons and so forth have, by dint of my memorization, formed themselves into some structure that is functionally equivalent to the actual calculator. It is simply the case that in order to input commands into this "calculator" which exists in my brain, it is necessary to first relay the commands to my conscious mind (by telling me which button presses I am to simulate), which then sends them into the calculator brain-space and receives the correct response. If a person could genuinely perform this feat without ever producing an error, then it seems clear that their brain really does effectively contain a calculator of that type.

How does this example relate to the case where Searle commits to memory the structure of the Chinese Room program? What happens when Searle tells the systems reply advocate to simply memorize the entire system? Recall the scale of the program in question. This program has the net effect of taking any written Chinese input and producing an appropriate response that a native speaker might give. This presumably includes questions about the past, about opinions, taste in art, mathematical problems, hobbies, employment, etc. In short, it must contain in principle at least as much information as an adult human can possibly convey verbally – and in addition to this information, it must include a framework to at least minimally simulate such things as

emotional states, to remember previous parts of the conversation, and so forth, as well as including the Anglophone instructions. In short, it is a very large program. In fact, this program would need to contain the same order of complexity as is contained in a human mind, if we assume that humans are capable of verbally communicating information on the same order as they are capable of experiencing it in general.

This is a program of staggering proportions, and would be perhaps the most ambitious computer program ever written if it existed today. Certainly it would make the memorization of the hardware of the pocket calculator from before seem utterly trivial by comparison, even though memorizing even that "simple" device in the way I described would probably be beyond human ability. Thus, when I memorize the instructions from the Chinese Room case, I am memorizing the formal structure of an entire personality, including a lifetime of memories, likes and dislikes, a certain quickness of temper (or lack thereof), a field of knowledge, a sense of humour, goals, desires, etc.[7] If a person could actually commit all of these things to memory, then, just as in the case of the person who memorizes the mechanical workings of a calculator, my brain would contain a structure that is functionally equivalent to the brain of another human, or at least a significant portion of one. If I could, indeed, always produce convincing, consistent answers to the written questions I received, then it seems it would be fair to say that there *is* a mind that understands Chinese – and that it is *not* my mind – but rather the virtual mind that exists in my brain, in precisely the same way that the calculator in the brain of our first case can be said to really exist. Margaret Boden (1988) makes a similar argument: that confusing a person with a brain is a category mistake on Searle's part, and that one's brain may

[7] Or at least, enough of these things that my verbal interactions will reliably pass for those of a human speaker.

understand without conscious understanding on the part of the person in question. I believe that the operator who has memorized the program is not only unconscious of the understanding, but that it does not exist in their mind, proper. Unlike Boden, I believe that this counterargument has deeper implications for Searle's reply than merely the possibility that a the operator could understand without being aware of it.

Insightful as it is, Boden's conclusion is hard to buy. It goes very much against human intuition to accept that two minds could share a brain, and more so that one could be nested in the other, as the Chinese-understanding mind would be in the English understanding mind (through which, after all, all inputs and outputs must pass). There seem to be cases in modern neuropsychology that provide evidence against this intuition: cases such as patients whose brain hemispheres have been separated by the severing of the corpus callosum[8], though this view remains controversial in neuroscience (see for instance: Sergent, 1987).

I, however, am perfectly happy to grant that it is implausible that two minds could share a human brain. The notion that two minds frequently or ever exist in the same brain is not necessary to my reply. If we find it unbelievable that two entire human personae could exist in the same brain, then this is simply a testament to the fact that no human could ever conceivably commit to memory a system as complex as the one in the Chinese Room case. Recall how implausible it is that a human could even memorize the circuit and operations of a simple pocket calculator, when the operations of a human brain must be many orders of magnitude more complex.

[8] See for instance *How to Count People* (Bajakian, 2010

Searle acknowledges that human minds are produced by the structure and nature of the brain. It therefore follows that a given brain structure is sufficient for a mind to exist. It doesn't seem that simply adding another identical brain structure in the same skull alongside the first is really problematic. Even Searle would have to agree that if some structure *x* is a human mind, then a single brain that happens to contain *2x* is not impossible in principle. However, in practice, it runs up against our intuitions. We are inclined to think of two complete minds sharing one brain as an impossibility, because we never (or rarely) encounter this sort of phenomenon in daily life, and we tend not to think of a single brain as the right sort of thing to contain more than one mind.

My true reply to Searle's memorization counterexample is not, in fact, that we should grant that Searle's memorizer has created a virtual mind when he commits the rulebook to memory. Rather, we should refuse to allow that a normal person *could* memorize it. Anyone with enough mental capacity to perfectly memorize and execute this mammoth program is sufficiently different from a normal human that it would then become reasonable to suppose they have two minds sharing a brain. What we understand to be a human brain has limited capacity for memorization. Searle's program exceeds this capacity by a great deal, and in a significant way: it contains the means to effectively emulate a human's responses. Therefore, the notion that the systems reply is defeated when our man or woman commits to memory the instructions ought to raise a red flag, for strange things are possible when a human mind has this sort of capacity.

The resulting human would be very different in practice from what we think of as a human, a being who could memorize an entire human lifetime's worth of information,

behaviour and all[9]. It seems that we must either grant that a virtual mind would exist that could in principle "understand" Chinese, or that describing a human as memorizing this information is an attempt to sneak in a contentious assumption. This latter option is my response. Such a capacity so far exceeds what is meant by the word "human" that it leads to a disanalogy. In other words, Searle's counterargument runs into a fork; either he grants that the "human" who memorizes the instructions has the brain capacity to completely duplicate the program, in which case their brain is the sort that can plausibly host two minds at once: or the operator cannot commit it to memory, in which case the systems reply advocate need not explain why the operator does not understand Chinese after memorizing. In either case, there is always a mind other than the operator's that can be said to be understanding Chinese.

Thus, the systems reply is not defeated by the counterexample in which the operator memorizes the entire rulebook, because there is always a place where a mind capable of understanding can exist.

[9] Or, if not a direct memorization of the behaviour and information, then the memorization of the mental structures necessary to *generate* the behaviour and information that we associate with a human mind.

2 Syntax and Semantics

Searle's first attack on the systems reply is his memorization counterexample. Having dealt with that in the previous chapter, there now remains his second, indirect attack on the systems reply: the impossibility of any syntactic program producing semantic meaning.

2.1 Blood From A Stone: Is it Possible to Draw Semantic Meaning Syntactic Interactions?

Even if one accepts the tricky notion of a virtual mind required by the systems reply advocate, there remains a further problem. One of Searle's central arguments against Strong AI is that any program is necessarily composed of a control unit reading a symbol or input, moving on to another symbol determined by the previous one, and producing outputs of a given type. In the Chinese room, the operator has no idea which Chinese characters mean what – or even that they are linguistic characters at all. All that the operator knows is that arbitrary symbol #1 is followed by arbitrary symbol #2, and so on. There exists, therefore, a *syntax* of how to manipulate the symbols, but no *semantic* content except that given to the inputs and outputs by the programmer and the people outside the room.

Searle puts it thus:

> "[…] 'Could something think, understand, and so on *solely* in virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?'
> This I think is the right question to ask, […] and the answer is no.'
> 'Why not?'

> 'Because the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even *symbol* manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them.' " (Searle, 1980, p. 422, original emphasis)

Searle argues that no matter what properties the syntax of a given language are, it can never *thereby* acquire any semantic content – and therefore because its most basic elements have only syntactic and not semantic content, that no computer program can have semantic content.[10]

However, Searle admits that humans are capable of understanding, and therefore that their thoughts or words can possess semantic content as well as syntactic. The human brain, however, is composed of parts that have no semantic significance. In the dialogue which serves as a conclusion to "Minds, Brains and Programs" Searle readily admits that a human brain can be accurately described as a machine instantiating one or more programs:

> "Could a machine think?"
>
> The answer is, obviously, yes. We are precisely such machines.
>
> […]"OK, but could a digital computer think?"
>
> If by "digital computer" we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiations of any number of computer programs, and we can think. (Searle, 1980, p. 424).

The smallest meaningful units of the human brain – nerve axons, neurons, gap junctions – are merely physical objects with certain rules of interaction. From these

[10] Although naturally the output of any program can be interpreted by a human and thereby (trivially) acquire semantic content. This happens, for instance, when a human reads the series of characters output by Schank's story-interpreting program.

patently meaningless components, however, the human brain obviously manages to derive the ability to assign semantic meaning to statements and have a mind and intentional states. It therefore follows that merely being made up of elements that have no inherent meaning does not preclude the system as a whole from having meaning, since we know that humans do this.

A possible reply to this objection is that I am misusing the terms "syntax" and "semantics" which can only properly be applied to purely linguistic systems. I grant this objection, with the caveat that if the meaning I assign above is incorrect then Searle's must also be incorrect in the same way. In the Chinese room, the room and operator form a system in which a control unit – for instance, the operator - is made such that it *necessarily* effects the proper changes or operations when given a set of inputs. Insofar as the control is simply an entity that necessarily causes certain events according to certain rules from certain initial conditions, it is analogous to the case of the physical components in a human brain effecting certain operations according to certain rules. Thus, if the physical structure of a human brain cannot be described as following a syntax, then neither can the operator in the Chinese room. Both are systems such that certain rules are followed given certain inputs. The rules to be followed are equally devoid of semantic content in both cases: physical laws in the case of the brain, symbol manipulation instructions in the case of the Chinese room.

The reply has been made to my argument that the terms "syntax" and "semantics" clearly apply in the case of the Chinese room but not in the case of a brain, for the simple reason that the Chinese room consists of actual linguistic symbols being manipulated while a brain does not. To this I reply that Searle is not limiting his application of his

argument to mere linguistic programs; in the case of the ambulatory robot (Searle, 1980. p. 420), Searle argues that information input to the robot through its visual receptors could be expressed as symbols to be modified by the interpreter, and the output used as control signals for arms and legs. In these cases, the visual-sense input and motor-control output, the inputs and outputs are clearly only "linguistic" in the most abstract sense: they can *in principle* be represented as symbols on a piece of paper. Searle clearly thinks that the syntax/semantics rebuttal applies even when the program in question is not applied directly to a linguistic problem. If we consider even visual information and movement to be things reducible to characters to be syntactically manipulated, then it seems like humans must, in fact, be subject to the same criticism as Searle's robot in that we receive input from our senses and produce output through our behaviour through some physical operation of our internal components - where, then, does semantic content come in?

In other words, we need not think of the operation of the control unit of a computer as purely a symbol manipulator – it is equally correct to think of it simply as a physical object following natural laws, just as is the human brain. Since the human mind consists of matter in a specific configuration, it follows that it is in principle possible for a computational device made of matter to be a mind, provided the correct program.

As an anonymous hacker once put it: "What people forget is that there is no software. There are only different hardware states." Any computer program has existence not only as lines of code in a compiler, but as specific states and patterns of a physical object (a computer). Because of this, it is perfectly legitimate to expect a program's syntax to be able to produce semantics in the same way that the human brain does, because in either case, it is a matter of the right physical state obtaining.

Searle believes that we should accept this intuitive assumption: no combination of semantically meaningless objects or symbol manipulations can produce meaning by virtue of their being sufficiently complex or organized. However, we have uncovered three points that run counter to this assumption:

1. Humans are capable of having thoughts with semantic content.
2. Human thoughts are caused entirely by the human brain and its inputs.
3. The human brain is composed entirely of components that, taken individually, have no semantic content at all.

Our earlier intuition is incompatible with these three points. Therefore either our intuition is wrong, or some combination of these three assumptions is wrong. I will now defend the view that these points are correct, and that it is our assumption that semantic meaning cannot arise from syntax that is mistaken.

I find it much more plausible that our intuition is incorrect than to overthrow any of the three premises. I freely admit that it is a considerable mystery how a mind, intentions, and understanding could somehow be produced when simple physical objects (axons, neurons, etc.) interact in the appropriate way. The mysterious nature of the human mind and brain is a well-known puzzle for philosophers and scientists, and is beyond the scope of this paper. What is quite clear, however, is that humans *do* have the ability to create semantic content, and also that the smallest meaningful parts of the human brain (whatever those might be) do *not* contain semantic content. Searle seems to have no objection at all to premises 1 and 2. He might even agree to premise 3, but then add that the human brain is capable of semantic content *not* in virtue of its components

being sufficiently complex or organized, but in virtue of their *being made of the right sort of material*. With regard to Searle's advocacy of the biological brain as the seat of true understanding, I offer my reply in sections 2.2 and 2.2.1.

Searle might object to my third premise – that the human brain is made of components which, taken individually, do not possess semantic meaning. The first half of this premise - that the smallest relevant parts of the brain do not possess semantic content - is, I think, uncontroversial. However one subdivides the brain into component parts and processes, these will always be physical objects that can, in principle, be understood mechanistically. What might raise objections is the notion that we can describe the interactions of brain cells as "syntax". In the introduction to chapter 2, I outlined why we ought to think of the physical structure of the brain as equivalent to an operator executing a program. I will now defend that assumption in greater detail in order to support my contention that Searle is incorrect in claiming that syntax cannot give rise to semantics.

2.1.2 Brains and Symbols

Let me begin my discussion of the brain by acknowledging the immense complexity of the human brain, and also the nascent state of our brain science. When I speak of looking at the "smallest relevant parts" of the human brain, it is controversial what these parts might be. Does human thought originate from the frequency and amplitude of waves of electrical impulses in the brain? From chemical balance in different parts of the brain? From the total system of firings of neurons? All of the above? Perhaps none of the above? Which features, if any, of the waves or chemical balances are important to thought, and

which are merely accidental by-products that could be abstracted away if one wanted to build a human thought simulator?

These questions are sufficiently contentious (and beyond the scope of this paper) that no firm answer can be given here. Nor does one need to be. All that is required for my argument is the assumption that there is some physical threshold below which parts of the brain do not possess understanding. Some components of the brain are too small or specialized to be said to have or be a mind – for instance, a single neuron or axon or other small brain component. Whatever units one chooses as the smallest, they will be physical objects or states. This, I think, will be granted by anyone who believes that human thought is directly caused by, or identical to, physical states of the brain.

Whether the smallest "important" parts of the brain are axons and gap junctions, or some yet-unnamed subdivision of these, is unimportant. It is clear that these parts exist; that these parts taken as a system are the cause of human thought; and that these parts do not, taken individually, possess any sort of semantic content. Each of these objects is one such that when it receives certain inputs (electrical firings, chemicals such as neurotransmitters or neuroinhibitors, etc.), then it produces certain effects, such as transmitting electrical impulses through certain output channels, increasing or decreasing the concentration of certain chemicals, etc. At their basic level, these components of the brain can be understood mechanistically, or at least in terms of a cause-and-effect relationship between possible inputs and outputs. From these physical components, semantic content can arise. Indeed, all human thought somehow arises from interactions of those components. Therefore, we have established that it is possible for thought, minds, understanding, and so forth, to arise from non-meaningful components if those

components are arranged in a certain way. Can we use this conclusion to defend the systems reply? How does Searle incorporate this fact into his own conclusions?

Searle argues that "what matters about brain operations is not the formal shadow cast by the sequence of synapses but rather the actual properties of the sequences." (Searle, 1980, p. 422) His motivation for assuming that non-brain systems do not understand seems to be the Chinese room example itself. I reject his conclusion because the Chinese room itself does not demonstrate that the machine or total system does not understand, for the reasons given earlier. Is there any other reason to believe that the brain may be special as compared to the hardware of a digital computer?

First, we need to compare a mind made of brain components[11] to a mind made from a symbol-processing engine. Are there relevant differences? When I discuss the argument laid out in my syllogism, a common argument arises against it: there is a fundamental difference between a brain whose components are physical objects, and a program whose components are abstract symbols. The former is defined in terms of both hardware and software (even if both are identical or inseparable), whereas the latter is independent of its instantiation method and exists primarily as abstract symbols rather than physical objects.

As mentioned earlier, it is a common misconception about computation that a computer program exists as a pure abstract string of symbols. Certainly it can be *represented* as such, and might even exist only in this state. For example, a program that is written on a napkin by a bored programmer, and never run on any system. But when a program is being instantiated by a computer, its nature is very different. In Turing's

[11] Individual cells, synapses, or any other physical unit(s) of the brain.

Computing Machinery and Intelligence, a digital computer (or Turing machine) is described as consisting of a store (memory), executive unit, and control (Turing, 1950, p. 437). The symbols on the tape, when read, correspond to physical changes which take place in the computer. The control is built such that these actions necessarily happen. This is an important point of Turing's definition, and one which still holds true of computer systems. The system is built such that the instructions of the program will happen through pure mechanical action. In other words, there exists a physical reality, when a program is operating, such that every command in the program necessarily leads to a certain physical state from which only one sequence of motions can occur. This means that if you write a program such that it produces the same outputs as a given person (or mind) would do, then there is a physical system (the computer) that is mirroring every physical process *necessary* to produce those outputs. I emphasize the word "necessary" because I wish to be clear that the program's computer might go about producing its results in a way that is mechanically quite different from what a human brain might do – however, the most basic structure of the computation will be present by definition, since the correct results can be produced. The abstract/physical distinction does not hold; any program being run *necessarily* involves some sort of physical structure that is capable of correctly executing it. Surely this applies even when the program is being executed on an unusual platform; even a program being run inside of a human mind, or a series of water pipes, or even a nation's telephone grid, must produce a physical structure capable of realizing its effects. This is assuming, of course, that the hardware is such that it makes few errors and faithfully implements the program. This condition may affect our intuitions about programs run in human brains or with humans as elements of the program, since we do

not generally think of humans in the mechanistic way in which we think about computer processors.

Recall the earlier observation: "What people forget is that there is no software. There are only different hardware states." Because it is obvious that the appropriate program does not produce any understanding or mental states whatsoever if it is not implemented, I intend only to defend the possibility of understanding being granted to programs being implemented on hardware of some sort. There is no argument that a brilliant program simply written on a piece of paper or sitting in a text file is inactive and generating neither understanding nor mental states. In this condition, the program indeed has no existence beyond being a series of symbols that can have power only when interpreted by a being whose thoughts have semantic content – such as a human programmer. On the other hand, the moment such a program is "run" on a system whose specifications are sufficient to correctly perform it, the processes and activities outlined in the program gain physical reality. A program that produces a list of prime numbers must have the physical capacity to store and represent prime numbers in some form, and the capacity to calculate further ones. There *must* exist some concrete system which can be thought of as the "brain" of a program. This is by definition true even of exotic computers (the water pipes, punch cards, etc.) – being capable of running the program *means* being able to represent all necessary elements, among other things.

We have established that a program being instantiated on a computer is not purely symbolic in nature; it must correspond to an appropriate physical system. The program determines the movements and changes that this system undergoes. Therefore, it does not seem apt to say that a computer program has only symbolic meaning, whereas human

mind has semantic meaning, if what is meant is that the computer program exists only in the form of abstract symbols and the brain does not.

The physical brain and the physical computer hardware have no important differences in this regard: both systems are based on a physical object which, through basic physical laws, effects the operations necessary for the correct outputs.

Steven Savitt (1982) makes a similar argument about the significance of a brain-simulating program. His argument begins by supposing that a human brain is composed of *n* neurons. If we remove one neuron and replace it with its simulation, it seems obvious that understanding does not therefore disappear. Call this state (one replaced neuron) *n – 1*. Continue the process until *n = 0* and you have the Chinese Room. Every neuron is replaced with its simulation in a large program. At what point does understanding disappear? Since there is no obvious point at which understanding can be clearly seen to vanish, why not grant that it is present even in this final state?

Unfortunately the impossibility of defining a specific line where the qualitative change happens does not prove that both extremes are equivalent. Savitt argues, as I do, that the distinction between the actual brain and the faithful simulation is unimportant. At the very least, this case illustrates how the distinction between simulation and genuine article can be reduced to a graded scale rather than an absolute division. This provides the basis for a sufficiently thorough simulation to be considered as good as the so-called real thing.

There is another possible meaning to the syntax/semantics objection. The argument exists that a computer program may possess syntax and a physical reality, but that programs cannot contain semantic content (or that it remains to be demonstrated

that they can). In response to this point I simply repeat my earlier argument: the human brain is capable of producing semantic content, the human brain is a physical object moving according to certain laws whose component parts have no semantic content, therefore it is possible for semantic content to arise from a physical object ordered in the proper way. Certainly it is not the case that all programs have semantic content - clearly there are certain minimum requirements for such content to be produced by a physical system – but we have established here that it is not, as Searle claims, *impossible* to obtain a semantically meaningful system from a syntactic one. Since computer programming syntax consists of a series of symbols which lead to certain physical states in a computer, then when discussing the possibility of obtaining semantic meaning from syntax with a computer program, we must allow that the definition of "syntax" can be taken to include a series of commands which cause physical objects to move in a certain way (the sense used when discussing computer programs).

The argument I wish to defend as to why we ought to grant the capacity of understanding to certain programmed computers is derived from the two points established above: first, that the human brain is capable of thinking, understanding, and attributing meaning (semantics), even though the human brain is no more than a very complex physical object, and second, that any software being executed on a computing system necessarily entails a certain structure in hardware – that is, physical objects. This hardware could be any computing device of sufficient power and capacity for the program in question: anything from a modern personal computer to a human brain to a series of water pipes and valves, properly configured.

If we accept these two premises, then we cannot rule out the *possibility* of a program capable of understanding simply on the basis that it is a program. The mere fact that the program's "smallest parts" are simply objects following rules is not a disqualification from its possessing understanding, as shown through the case of the human brain. What criteria should we then use when we are deciding whether or not to attribute understanding to a given subject?

2.2 In Defense of Behaviourism

I believe, contra Searle, that the answer is: observe their behaviour. Given the premise that both human brains and computers are physical objects, I will show that we ought to base our decision whether to attribute "understanding" to a given subject based on its observed behaviour. Thus, any computer system that can satisfactorily emulate some capacity of the mind *has* that capacity - understanding or otherwise. It is worth noting that it may make a very big difference which capacities are being claimed of a given program and which are not, as I discuss later in my section on concessions to Searle. In order to steer clear of certain issues that arise when discussing computers mimicking only a narrow aspect of a human mind (such as the ability to understand stories, or play chess, but nothing else) I will employ my own thought experiment for this section. Later in this thesis, in the section "Concessions to Searle", I will examine in greater detail why we ought to be cautious about granting such properties as "understanding" to a program that produces only a narrow range of behaviour types, such as an accounting program, a simple linguistic translation program, or even a story-comprehension program. Suppose

that we had a human-imitating program of a quality utterly unheard of as of this writing, a program that duplicated every human capability which can be experienced through the medium of a modern personal computer. This program is therefore able to communicate through textual and voice chat, and perhaps even create a video image of itself speaking. This program, we suppose, has no trouble at all with the Turing test in any form. It flawlessly demonstrates the ability to handle complicated or inter-referential sentences, remembers details, takes initiative, occupies itself when not being directly interacted with by a human, appreciates and even produces written stories, articles, poetry, songs and other such creative media, is able to "get" humour and sarcasm as least as well as the average human, etc. In short, this computer program duplicates every "output" of an ordinary human except of course the possession of an organic brain and body. Let us call this program Alicia for ease of reading. Should we believe that Alicia has a mind, is capable of thought or understanding? After all, it is not uncommon in today's world to have acquaintances one has never met or interacted with except through the media of text, voice and video correspondence. Presumably we grant these acquaintances understanding even though we have not experienced firsthand their possession of a biological body. Therefore, having known and interacted with Alicia for many years, should we attribute to her the power to understand?

Searle has a reply to a similar question (the robot reply), in which he answers that *of course* we should grant Alicia, or any convincing human-imitator, the status of understanding as well as any other mental properties we ordinarily grant to other humans – so long as we do not know that she is a fake. The instant that we know that her mind is not derived from a flesh and blood brain, we should no longer consider her to be a person,

nor possessed of a mind, nor able to understand. (Searle, 1980. p. 420) The reason, he implies, that we assume that all things which look or behave like humans possess minds and thoughts is one of convenience – and an application of Occam's Razor. One *could* be a skeptic about minds other than one's own, but this is impractical, philosophically unattractive and rude. Rather, we take the stance that anything that passes for human is a human more or less like ourselves – therefore possessed of a mind. This assumption is a very safe one because, in the past, the risk of encountering something that was *not* a human but was able to pass for a human even for a brief time was extremely rare. Chatterbot programs such as ELIZA or ALICE did not exist. There was no conceivable way that a being that was talking to you for more than a few words would not be a human. The best candidates for passing the Turing Test prior to the 20th century would probably have been a few unusually talented parrots. But this assumption is based on a premise which is no longer entirely true; things that can pass for human, at least for a time and under the right conditions, *do* exist, and become more common every day. We should therefore be more cautious in our attribution of such properties as having a mind, and be quicker to withdraw them when we have reason for suspicion. Therefore a program's mere ability to *pass* for human is insufficient to demonstrate that it possesses understanding. We grant understanding only to humans, and the moment evidence arises that Alicia is not a *bona fide* human, we withdraw our assumption that she has a mind, whatever other properties she may have.

What is wrong with this position? In order to attack this argument, we need to examine what exactly it is about humans that we think distinguishes them from Alicia. We grant minds to other humans as a matter of convenience. Is this based entirely on our

belief that they are human and not on their behaviour? I intend to show that our tendency to grant that other beings have minds is based largely on their *behaviour* and not merely our belief that they are biologically human. – therefore demonstrating that this tendency is not leading us astray when we apply it to Alicia and the like.

Consider the case of a human who is dead, or brain-dead. This is a being indistinguishable in form, mass, and basic structure from a human. In purely physical terms, a corpse or a brain-dead human is infinitely more similar to a living human than is a computer to a living human – and yet we would be foolish to say that a dead human, or a human with sufficiently extreme brain damage, has a mind. Why would we assume that the humans in these cases do not have minds? The first reason is that we intuit that a mind requires some sort of thinking thing – a working brain. Neither case has a working brain. But more primitively, we deny them a mind because we are not convinced by their behaviour. In both cases the human fails to display intelligent or intentional behaviour. This is one case in which a human with a brain is not granted the status of mind based purely on their being human, but also based on their behaviour.

Now consider the case of a person with a severe mental illness, who appears not to recognize or remember other people, whose plans are irrational or inconsistent, who is unable to communicate meaningfully, form or execute plans, etc. despite their body, other than their brain, being sound. There are surely some cases of this nature in which a reasonable observer would refrain from granting understanding and perhaps even the possession of a mind, in the most extreme cases (complete persistent vegetative state with no visible brain activity beyond the minimum required for life signs, for instance). Again, it

seems that observed behaviour is an important factor in whether or not our intuition to grant "mind" to an entity applies.

Thus, our intuition to grant minds and understanding to other beings is not entirely motivated by our belief that the beings in question are humans. It seems that behaviour is a factor as well. Searle would probably be happy to grant this; it does not weaken his position, because he need only stipulate the caveat that both being human and demonstrating intelligent behaviour are necessary conditions, and that what is *sufficient* for understanding is that *both* are present. This still rules out any computer program run on a non-human computer, however sophisticated its behaviour. All of this labour to show that both conditions apply, however, is not in vain. We now ask, "Why is being biologically human a relevant feature?" I will answer this question in Chapter 3, in which I establish ability to pass a revised Turing test, rather than biological status, as the proper criteria for detecting the ability to understand.

2.2.1 The Martian Problem

In *Minds, Brains and Programs*, Searle briefly addresses the possibility that humans could encounter beings who possess intentionality but whose brains are biologically very different from human ones, in this passage:

> "It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality (I am, I suppose, the instantiation of any number of computer programs), but as far as we know it is because I am a certain sort of organism with a certain biological (i.e. chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena. And part of the point of the present argument is that only something that had those causal powers could have that

> intentionality. *Perhaps other physical and chemical processes could produce exactly these effects*; perhaps, for example, Martians also have intentionality but their brains are made of different stuff. *That is an empirical question*, rather like the question whether photosynthesis can be done by something with a chemistry different from that of chlorophyll." (Searle, 1980, p. 422, my emphasis)

Here Searle seems to want to make it clear that he does not believe that human brain is the only *possible* way to achieve genuine intentionality – rather, he wishes to limit his conclusion to denying the possibility of a formal program being sufficient for understanding. The problem with his assertion is his statement that the Martians' being capable of possessing intentionality is an empirical question.

Searle grants that it is possible to exactly duplicate the outputs of an intentional being using methods that do not generate genuine understanding – hence, he argues that behavioural tests are insufficient as a test of understanding. Since he grants that the mere exhibition of apparently intentional behaviour is insufficient to detect the presence of real understanding, how could we determine if the Martians were in fact intentional beings and not a sort of zombie, no more intelligent than the Chinese room itself?

Searle might be arguing that we need only test any Martians we encounter to see if their brain processes are purely formal programs. If they are, then presumably he would say we should not grant them intentionality, and treat them like convincing forgeries of intelligent life. It is not evident to me what sort of tests would determine as a certainty whether or not the Martian brain was, in fact, executing a formal symbol manipulating program in the same way that, say, a digital computer might.

In other words, Searle's position on what constitutes the proper sort of thing to have understanding or intention appears to make it impossible for him to ever establish

whether or not he ought to treat a Martian as an intentional being. He cannot determine based on their *behaviour* whether or not they understand, since he already grants that in principle a program might exactly duplicate the outward behaviour of a human and not understand a thing. What further criteria could he apply when examining their brains to decide whether or not he would grant them understanding? If their brains were made of too strange a substance, or if the operations of these brains were too similar to those of a digital computer, how would Searle react? After all, Searle seems to have no objection to the notion that a being might act like an intelligent creature without possessing understanding. He therefore has no reason to suppose that the Martians do or do not possess genuine understanding, and cannot assume the position that they have intention by default.

It is also interesting to consider this scenario: suppose that a Martian – or even, Mr. Spock of Star Trek – were to land on earth tomorrow and announce his goodwill toward humankind. After humans scan Spock's brain, it is discovered that his brain processes can be described in terms of formal symbol manipulations, and that his brain much more closely resembles a digital computer than that of a human. Would Searle then decide that humans had no more obligation toward Spock and his race than we have toward a personal computer or a filing cabinet? The possibility that, somewhere, intelligent-seeming beings might have evolved brains that do not qualify as intentional by Searle's standards is a case that shows that his intuitions about understanding may lead to some very confusing conclusions when applied to cases beyond humans and digital computers.

3 The Necessity of the Turing Test

The Turing test as originally proposed by Turing is imperfect. There is a need for the somewhat speculative format of the test in Turing's original paper to be modified. There are several reasons for this: philosopher Stevan Harnad puts forward a number of revisions to the test in his article *Minds, Machines and Turing*. I will outline Harnad's basic objections to the standard "as-written" Turing test insofar as they mirror my own position. Harnad makes a large concession to Searle when outlining his proposed test of machine understanding, however – a concession that I believe to be in error, and will rebut here as I outline my own version of the Turing test.

3.1 Basic Objections to Turing

Why should we care whether or not a given being is biologically human if its behaviour appears to be very human-like? There are many possible reasons to believe that the ability to pass for human is not a reliable indicator. I consider the more common ones here.

The first is one that Turing considered in his original defence of the Turing Test - the argument that humans are capable of emotions, feelings, understanding, mental states, pain, pleasure, love, being resourceful, taking initiative, being creative, etc. which, an opponent of the Turing Test might claim, are not necessarily going to be present even in a computer program that perfectly mimics human outputs. Alan Turing identified this brand

of objection as "The Argument From Various Disabilities"[12], and I will not consider it in great detail. Rather I will quote Turing's reply:

> "No support is usually offered for these statements. I believe they are mostly founded on the principle of scientific induction. A man has seen thousands of machines in his lifetime. From what he sees of them he draws a number of general conclusions. They are ugly, each is designed for a very limited purpose, when required for a minutely different purpose they are useless, the variety of behaviour of any one of them is very small, etc., etc. Naturally he concludes that these are necessary properties of machines in general. […] A few years ago, when very little had been heard of digital computers, it was possible to elicit much incredulity concerning them, if one mentioned their properties without describing their construction.[…] The works and customs of mankind do not seem to be very suitable material to which to apply scientific induction. (Turing, 1950, p.335)

In other words, many of the properties of intelligence, creativity, initiative, and so forth, which we today deny to computers, could be attributed to the fact that no computers or machines to date have displayed these properties. This, as I hope the history of computation has shown, is a very poor reason to believe it is impossible in principle for computers to display such properties. It was once believed that computers would never be capable of playing Chess at the level of a human expert. Such beliefs about the limitations of computers have a way of embarrassing those who support them once ten or

[12] Turing, 1950

twenty years have passed. In the same way, the idea that a computer cannot think (which I estimate is fairly popular at the moment) may seem just as foolish in a few decades as the belief that a computer could never defeat a Kasparov or pilot an airplane, when perhaps the sight of a computer communicating at a near-human level of apparent intelligence becomes more commonplace. As for those "deficiencies" of an emotional nature, such as the presumed inability of a computer to feel love, anger, sorrow or joy – or, for that matter, pain and pleasure –there are two avenues of response. First, we could argue, as Turing does, that these are frivolous requirements. There exist many humans who, allegedly, do not experience certain emotions (or in some cases, *any* emotions). We do not therefore claim that they do not possess minds or the general capacity for thought, mental states and understanding. Before such an objection can be raised, the case must be made that emotions are a necessary part of having a mind and not merely an accidental feature that does not directly contribute to the ability to understand. Second, even if one insists that emotions and the like *are* essential parts of mental life, I do not think that this need trouble a programmer overmuch. Of all of the elements of human experience, emotions are probably one of the most primitive and basic features; as evidenced by the fact that many nonhuman animals whose brains and behaviour appear much more simple than that of humans (consider cats, dogs, birds, etc.) seem to experience them. At minimum, all that is necessary to make any artificial person display emotions is to make their emotional state have the "correct" (plausible) effect on their behaviour. This means that designing a chatterbot to "feel" anger could be as simple as having an "anger" variable whose value effects the text outputs given. It is not necessary to wonder whether the robot can actually subjectively experience these emotions, so long as their behaviour is consistent with a

human's under similar circumstances –. In fact, the 1998 interactive fiction computer game *Starship Titanic*, written by Douglas Adams of *Hitchhiker's Guide to the Universe* fame, boasted several chatterbot artificial intelligence characters that interacted textually with the player. These bots had several separate emotional dimensions which were influenced by the player's inputs and also influenced the bot's outputs to the player. Naturally these were constructs of a very rudimentary nature, but the essential point is that it is not difficult to imagine in principle how to duplicate the effects of emotion on behaviour when programming an artificial mind. Even subtle emotions such as amusement, love, hatred, etc. seem to possess nothing that might put them off-limits of computer thought, however flattering it may be to us as humans to imagine that they are.

Having dealt with the objection from various disabilities as to why being human ought to matter when determining the possession of a mind, we move on to Searle's objection; that something about the *brain* matters. Searle's arguments about programs and computers being fundamentally different than minds and brains are relevant here. In particular:

> "[…] the distinction between program and realization has the consequence that the same program could have all sorts of crazy realizations that had no form of intentionality. Weizenbaum (1976, Ch. 2), for example, shows in detail how to construct a computer using a roll of toilet paper and a pile of small stones. Similarly, the Chinese story understanding program can be programmed into a sequence of water pipes, a set of wind machines, or a monolingual English speaker, none of which thereby acquires an understanding of Chinese. Stones,

> toilet paper, wind, and water pipes are the wrong kind of stuff to have intentionality in the first place – only something that has the same causal powers as brains can have intentionality – and though the English speaker has the right kinds of stuff for intentionality you can easily see that he doesn't get any extra intentionality by memorizing the program, since memorizing it won't teach him Chinese." (Searle, 1980, p. 423)

This conclusion is, I think, at the heart of Searle's position. I also think that it is based on false premises. First, the offhand claim that "toilet paper rolls, wind, pipes, etc." are the *wrong sort of thing* is flawed in two very important ways which Searle neglects to mention. First, *scale*. Let's discuss a program that is as good a candidate as any for the title of being a program that, when instantiated, creates a mind: Alicia. Recall that Alicia can display an incredible range of human abilities, including convincing verbal communication over any scale of time, the ability to write songs, evaluate art, learn new skills and make acquaintances, etc. All of these tasks require immense computational power. The program required merely for the Chinese room (which would be much easier, I presume, than an Alicia program), in order to produce a convincing answer for literally any Chinese question, would have to be extremely complex and, given the necessary operations, probably on a greater order of complexity than the any existing computer program. Now, keeping in mind this staggering complexity, imagine that you were required to build a computer out of toilet paper rolls and stones, and then cause it to execute a program such as Microsoft Word or even the scientific calculator program found on some PCs. The contraption it would be necessary to build – the amount of stones and time and the scale involved – would result in either a device the size of a house, or a running time of the age

of the universe, or any other absurdly large figure you care to name. Now scale it up to a program for all human behaviour like Alicia, or even just the Chinese room. We are not dealing with a simple cardboard tube with some pebbles in it – we are dealing with a network of pipes, valves, and pumps the size of a solar system. By definition such a construct would also be able to carry on a conversation and display a wide range of human traits such as memory, outward emotion, etc. Searle appeals to our intuitions that these "novelty" implementations are simply absurd little inanimate objects that clearly cannot contain intention – but recall that the programs they are being required to instantiate are of staggering complexity - and, more important, that you can *talk to them*. Again, the solar-system-sized maze of pipes and water that is instantiating Alicia could, *by definition*, carry on a conversation (provided you know how to convert its outputs to human language and vice versa). When dealing with an entity such as this, we should definitely be extremely cautious of our intuitions before making sweeping statements such as that they obviously have no form of intentionality. This is especially true when the only plausible way to make sense of an Alicia program, even running on water pipes, would be to take the intentional stance[13] with it, due to its complexity. Searle's (unjustified) statement that these objects are simply the wrong sort of thing is begging the question; it is an axiom of the systems reply that any physical system could, in principle, host a mind. Searle, I daresay, is hoping that the reader will conjure to mind a pitiful toilet-paper roll with some stones in it, or the image of a few gears turned by wind-powered fans. These simple constructs are indeed as incapable of understanding, thinking, or

[13] That is to say, we would predict the behaviour of the object by assigning intentions and goals to it, in the same way as we attempt to predict the behaviour of other humans. To predict the responses of the complex pipe system using brute physics would likely be as useless as it would be with a human brain.

having a mind as are a cluster of three or four brain cells, since in order to run even a simple program they would require time on the order of years – for a program like Alicia, they would require the lifetime of the universe to produce two or three sentences worth of computation (again, this is almost surely an *under*estimate). Analogously, we do not conclude from the fact that a small number of brain cells do not seem to be capable of understanding that a complete human brain is likewise incapable; why should we have a double standard when considering media such as circuit boards (or even water pipes)?

Therefore Searle's claim that certain non-brain objects cannot understand requires more support than he offers.

3.2 Harnad's Objections

Stevan Harnad examines what criteria should be used for a more conclusive Turing test. I agree with his basic outline of the test, and will include his thoughts on that subject. However, Harnad also makes what I believe to be a crucial error in defining his test.

Harnad identifies five levels of detail for Turing tests:

> "Turing's celebrated 1950 paper proposes a very general methodological criterion for modeling mental function: total functional equivalence and indistinguishability. His criterion gives rise to a hierarchy of Turing Tests, from subtotal ("toy") fragments of our functions (t1), to total symbolic (pen-pal) function (T2 – the standard Turing Test), to total external sensorimotor (robotic) function (T3), to total internal microfunction (T4), to total indistinguishability in every empirically discernible respect (T5). This is a "reverse-engineering" hierarchy of (decreasing) empirical underdetermination of the theory by the data." (Harnad, 2000, p. 1)

Harnad believes that T1, the "toy" level, is plainly too lenient a test of understanding and possession of a mind. In section 4.3, I examine in greater depth the idea that programs that emulate only some small part of human ability ought not to be said to understand – even if the understanding that we attribute them were to be limited to their particular specialties (for instance, we should be wary of granting that a chess program "understands" chess merely because it can produce strong chess moves).

Harnad also rejects the T2 model: the "pen-pal," with whom we exchange symbols to communicate. He says that the T2 model is satisfactory enough, except that it can be revealed as a fraud when we subject it to what he calls the "Searlean Test": if you were to read and execute the program that passes the T2 test, you would not yourself have any understanding of the supposed "mind" being produced. Indeed, the fact that a person executing the code does not gain any understanding of the program's mind is, on Harnad's view, the death stroke for the T2 test. When he considers the position of the systems reply – that the code might produce a "mind" even if the person running the program by hand is not aware of it – he replies thus:

> "This is spooky stuff. We know about multiple personality disorder – extra minds cohabiting the same head, not aware of one another – but that is normally induced by early sexual abuse rather than by learning to manipulate mindlessly a bunch of meaningless symbols. So I think it makes more sense to give up on cognitive computationalism and concede that either no implemented code alone could ever successfully pass the lifetime T2, or that any code that did so would nevertheless fail to have a mind; that it would indeed be just a trick. And the way to test this would be via the "Searlean Test" (ST), by being the candidate, executing the code oneself, and failing to experience the mental state (understanding) that is being attributed to the pen-pal. As one has no intuitive inclination whatsoever to impute extra mental states to oneself in the face of clear 1st-person evidence that one lacks them, the ST should, by the transitivity of implementation-independence, rule out all purely computational T2-passers. A purely computational T2-passer, in other words, *is* Turing distinguishable from a candidate with a mind, and the way to make the distinction is to be the candidate, via the ST." (Harnad, 2000, pp. 13-14)

I think this is a curious conclusion to adopt. What would a program look like, I wonder, that *did* impart new understanding on the person executing it? If such a program were written, would Harnad grant that it possesses a mind even when being executed by a system of water pipes that formed a makeshift computer? Presumably, a program that imparted understanding of its operations to a person who was implementing the code in the style of the operator in the Chinese room would simply be a program with a great deal of "commenting" – messages to the reader interspersed in the code, explaining what each operation does. For instance, "transpose symbol X to position Y" might become something like "We have now determined that symbol X is the name of the interrogator, who has asked us our name. We will transpose symbol X into our response such that we may reply in the form 'nice to meet you, X'. Therefore, transpose symbol X to position Y. " It is certainly not obvious that the presence of these notes to the reader, which need have no actual effect on the outcome of the program when run on a conventional computer, would add any greater plausibility to the notion of a given program being sufficient to generate understanding or not.

There is a further objection which is even more problematic to Harnad's premise that a program can be disqualified from possessing understanding for failing to impart understanding to its operator in the Chinese room. Namely, since Searle is happy enough to admit that the human brain can be conceived of as a machine instantiating a program (or programs), it seems doubtful in the extreme to me that this program, when executed in the Chinese room, would impart the least understanding to the operator. This objection is of no trouble for Searle himself, naturally, since he insists that the medium of implementation is all-important, and that even the correct "program" of a human mind

achieves understanding in virtue of the brain itself, not the mere fact of being executed. Harnad has no such stipulation. Since we grant that the human mind program produces understanding when it runs in a human brain, and it seems obvious that it would impart no understand whatever on a person implementing it in the Chinese room, it must be the case that the operator's understanding anything when implementing a given program is neither a necessary nor sufficient test of whether the program in question can generate understanding simply by being implemented. Thus, I conclude, contra Harnad, that a program need not produce understanding in a human operator in a Chinese room scenario in order to be a potential candidate for the generation of genuine understanding. With this worry assuaged, we are able to focus solely on whether or not the program produces outputs which can pass for human on a long time frame with a convincing level of (apparent) intelligence, creativity, emotion, and other human characteristics.

3.3 Against Turing: Blockheads and Understanding

Ned Block has argued against the Turing Test as a standard for detecting intelligence. In his 1981 paper "Psychologism and Behaviourism" he sets forward a thought experiment meant to demonstrate that the Turing Test will "pass" entities which are self-evidently not intelligent ones. Block's arguments are relevant not only to my advocacy of the Turing Test, but also the functionalism and behaviourism I endorse in my reply to Searle. I shall therefore address Block's argument and show that his proposed counter-example does not in fact succeed in debunking the Turing Test as an adequate measure of intelligence or understanding.

Block discusses what would constitute a fair and charitable definition of the criteria of the Turing Test and arrives at this: "Intelligence (or more accurately, conversational intelligence) is the disposition to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be." (Block, 1981. p. 4). Block calls this amended version of the Turing Test thesis the "Neo-Turing Test". This definition is *nearly* acceptable for the purposes of this thesis, given that the use of "disposition" addresses counterfactual cases such as those I discuss in section 4.2.There is, however, one important detail that Block fails to note in his definition of intelligence as put forward by the Turing Test, and this is the fact that the test is to be carried out with a limited timeframe. Specifically, the machine attempting to pass the test must consistently answer quickly enough to pass for a human. This detail will become very important in my response to Block.

Taking the Neo-Turing Test as the criterion for intelligence, Block defines a machine that will pass the Turing Test while failing to be intelligent. The principle of the machine is to contain in advance a table of every possible sensible exchange of phrases in a dialogue. Here, "sensible" is taken in the broadest possible sense – it excludes only completely absurd statements. Given that there are a finite number of sensible English sentences in a given conversation between two speakers for a finite number of exchanges – for instance, suppose one thousand total phrases between both speakers – then all of these possible conversations are also finite in number. Thus, it is in principle possible for a team of clever programmers to create a branching "tree" of all possible conversations, matching each possible statement by the machine to another statement by the human

judge, and then further possible responses, and so on. Block gives this example of its functioning (using "string" interchangeably with "sentence"):

> "Imagine the set of sensible strings recorded on tape and deployed by a very simple machine as follows. The interrogator types in sentence *A*. The machine searches its list of sensible strings, picking out those that begin with *A*. It then picks one of these *A*-initial strings at random, and types out its second sentence, call it ``*B*''. The interrogator types in sentence *C*. The machine searches its list, isolating the strings that start with *A* followed by *B* followed by *C*. It picks one of these *ABC*-initial strings and types out its fourth sentence, and so on." (Block, 1981, p. 10).

Block's program (also affectionately known as a "Blockhead") is thus a lookup table. Every possible input is simply matched to a list of suitable outputs, from which one is selected at random. By definition, the responses produced by this program will be as convincing as those of the programmers themselves, or perhaps even *more* so, since they will be the result of the collaborative work of several authors. Therefore we must grant that this program will pass Block's definition of the Turing Test. Simultaneously we must admit that no possible intelligence could exist in so simple a structure. Thus, the Turing Test is defeated. As Block puts it,

> "[…] the machine has the intelligence of a toaster. *All the intelligence it exhibits is that of its programmers.* Note also that its limitation to Turing Tests of an hour's length is not essential. For a Turing Test of *any* given length, the machine could in principle be programmed in just the same way to pass a Turing Test of that length." (Block, 1981, p. 10, original emphasis)

Block compares the functioning of such a program to a two-way radio. If a two-way radio were entered in a Turing Test, it would surely "pass," provided that a human operator was providing the responses for the radio. Yet no-one would grant that the radio

was therefore intelligent. "All of the intelligence here," we would object, "is that of the human operator. This machine is just transmitting their responses to us." Block then proposes that his program is just the same as the two-way radio, except that the conversation has been stored in advance. It is nevertheless merely reproducing word-for-word a response thought up by a human operator, and is therefore no more intelligent than is the radio. Block continues:

> "I conclude that the capacity to emit sensible responses is not sufficient for intelligence, and so the neo-Turing Test conception of intelligence is refuted (along with the older and cruder Turing Test conceptions). I also conclude that whether behavior is intelligent behavior is in part a matter of how it is produced. Even if a system has the actual and potential behavior characteristic of an intelligent being, if its internal processes are like those of the machine described, it is not intelligent." (Block, 1981, p. 11)

If the capacity to emit sensible responses is insufficient for intelligence (or, we may infer, understanding), then naturally my position that a suitably thorough Turing Test is untenable. It therefore falls to me to answer: what is wrong with Block's argument?

Block lists a number of possible objections to his conclusions in his paper. My objection is the one he lists as #6:

> "Combinatorial explosion makes your machine impossible. George Miller long ago estimated (Miller et al., 1960) that there are on the order of 10^{30} grammatical sentences 20 words in length. Suppose (utterly arbitrarily) that of these 10^{15} are semantically well formed as well. An hour-long Turing Test would require perhaps 100 such sentences. That makes 10^{1500} strings, a number which is greater than the number of particles in the universe." (Block, 1981, p. 16)

Block replies that he only claims his machine to be *logically* possible, not realizable in actuality. Since the Neo-Turing Test is not meant to be a conceptual claim, rather than an empirical one, then the mere logical possibility of an unintelligent machine capable of passing the test is a sufficient refutation. (Block, 1981, p. 17)

Block's reply fails to adequately address the one crucial difference between the performance of his machine and the performance of an actual human when subjected to a

Turing Test: response time. Given the astronomical size of the list of possible statements used by his program, it is certain that in practical terms it would be exceedingly slow in its responses if the hardware instantiating it were any sort of currently extant processor – slow enough that a one-hour test would likely be insufficient for even a single response from it. This is obviously the sort of problem that would cause the program to fail an actual Turing Test, even though the responses it would eventually look up might be very convincing ones. In order to pass for human in an actual test (which, as I contend, is a crucial element of a reliable Turing Test), Block's machine's responses must be sped up. There are two main ways to accomplish this. I argue that no matter which of these two approaches one takes, speeding up the program enough to pass for human entails a qualitative change that makes it no longer obvious that the machine is unintelligent. For both of my responses I will assume that it is somehow possible to store all of Block's machine's possible statements on some sort of computer storage device, even though this is very unlikely given the numbers above. Instead I will concern myself only with the amount of *time* necessary to look up responses from this staggering list.

The first way to speed up the response of the machine is to improve the quality of the program. This could be done by adding processes to analyze the incoming statement and rapidly narrow down the possible responses based on the type of statement that was input. For instance, having possible responses grouped into "answers to questions", "responses to non-interrogatives", "questions", etc. Each of these categories could, itself, be further subdivided (responses to questions about myself, responses to questions about current events, etc.) to a great degree. In order to speed the process of sifting all possible statements to merely the list of suitable ones, it would be necessary to subdivide the statements very thoroughly, and to develop a sufficiently capable parser capable of analyzing an incoming

statement and determining which categories of responses it ought to consider first. In the end we arrive at a program which not only contains all possible statements but also possesses a system to analyze incoming statements and sort through its own possible categories of stored statements. In order to pass a Turing Test this program must be so effective in its analysis and categorization (or whatever efficiency techniques its programmers can devise) that it can sort through 10^{1500} sentences quickly enough to respond with *human* speed - and that figure is merely for a *one hour* test. When the analysis and sorting system of the program reaches this level of sophistication, I propose that it is no longer obvious that the program is unintelligent. It seems plausible to me that the action of parsing the incoming statement and comparing it to the various subcategories of possible responses, when it operates at this level of sophistication, is the sort of thing we look for when describing intelligence, the possession of a mind, or the ability to understand. Since such an improvement is necessary before Block's program could pass a Turing Test, one of his crucial premises – that his machine is obviously unintelligent – goes out the window. The ability to manage the list of possible responses with sufficient speed requires what could plausibly be called intelligence.

The second way to speed up the response of the machine is to improve the quality of the physical processor so that it can consider more responses in a given amount of time. The quality of the processor might be improved by structuring it differently: creating a vast, efficient network of connected nodes corresponding to different branches the conversation might take, for example. Given the number of nodes required for the one-hour test, this network of nodes would vastly exceed the human brain in complexity – by a thousand orders of magnitude. When one considers this massive, complex web of nodes processing inputs and producing responses, it becomes less obvious that it is not in fact intelligent. Just as in the case above, the key premise that the machine is obviously unintelligent is no longer true.

Another method of improving the processor is making the processor out of the right kind of materials, capable of transmitting signals very quickly, or using the right techniques of engineering, to move the necessary physical parts very quickly. The amount of improvement necessary on current techniques to sort through a list of sentences - whose number exceeds the number of particles in the universe – is so great that I cannot imagine it. The hardware which would be capable of such processing would be so far beyond what currently exists that I am reluctant to agree that the new, human-speed machine would be obviously unintelligent, if it is even physically possible for it to exist.

Block addresses this line of reply in his paper. He explains that the physical difficulties in realizing his program stem from the physical constants that happen to exist in our universe – for instance, the speed of light, number of particles extant, etc. These physical laws, he claims, are different from the *psychological* laws of our universe:

"But a situation can contravene laws of physics without contravening laws of human psychology. For example, in a logically possible world in which gravity obeyed an inverse cube law instead of an inverse square law, our laws of physics would be different, but our laws of psychology might not be." (Block, 1981, p. 18)

He continues:

"But if my machine does not contravene laws of human psychology--if it exists in a possible world in which the laws of human psychology are the same as they are here--then the neo-Turing Test conception of intelligence is false in a world where the laws of human psychology are the same as they are here. So the neo-Turing Test conception of intelligence cannot be one of the laws of human psychology." (Block, 1981, p. 18)

There are two things I wish to say in reply to this argument. First, I have some difficulty accepting the premise that there are universes in which information can be

transmitted much faster or stored more densely, but that the human intellect does not change to reflect this. It seems to me that in universes in which information-processors happen to be much more powerful than in our own, it is natural to expect the human brain to have evolved differently in order to benefit from these physical principles.

Second, even if one grants that there exists a universe in which computing devices can have much greater storage and processing power while the human brain remains unimproved, this does not affect my position. As I explain in section 4.3, the Turing Test is simply a practical method which happens to work in our universe. Thus, Block's statement of the neo-Turing thesis might be amended to include a disclaimer such as "in this universe".

In conclusion, it seems to be a fact that a brute-force program such as Block's Blockhead is unable to pass for a human in any practical Turing Test – i.e. it will always require too much time, and/or be incapable of passing a detailed, long-term test such as the one I advocate. Thus, Block's Blockhead is not a counterexample to the view I advance in this thesis. The inability of the Blockhead to pass an actual test is comparable to the problem with Searle's operator in the Chinese Room – the scale and slowness of these brute-force approaches makes them unsuitable as counterexamples to a functionalist test of understanding.

4 Concessions to Searle

For all of the antagonism it has drawn, *Minds, Brains and Programs* raises extremely interesting and relevant points of discussion. Searle raises cases where one might attribute understanding to a program, but ought not to. These facets of his argument are worth noting, even if his main thesis is incorrect. These points are:

1. It is not clear that a program that emulates only some small subset of human behaviour *does* understand by Searle's definition (which requires the possession of a mind) even if the activity the program is duplicating is called "understanding" in ordinary parlance ('understanding' Chinese, "understanding" a story, etc.)
2. Related to 1, where do we draw the line when attributing understanding? Does a calculator *understand* mathematics, does a thermometer *believe* that it is cold? Is an advocate of the systems reply forced to grant these objects mental states?

With regard to point 1: Searle's argument is one that relies on our intuitions about minds and understanding. He appeals to our intuitive sense that a system consisting of a filing cabinet being manipulated by a clueless operator does not perform the action we call "understanding" on the grounds that one of the important components of this system is, in fact, a being which is itself capable of understanding.

It is therefore important to examine what it means for a human to understand, in the ordinary sense, because this will inform our intuitions about what we do or do not count as instances of this activity. "Understanding", when applied to humans, is something that humans only *do* with the possession of a full human mind. It is difficult to conceive of what it would mean for me to, say, "understand Chinese," when the only activity that is

being done by my mind – and ever *has been* done – is this specific understanding. This does not seem compatible with a standard definition of "understand". In order to shed light on the question of what standards one ought to set regarding understanding, we will briefly re-examine Searle's use of the word.

When first mentioning "understanding" Searle specifies: " 'Understanding' implies both the possession of mental (intentional) states and the truth (validity, success) of these states. For the purposes of this discussion we are concerned only with the possession of these states" (Searle, 1980, p.424). He further elaborates: "Intentionality is by definition that feature of certain mental states by which they are directed at or about objects and states of affairs in the world."

Understanding therefore requires the possession of intentional states which relate to the world in the proper way. What are the criteria we normally use to grant that these states exist? We know that we ourselves have intentionality. Each individual experiences their own intentions toward the world, their beliefs about people, objects, states of affairs, etc. We also grant intentionality to other humans on the ground that they display behaviour for which the simplest explanation is that they, like ourselves, possess intentionality. Searle argues that we should not extend this privilege to digital computers. In fact, he specifically argues that "[…] the brain's causal capacity to produce intentionality cannot consist in its instantiating a program, since for any program you like it is possible for something to instantiate the program and still not have any mental states" (Searle, 1980).

I have given reasons in this paper why we ought not to agree to this premise: our opinion that other humans possess intentionality is based on a combination of our own

experience of intention and the observed behaviour of other humans. Because both the brain and the digital computer are simply objects moving through certain physical states, there is no compelling reason to believe that a given program is any more incapable of mental states than the human brain itself. Searle is basing his argument that no program is sufficient for mental states on thought experiments such as the Chinese Room, which I have argued are false analogies and do not support his conclusions.

4.1 Can a Rock Think?

In "The Rediscovery Of the Mind", Searle argues that there is no clear distinction between computation and non-computation, and that on standard definitions of computation any object can be considered a computer executing a program (Searle, 1992). The swirl of air molecules in a room could be understood to perform a computation, assuming that some interpreter had the correct framework to understand the movements of the particles. This argument poses some serious challenges to the functionalist or adherent of the intentional stance, because such a person is compelled to grant that understanding is literally everywhere, since a complex, Turing-test-passing program is being executed in any object one cares to choose. The complex motion of water molecules in a cup of coffee could be interpreted to be processing any arbitrary inputs one cares to choose, and producing the "correct" responses according to some framework (or at least, on Searle's view, this is so). Thus, if one embraces the principle that there exists at least one program that is sufficient for understanding, accepting the conclusions of Searle's paper means granting that all objects possess minds and understanding.

In response to this challenge, I side with Chalmers' (1996) reply that although any open system can be said to be instantiating any program provided one correctly maps the inputs and outputs, a system such as a rock, a cup of coffee, etc. will not be able to correctly compute queries other than the specific one addressed by that framework. For example, if I had worked out an interpretation by which a lake, when its surface is touched at two points, interprets this as the input "two plus two", and then the resulting ripple pattern is understood to mean "four", then I could not keep the interpretation constant but ask a different question and receive a coherent answer. For instance, suppose I wanted to instead input the problem "two plus three", which takes the form of touching the lake surface at three different points: the framework I used to calculate "two plus two" would no longer produce the correct answer. Every specific question must have its own interpretation, and cannot perform an arbitrarily large number of calculations using the same constant framework. This is a property which conventional computers, however, *do* possess; I can just as easily solve two plus two as any other basic mathematical operation on a given pocket calculator, using the exact same framework for understanding inputs and outputs every time. Furthermore, the method by which the calculator is understood to be computing the answers does not radically change depending on the problem posed to it – unlike the impromptu computers such as those produced by the lake or the cup of coffee or the stone. The right "types" of computers to possess understanding or intelligence are those that, like the human brain, have a more-or-less consistent input-output format and whose methods of calculation are not arbitrary but rather fixed (or "reliable") in that they can be generalized. My particular response to Searle's argument is not troubled overmuch by the possibility that any given object could be said to be implementing any finite

program, because the standard that I propose as definitive of understanding or the possession of a mind is a form of a revised Turing test as outlined in Chapter 3. A convincing, long-term interaction with a human being, including questions on topics requiring creativity, memory, and so forth is what I advocate. Such a test might be passed given the proper program and a minimally adequate computer. It is, however, difficult to envision such a test being foiled by a rock or a cup of coffee which happens to be given a framework by which it is understood to be implementing the "mind-having" program.

4.2 The Threshold of Understanding

Given Searle's definition of "understanding", if one accepts my arguments against the Chinese Room, should one therefore reject his position that no program can, of its own virtue, produce understanding, even in principle? Yes, absolutely. Is there any conclusion worth keeping from Searle's argument? Again, yes. Searle does score one important victory by attempting somewhat to limit the scope of machine "understanding" and mental states.

What other conclusions ought we to draw from the Chinese Room case, if not those Searle proposes? Consider Searle's conclusion that we ought not grant "understanding" to a given program merely because its actions appear to emulate some facet of human life. While I do not agree with his strong conclusion that no purely formal program could produce a mind in virtue of being executed, I do believe a weaker formulation: that the criteria for assigning "understanding" - mental states and intention – to an instantiated program should be much more specific than those he lays out as the position of "Strong

AI" in his introduction. The criteria I advocate are, however, still much more permissive than those he himself advocates.

Consider Searle's first example of a candidate program for "understanding" according to Strong AI; the story-comprehending program from Yale. Should we grant mental and intentional states to this program after seeing it in action? The answer is no. Despite the program's remarkable ability to produce correct answers to questions of a very specific nature under very specific circumstances, it does not exhibit the kind of behaviour that is properly termed "understanding".

Harnad expresses similar worries in *Minds, Machines and Turing* when he describes the T1 Turing test, or "toy" model. The "T1" Turing test is a test where the program in question is able to duplicate some limited subset of human ability – for example, the ability to play chess, but nothing more. Harnad cites Schweizer's article *The Truly Total Turing Test* in explaining why we ought not to grant understanding to automatons that only replicate small fragments of human ability:

> "Toy models, that is, subtotal models, of anything at all, have the liability that they have more degrees of freedom than the real, total thing that they are modelling. There are always more different ways to generate a fragment of a system's function (say, chess-playing – or simple harmonic motion) than there are to generate the system's total function (say, everything else a human mind can do – or all of Newtonian or Quantum Mechanics), if for no other reason than that the Total model must subsume all the functions of the subtotal toy models too." (Schweizer, 1998).

This position has direct implications for Searle's choice of example: we should be cautious of saying of the Chinese-speaking program that it understands Chinese. Suppose for instance that Bob is a human whose outward behaviour was exactly that of the story understanding program: he has no initiative, and no ability to hold any sort of

conversation or interaction outside the context of being fed a short story followed by questions. As observers, it would not be obvious to us that Bob did in fact understand much of anything. Even his behaviour in answering the questions is of such a narrow scope (he could not, presumably, answer such questions as "What would you like for lunch?" or "Why are you only answering questions about that story?"). Bob would be a sort of extreme idiot savant, and we might be hesitant to correlate his robotic, narrow replies with genuine understanding. Unlike a human who is highly specialized, Bob would not only be incompetent in areas beyond his expertise, he would be completely unable to function. For added clarity, imagine Bob's capabilities were not those of the story-analyzing program, but rather a simple chess program, or calculator program. Bob's inability to perform any activity outside of playing chess or performing mathematical problems would hint at some sort of bizarre mental hard-wiring for a specific task rather than what we consider understanding. If programs such as these do not qualify as understanding, then naturally we will also rule out absurd cases such as those brought up by Searle when he criticizes McCarthy's 1979 paper "Philosophical perceptives in artificial intelligence" (Searle, 1980), e.g. cases such as thermostats having beliefs (therefore mental states) about the temperature, etc. Even a relatively complex program such as Schank's story-question-answerer fails to unambiguously appear to understand much of anything. A human who was only able to shout out the ambient temperature every few seconds while otherwise being comatose might also have difficulty qualifying as understanding anything.

What do these cases show? Intentionality in the sense used by Searle is something we humans tend to ascribe to those whose behaviour leads us to infer that they are beings

who have a mental existence like our own. In the case of humans we are quick to grant understanding because we reason that all humans, having similar brains, will possess intentionality as the rule rather than the exception. The simplest explanation for other humans demonstrating this complex behaviour is that, like oneself, they possess understanding. Given the great number of humans who display the signs of possessing understanding, we grant it to other humans as a matter of course. Therefore our starting point for granting understanding ought to be, as Turing proposed, a test to see if a given entity can pass itself off as a human mind through sustained interaction. We grant other humans mental states on the basis of their presumed ability to demonstrate the full range of human behaviour; therefore this standard ought to be sufficient for any other being.

Thus, we ought to be very wary of saying of a program that it possesses intentionality unless it is capable of passing for a human under Turing Test conditions. Searle argues that even a Turing test is insufficient to demonstrate intentionality in a case in which we are dealing with a symbol-manipulating program, but I have offered reasons why this position is unwarranted. A program such as Alicia, which can carry on conversations at length on many topics, which consistently forms friendships and relationships, and which has likes and dislikes and tastes in books and films, should be granted intentionality for the same reasons that we grant it to any of our human acquaintance. On the other hand, in the case of a thermostat, calculator or even constructs such as a story-understanding program or a language-translating program, this condition is not met. These objects and programs cannot pass themselves off as humans in prolonged interaction, and so we do not grant them the status of "understanding beings," even though they can duplicate some fragment of human behaviour or ability. A Turing

Test (over an appropriate time frame and with competent judges) is the minimum criterion to conclusively establish the presence of intentionality/understanding in a given entity.

That said, is it possible that we could grant understanding to something/someone who demonstrates less mental ability than required for a Turing Test? There seem to be a number of counter-examples to this standard, cases that suggest that beings ought to be considered to have understanding, but who would likely fail a Turing test. I will briefly examine a few such cases.

First, consider a human who chooses not to speak, or who speaks very little. A particularly reserved and taciturn human might well fail a Turing test. Should we deny them understanding? The answer is no. As I have mentioned above, humans possess a special status as beings possessing understanding because any human has their own mental experiences as a guideline for the subjective experience of being human, and therefore can imagine the thoughts and intentions of a human who chooses to remain silent. Furthermore, there are of course an overwhelming number of instances of humans demonstrating intelligent behaviour requiring intentionality, such that we conclude by induction that all living humans have intentionality by default. In the case of a silent and reserved human, we infer by application of Occam's Razor that they have understanding. In the case, for example, of a powerful computer running a speech program that never produces a response, we have no such grounds for assuming that it possesses understanding. Unlike cases involving other humans, we do not in this case have pre-existing evidence that any given computer is capable of understanding, or that computers more generally act in a way the requires understanding. What's more, we cannot use our

own subjective experiences as sometimes silent humans to imagine what an unresponsive computer might be thinking. Thus, for practical reasons, we require computers to speak and to interact in a familiar (human-like) way, and we would be unjustified in granting understanding to non-communicative computers as to non-communicative humans.

Second, consider a non-human animal that demonstrates complex and apparently intentional behaviour, but is not capable of communicating in the usual human ways. For instance, might we be justified in concluding that a dolphin has intentionality? The case of a seemingly intelligent animal is a delicate one due to the slippery slope that is entailed when one grants intelligence to a single non-human animal. I daresay it is obvious that such animals as dolphins, most types of whales, cats, dogs, primates, etc. do indeed possess intentionality and thus understanding. As evidence to support this assertion I offer the many experiments in psychology which demonstrate an apparent capacity on the part of these animals to perform abstract reasoning, apply induction, solve various puzzles, communicate with humans, etc. as evidence that at least some non-humans have mental states "[…which] are directed at or about objects and states of affairs in the world." (Searle, 1980, Notes – defining "intentionality"). A dog performing a complex series of tricks in order to receive affection, an octopus learning to unscrew a man-made jar containing a crab, and a magpie recognizing its own image in a mirror (Prior, Schwarz, *et al.* 2008) all appear to possess beliefs about the world and themselves qualifying as intentional. Judging animals to possess understanding of a sort in higher-order cases such as these and others is easy to accept, but as one moves down the continuum of animal intelligence, it becomes increasingly difficult to estimate to what degree the observed behaviour constitutes intentionality. Consider, for example, the jellyfish. This animal may

react to stimuli and even appear to make decisions of a sort, but it lacks a brain altogether, and demonstrates only the most minimal awareness of its environment and does not appear to possess any ability whatsoever to reason, nor does it appear to hold any beliefs whose consequences extend past instantaneous reactions. Moving up the scale to such animals as shrimp, insects, small fish, birds, and various rodents, there must come a point at which one begins to grant intention to what is observed. I believe that since it is impossible to formulate an accurate, objective standard for the possession of intentionality that can cover beings unable to communicate with humans, the assignment of intentionality among such animals is subjective, with some cases lying clearly on one side or the other of any plausible delineation –with gorillas and similar primates clearly possessing at least a minimal understanding, the sea anemone or venus flytrap possessing none. How does this conclusion affect our position on the possibility of an intentional program? For one, it indicates that we ought to grant understanding to some beings that are not capable of communicating with humans and therefore could not pass the Turing test. Does this mean that we should be willing to grant intentionality to some computer-run programs that cannot pass the Turing test? As in the case of uncommunicative humans, I believe we are justified in holding a double standard for programs and biological creatures. We know that animals possess biological brains relatively similar to our own: we recognize what appear to be human emotions and attitudes in animals (greed, fear, joy, love). We possess a wealth of experimental evidence about the adaptability and ability to perform minimal reasoning and abstraction possessed by various animals under changing conditions. Our belief that animals possess mental states is strengthened by induction; we know that we think, and for reasons outlined above, we

conclude that other humans think. When we encounter an animal whose brain appears similar to ours, and whose behaviour is sufficiently familiar, we grant that it likely possesses an understanding at least minimally similar to ours. We do not extend this same courtesy to programs because we are naturally suspicious of their ability to understand the world as we do due to their very different "brains", and also because we are generally unable to have the same depth of non-verbal communication with a program as with, say, a dog. An animal possessing a familiar body has the advantage of being able to communicate its thoughts and desires through body language which resembles our own, something that naturally computer-instantiated programs will tend to lack. Nevertheless, if a computer-controlled robot were produced that could reliably display the same depth of behaviour and adaptability as an animal candidate for the capability to understand, then we ought to at least consider the possibility that it may be capable of understanding. How we ought to test such beings as robotic dogs and the like - those that communicate with humans in ways not suitable for a Turing test but that might still convey intentionality - is beyond the scope of this paper. I acknowledge only that such a test would of necessity be difficult to devise and agree upon, and that it is in principle possible for such a being to understand on the grounds that it appears possible for such animals as dogs or dolphins to qualify also.

References

Bajakian, Mark. “How to Count People,” *Philosophical Studies* 20 (2010)

Block, Ned, “Psychologism and Behaviourism” *The Philosophical Review* 90(1981): 5-43

Boden, Margaret. *Computer Models of Mind*. Cambridge: Cambridge UP, 1988.

Chalmers, David. "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108.3 (1996): 309-33.

Harnad, Stevan. *Minds, Machines and Turing,* Journal of Logic, Language, and Information, 9 (2000): 425-445.

Miller, George, Eugene Galanter, and Karl H. Pribram.
Plans and the Structure of Behavior. New York, NY: Holt, Rinehart, and Winston, 1960.

Prior, Helmut, Ariane Schwarz, and Onur Güntürkün. "Mirror-Induced Behavior in the Magpie (Pica Pica): Evidence of Self-Recognition." *PloS Biology* 6.8 (2008).

Savitt, Steven. "Searle's Demon and the Brain Simulator." *Behavioral and Brain Sciences* 5.2 (1982): 342-43.

Schank, Roger, and Robert Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Hillsdale: Lawrence Erlbaum, 1977.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Science* 3.3 (1980): 417-57.

Searle, John. *The Rediscovery of the Mind*. Cambridge: MIT, 1992.

Sergent, Justine. “A New Look at the Human Split Brain,” *Brain* 110 (1987): 1375-1392

Schweizer, P., “The Truly Total Turing Test,” *Minds and Machines* 8 (1998): 263-272.

Turing, Alan. “Computing machiner and intelligence.” *Mind* 59 (1950): 443-460

Weizenbaum, J. “Eliza - a computer program for the study of natural language communication between man and machine.” *Communication of the Association for Computing Machinery* 9 (1965): 36-45. [JRS]