Zombies in Searle's Chinese Room: Putting the Turing Test to Bed

Louis J. Cutrona, Jr.

Teleonetics Ltd.
625 North Monroe Street
Ridgewood, New Jersey 07450 USA

May 2005 - Revision 4.1

Technical Report 05-002

Correspondence should be addressed to:
Louis J. Cutrona, Jr.
Teleonetics Ltd, 625 N. Monroe St., Ridgewood, NJ 07450
(201) 447-3270, FAX: (201) 447-2547, teleonetics@aol.com
http://www.teleonetics.com/publications/TR-05-002.pdf

Zombies in Searle's Chinese Room: Putting the Turing Test to Bed

LOUIS J. CUTRONA, JR.

Abstract: Searle's discussions over the years 1980-2004 of the implications of his "Chinese Room" *Gedanken* experiment are frustrating because they proceed from a correct assertion: (1) "Instantiating a computer program is never by itself a sufficient condition of intentionality;" and an incorrect assertion: (2) "The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program." In this article, I describe how to construct a *Gedanken* zombie Chinese Room program that will pass the Turing test and at the same time unambiguously demonstrates the correctness of (1). I then describe how to construct a *Gedanken* Chinese brain program that will pass the Turing test, has a mind, and understands Chinese, thus demonstrating that (2) is incorrect. Searle's instantiation of this program can and does produce intentionality. Searle's longstanding ignorance of Chinese is simply irrelevant and always has been. I propose a truce and a plan for further exploration.

Key Words: zombie, Searle, Chinese Room, consciousness, understanding, instantiation, programming, Turing test, mind, intentionality, brain, computation, solipsism.

Writing a Program for Searle's Chinese Room

"[C]ould 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?" Searle 1980 poses this question and says the answer is no. By way of explanation, he describes a *Gedanken* experiment that he adduces as evidence of the correctness of his answer. It is serious understatement to say that there has been considerable discussion in the literature of Searle's essay, but as of Searle 2004 his view remains unchanged.

Searle describes what has come to be called the Chinese Room experiment as follows:

> [I]magine that I am locked in a room with boxes full of Chinese symbols, and I have a rule book, in effect, a computer program, that enables me to answer questions put to me in Chinese. I receive symbols that, unknown to me, are questions; I look up in the rule book what I am supposed to do; I pick up symbols from the boxes, manipulate them according to the rules in the program, and hand out the required symbols, which are interpreted as answers. We can suppose that I pass the Turing test for understanding Chinese, but, all the same, I do not understand a word of Chinese. And if I do not understand Chinese on the basis of implementing the right computer program, then neither does any other computer just on the basis of implementing the program, because no computer has anything that I do not have.

There are lots of ways to write a program, but they all start from a set of specifications. Searle's specifications require that the behavior of a Chinese Room program will be indistinguishable from the behavior of some (human) native speaker of Chinese who receives messages written in Chinese that are passed in to him or her and who composes whatever reply seems appropriate and sends it back outside. That is, the program will be able to pass the Turing 1950 test (in Chinese).

Since this is a *Gedanken* experiment, we can proceed simply, without worrying about real world practicalities. First let us find a native Chinese speaker (call him or her Zhang). The programs we write will simulate Zhang's behavior. They will differ only in the way we create them.

For the first version of the program, we have to choose a time period T for the experiment to be run. This is the length of time Searle will be in the room simulating Zhang. To make things a bit easier, rather than physically pass messages written in Chinese characters back and forth, we will transmit the characters via some standard computer encoding.[1] To prepare the program, we choose a time granularity Δt, put Zhang in the room, and systematically record Zhang's outputs for every possible sequence of inputs. There are a very large number of possible inputs, but only a finite number of them because there are only a finite number of Chinese characters, there is only a finite amount of time, it takes time to compose a message, it takes time for Zhang to compose a response and so on. Finally, (this is a *Gedanken* experiment, you will recall), we will reset Zhang at the start of each data collection run. Thus, the initial conditions will be the same every time, that is, Zhang always starts at the same age and believes the current run is the first and only run.

When we're done, we have a complete input-output specification that describes what output Zhang will produce to every possible input at every possible time. It is now a trivial matter to incorporate this information into a computer program in the form of an extremely large (humongous, to use the technical term) decision tree. The decision tree has $L = T / \Delta t$ levels. At each level n, which corresponds to time $n\Delta t$, there are as many subtrees as there are possible events at time $n\Delta t$, one labeled "no input received" and each of the others labeled with an input character. Associated with each subtree is either a single character to be output at time $(n+1)\Delta t$ or the notation "no output". The operation of the program is straightforward: at time $n\Delta t$, in the currently selected subtree, select the subtree corresponding to the (external) event at that time, which will either be nothing or a character; at time $(n+1)\Delta t$, output the character associated with the selected subtree (or no character if "no output" is associated with the selected subtree); repeat until $n = L$, which will be time T; then stop.

With suitable choice of T, say Zhang's lifespan, and Δt, say 1 millisecond, the behavior of this program will be (by construction, as they say in geometrical proofs) indistinguishable from that of Zhang in the Chinese Room. When Searle wants to do his Zhang imitation, we can print up the decision tree and let him hang out in the Chinese room doing the lookup, subtree selection, and output operations based on the printout. Searle has already stipulated that he is fast enough to do whatever is necessary in real time, and we have made it easy for him.

So, this program will pass Turing's test, assuming that Zhang would pass it (and if Zhang wouldn't pass it, we need to select another Zhang).

[1] This is just to elide the problem of character recognition, which isn't central to anything that follows.

The Zhang Zombie

This is surely the ultimate behaviorist Chinese Room program. It consists entirely of direct S-R (Stimulus-Response) links. There is absolutely nothing in the structure of this program that can be construed as *understanding* or *consciousness*. Nonetheless, we can use the approach elaborated above to construct (*pace* Disney) *Gedanken* audioanimatronic mechanisms that are indistinguishable from real live people. In fact with a little *Gedanken* biomedical engineering we can make a plug-in replacement for a human brain and create a zombie.[2] This is at least vaguely interesting because the foregoing constitutes a constructive existence proof for zombies and there are philosophers and others who have claimed in print that zombies *could not* exist.[3] But I digress.

All right. So the Turing test cannot in fact distinguish between a gigantic table-lookup system and a human being because (by construction) the table-lookup system and the human would produce identical responses. If we built Zhang zombies, they would be behaviorally indistinguishable from the real Zhang even if an x-ray disclosed that Zhang had a positronic brain[4], because by construction, Zhang's reactions to such a discovery would be in the Zhang decision-tree data.

As a practical matter the Zhang decision tree data (even the one for the Chinese Room, never mind the Zhang zombie) probably requires more bits of data storage than the number of atoms in the universe.[5] So Searle can't possibly take a printout of it into the Chinese Room with him. Moreover, gathering the data to build the decision tree

---

[2] "According to common agreement among philosophers, a zombie is or would be a human being who exhibits perfectly natural, alert, loquacious, vivacious behavior but is in fact not conscious at all…." (Dennett, 1991, p.73). Note, however, that a zombie created by the method described above would be a single-use zombie. It would always behave like Zhang. If you wanted e.g., a Searle zombie, you would have to create it by applying the technique of the previous paragraphs to Searle.

[3] Well, maybe all we've created is a Golem or an android because it is possible to distinguish it from a human being by some kind of straightforward physical analysis (it has something other than a normal human brain in its head). Zombie mavens want their zombies to be indistinguishable from human beings in all particulars except the absence of consciousness. If you want to believe in that kind of zombie, you are committed to believe that consciousness depends solely on the presence or absence of some kind of *Denkstoff* which is not (or not yet) physically detectable. If you believe it is not physically detectable, then you're a mystic of some coloration. If you believe it is physically detectable (we just don't know how to detect it at this point), you think we will eventually be able to distinguish between zombies and their (conscious) human twins by using a *Denkstoff* detector (but only a *Denkstoff* detector). That said, I don't for a second believe in *Denkstoff* of either kind and I'm still going to talk about a Zhang zombie because I think something that is physically indistinguishable from a human being is a human being and as I will discuss below, I don't think it's at all useful to argue whether human beings are conscious. Since you've managed to read this far in this footnote, I'll tell you in advance that I think arguments about zombies are really arguments about solipsism and every bit as worthwhile, viz. not.

[4] Positronic brains first appeared in Isaac Asimov's writings sometime before 1950, the year *I, Robot*, a collection of nine of his short stories, was published. The *Star Trek* character Data postdates Asimov by many decades.

[5] For starters, assume there are about 15,000 Chinese characters to choose from every millisecond and go from there for 80 or 90 years. Rough calculations suggest there are about 10^{79} hydrogen atoms in the visible universe. You do the math.

probably requires more time than is left before the end of the universe. So although we have demonstrated how to construct a Zhang zombie, we know it to be utterly impracticable.

That makes it all the more impressive that the original Zhang, with a paltry 200-250 billion active brain elements (neurons and glia) and, say, only 200-250 trillion synaptic sites (element to element data transmission points) is able to imitate the Zhang zombie perfectly and effortlessly in real time.[6] What that means is that if we built a computer that simulated Zhang's brain and it took 1 million bits to distinguish each active element and 1 million bits to distinguish each transmission point, and 100 trillion bits to describe the operation of neurons, glia, synapses, and everything else, then we're talking of a program only 200-250 x 10^{18} bits in length, which, although large, doesn't even begin to challenge the number of atoms in the Earth, which is estimated to be about 1.33 x 10^{50}, give or take.

The Chinese Room and the Turing Test

Searle seems to think that the significance of the Chinese Room *Gedanken* experiment is that it demonstrates that computers as he defines and understands them can't have minds. As we've seen, the construction of a table-lookup version of the Chinese Room program makes it clear that it is not possible to tell from the input-output behavior of the Chinese Room whether Zhang is inside or a space alien with a colloquial knowledge of Chinese or a digital computer or John Searle working in blissful ignorance of things Chinese. So one conclusion we can draw from the Chinese Room is that the Turing test doesn't buy us anything at all by itself. Too bad. It was kind of nice to think that if someone could create a program that could fool all the experts all the time, then we could feel confident that the machine was artificially intelligent.

Searle already said in 1980 that passing the Turing test may be necessary but it is not a sufficient condition for a thinking machine, but his argument was and remains flawed.[7] There is a lot of potential for confusion here. Eric Baum 2004 failed to notice the zombie solution to the problem of writing a program that can pass the Turing test, so he thought Seale was wrong to conclude that "a computer that did pass the Turing test would not be a mind." (p.76) But, as we have seen, it is perfectly possible to create a *Gedanken* program that can pass the Turing test but is, by construction, not a mind. So Searle got that one right: passing the Turing test may be necessary, but it isn't sufficient. Jeff

[6] Suppose each neuron in the brain is connected to about 10,000 other neurons, and suppose arbitrarily that on the order in half of those connections are afferent and the other half are efferent. Then if there are about 20 billion neurons in the brain and each receives input from 5000 other neurons, there must be about 100 trillion synapses in the brain and who knows how to factor in the 200 billion glial cells that cluster around synapses.

[7] This is one of those so-called Gettier 1963 cases: Searle's justified true belief doesn't add up to knowledge because he got the right answer for the wrong reasons.

Hawkins 2004, too, gets it right on one page but on the next page he turns around and says it wrong.[8]

Where that leaves us is where we were when the first ur-philosopher cum comedian said, "*Cogito ergo sum*, but I'm not so sure about you." That is, the only mind of which we have direct evidence is our own, and we have the same problem if the putative other mind is the instantiation of a computer program. This way lies solipsism. The problem with solipsism is that it is a self-consistent system. Once you are in it, you cannot prove your way out.

Searle sees the danger, and tries to avoid it with the more-or-less standard philosophical meta-argument[9] adduced to fight the temptations of solipsism, to wit: we obtain support for the surmise that others are conscious from the knowledge that they have a similar biological makeup to our own.[10] But when Searle makes the argument in this context he uses it not to invalidate or avoid solipsism per se, but only to provide a carve-out, an exception that saves him from having to apply his arguments to the question of the possible existence of other human minds, but allows him to apply those arguments to the question of the possible existence of computer minds. It's cheating.

Let's look in detail at Searle's argument. What he really objected to in his 1980 essay was the assertion by certain well-known researchers that their favorite AI (Artificial Intelligence) programs could think. Good for him; some of those claims were a bit—how shall I put it?—overreaching.

He is willing to accept that a causal mechanical process (viz. the one that takes place in the human brain) can think. So far, so good.

He allows that obviously "an artifact, a man-made machine" could think, understand, and so on, "assuming it is possible to produce artificially a machine with a nervous system … sufficiently like ours." He also grants "of course" the possibility of thought to a digital computer "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." I'm not sure what else a 'digital computer' might be under this description, but it's clear that Searle thus allows that a digital computer running a program is a 'digital computer' under this definition.

Now, a computer, suitably programmed, can simulate an arbitrary causal mechanical process. So a computer could simulate the process that takes place in the human brain[11]

---

[8] p.20: "[U]nderstanding cannot be measured by external behavior." p.21: "The only way we can judge whether a computer is intelligent is by its output, or behavior." It is apparently very hard to relinquish the Turing test.

[9] I take it to be a meta-argument because you have to reject solipsism first in order to make it.

[10] I know this is a paraphrase of Searle, but I can't find the reference.

[11] There is an important caveat to this assertion: Results from chaos theory indicate that even simple-seeming physical processes may be sensitive to small differences in initial conditions. This is a mathematical result that is prior to any computational result. Thus, the results of a computer simulation could be expected to diverge over time from the behavior of a particular physical system even though the

and such a computer could think. Searle doesn't actually say this, but he certainly has left the door open for it.

Now, I get really perplexed, for Searle also insists that nonetheless "instantiating a program, the right program of course," would not "by itself be a sufficient condition of understanding," and that AI programs have "little to tell us about thinking, since [they] are not about machines but about programs, and *no program by itself is sufficient for thinking*," (my italics) because (2004)

> Computers are defined in terms of symbol manipulation, and symbol manipulation by itself is neither constitutive of nor sufficient for meaning. (p.101)

Every time I reread this, I am struck by its ambiguity. There is an untendentious interpretation which is

(Interpretation 1) If I take some arbitrary bunch of symbols (the alphabet, say) and move them about in some arbitrary way (manipulate them), then that is neither constitutive nor sufficient for meaning.

In other words, instantiating a random program doesn't suffice for understanding. Surely that is obvious and not what Searle intends. Alternatively, he might mean

(Interpretation 2) A particular selection of symbols manipulated in a particular way *could* constitute and be sufficient for meaning, but it wouldn't be just any set of symbols and not just any way of manipulating them.

I happen to believe that, but I don't think Searle does. If he did, why would he insist that the Chinese Room *Gedanken* experiment demonstrates conclusively that if he is the one turning the crank on the sausage machine, and he doesn't know how sausage is made, it doesn't make sausage?[12]

Finally, and I think this is what Searle intends, it might mean

(Interpretation 3) Symbol manipulation by itself *can never be* constitutive of meaning and it *can never be* sufficient for meaning.

---

simulated behavior could be made accurate to any desired degree. As a practical matter, it is impossible to predict with certainty the positions of the planets in the solar system a million years into the future, even though the system is governed by known deterministic physical principles. Nonetheless, NASA is able to compute using these principles with sufficient accuracy to guide a spacecraft from a launch from Earth to a landing within a 20 mile radius on Mars, after a voyage of over 300 million miles.

[12] "If I do not understand Chinese on the basis of implementing the right computer program, then neither does any other computer just on the basis of implementing the program, because no computer has anything that I do not have."

That is, if you want something to be constitutive of and sufficient for meaning, then you can just forget about attempting it with symbol manipulation, i.e., computation.

Now, why would Searle think this in light of the fact that, as we have seen, he effectively grants that a computer suitably programmed can think? The answer appears to be that for some reason he thinks there is a difference between symbol manipulation and what a computer does. This is just plain incorrect.

Symbol manipulation is all that computers (Turing machines and their practical progeny, von Neumann, stored program, digital computers) ever do. The symbols they manipulate are more or less trivially defined (in the simplest case: 1, 0) and so is manipulation. Everything else is implementation details.[13]

The Turing test can be seen as an effort to make an end run around solipsism, but you can't avoid solipsism except by rejecting it on faith—it's not possible to prove there are other minds than our own, and that extends to putative minds created as instantiations of computer programs. Having killed off the Turing test in all of its variants and rejected solipsism, we still have a problem. We'd still like to be able to say with confidence whether the instantiation of a particular computer program can think.

We have concluded that we cannot tell from behavior alone whether something has a mind, is sentient. Indeed, in a brief paper, Abolfazlian 1995 proves that if human beings are Turing machine equivalent, "*Human beings cannot answer the question 'Given a Human being, is he / she really a Human being?!!'*"[14] (italics and dramatic punctuation in original) Philosophers have been dealing with this for a long time, but it's kind of bad news for researchers seeking to set the world on fire. And all it really boils down to is that solipsism is self-consistent; that is, it is not possible to prove that somebody or something else is or has a mind.

So, what question have we been arguing about, anyway? If Searle was just saying that an ability to simulate human behavior doesn't prove anything about human behavior other than that the simulation simulates it, then I'd certainly grant that and we could all go home. But that's not all he's saying. He says "Instantiating a computer program is never by itself a sufficient condition of intentionality" (1980, abstract). Yes. It is not

---

[13] Searle's "explanation" of why symbol manipulation / computation can neither be constitutive of nor sufficient for understanding is incorrect anyway. His argument, reduced to a sentence is:

(S) Syntax is not semantics *and can never be*.

He is far from the only one who takes (S) for granted. Now, I grant you syntax is not semantics in a large enough number of cases to make (S) a good rule of thumb; however, it is emphatically not true in every case. I guess I can't claim, as I'd like to, that this is obvious, because (S) has been around for a long time. It's time to drive a wooden stake through its evil heart. I'm doing my part. *See* Cutrona 2005 (in prep.).

[14] Abolfazlian thinks he is defending connectionism by rigorously proving that assuming that the brain is equivalent to a Turing machine entails this and things like, "*Human beings cannot answer the question, 'Given a human being and an example of his/her behaviour, what has caused this particular behavior?'*" The problem is that connectionism, which can be simulated by a Turing machine to an arbitrary degree of accuracy (which is generally how it's done, anyway), has the same problem, so his results don't actually attack or defend anything. Nonetheless, it is always refreshing to see a set of remarkably convincing proofs of the impossibility of accomplishing certain cherished philosophical, psychological, and AI goals.

true that *every* program that can pass the Turing test is sufficient for thinking. The Zhang zombie program is just such a program; but, Searle has conceded (*vide supra*) that we can construct a *Gedanken* program that can pass the Turing test *and* is sufficient for thinking, and that's what we're about to undertake. But first a cautionary admonition.

I used a shorthand in the previous paragraph (Searle and others use it, too). I wrote of a "program that can pass the Turing test." A program *tout court* doesn't *do* anything, and neither does a computer *tout court*. A program is a sequence of instructions which, when loaded into a computer, determines the actions the computer will take when the computer runs the program, and the physical machine that results is the *instantiation* of the program. Referring to the ensemble as a program is a synecdoche motivated by the underlying fact that in the ordinary course one keeps the computer constant and varies the program.

Even the preceding description isn't precise. At the level of physics, a program is an organized physical ensemble of things (at times a collection of selectively oriented magnetic domains, as on a magnetic disk drive; at times a pattern of statistical excess or scarcity of electrons at various physical locations, as in certain integrated circuit memories). The physical ensemble that is the program, through the intermediation of other physical objects (e.g., structures that causally channel and control the movement of statistical excesses or scarcities of electrons, as in a computer's central processing unit chip) governs the physical behavior of still other physical ensembles that serve as transducers between what is outside the computer and what is inside the computer. The whole shebang is the instantiation of a computer program: viz. the physical machine that results when a device that can execute a program does so.

Additional caveat: Human beings are clearly special-purpose machines. As such they do not analogize easily to general-purpose, digital computers.[15] Computers (general-purpose, digital to be understood from now on) make a clear distinction between the program and the device that executes it. That doesn't happen in human beings. Even granting that we are deterministic machines, it's still difficult to say with confidence what the program is that we are running or even what constitutes it. Eric Baum suggests that the only program human beings have is encoded in our DNA, and everything else is the instantiation of that program. I take it that he's literally correct and that the insight is significant, but I don't think we have to step that far back to see the forest.

Writing a Chinese Brain Program

As we have seen, it is possible (at least as a *Gedanken* experiment) to create a program that is input / output equivalent to the behavior of a human being without in any way involving processes anyone would want to characterize as thoughtful. But, inputs and

[15] Searle 1992 makes the same point, or at least gives the impression of making the same point, which is good enough for me.

outputs are not the only place we can start from when looking for a way to describe human behavior in computational terms.

Between inputs and outputs there is some (presumably well-behaved) transformation function. What we would like to find is a way to describe the transformation function exhibited by typical human beings. But, unless we articulate additional constraints (on the implementation of the transformation function), we will remain with the problem that a gigantic table look-up decision tree can (literally) mindlessly reproduce an arbitrary transformation function and satisfy an arbitrary input / output relationship. It is not sufficient to be able to duplicate *what* human beings do, it is necessary to duplicate *how* they do it—particularly if what we want to do eventually is create a program that is not simply a duplicate of a specific individual.

So let us write a program that simulates Zhang at the level of microbiology, or molecular chemistry if necessary. After all, the human brain is a finite physical device that operates according to knowable physical laws. A computer can simulate an arbitrary process to any desired degree of accuracy. So simulating a human being is just a special case. To make things work, we need to select the level at which to implement our simulation. It's a *Gedanken* experiment, so I don't really care what level we choose so long as the program models Zhang in every microbiological and molecular particular to whatever level of detail is necessary for Searle to stipulate that the program simulates "a machine with a nervous system … sufficiently like ours"[16] and it exhibits the same transformation function that Zhang does (and consequently that the Zhang zombie does).

The result is a program that, when instantiated, not only passes the Turing test, but *it does it as Zhang would have done it*, not in terms of rigid, pre-stored input / output linkages, but in terms of the thoughts and feelings Zhang would have had. This is because the program has been constructed in such a way that there exists an isomorphism between it and Zhang such that we could, in principle, replace any portion of Zhang's brain with the corresponding simulated portion and Zhang's behavior would not change.[17] The instantiation of this program has a mind, thinks, has intentionality, and is undeniably a person. In short, instantiate a program that simulates a normal brain in sufficient detail, and you get a machine that has a mind.[18]

---

[16] Recall that this is Searle's declared requirement for a machine that can "think, understand, and so on."

[17] Obviously, we have to respect the level of our simulation in the replacement. If we are simulating at the level of neurons, we have to replace whole neurons because the isomorphism only extends to the level we had John Searle agree would suffice when we constructed the program.

[18] Assuming (as I do) that brain processes are mathematically chaotic as described in a previous footnote, our Zhang brain simulation might not behave *exactly* like its prototype (Zhang's brain) over its lifetime. So it's not actually input / output equivalent. Presumably, where the two might diverge is at some point where a small difference between a modeled process and the actual process has a macrobehavioral consequence. Such a difference might arise if the modeled diffusion of potassium ions at a synapse were to differ in the exact arrival time of a particular potassium ion such that arrival just before or after a particular moment in time makes the difference between the firing of the post-synaptic neuron and not. And further, the firing of that neuron may make the difference between the firing or not of a larger group of active brain elements with which that neuron synapses, and so on to the presence or absence of some macrobehavioral difference.
The point is that Zhang could have behaved either way and still have been the same, thinking, understanding Zhang. The Zhang brain simulation under these circumstances turns out to have a mind of

Now, the two *Gedanken* programs we have constructed surely bracket the ultimate target, viz. an actual (not *Gedanken*) program that can pass the Turing test, has a mind, and can think. We know that the Zhang zombie program by construction does not have a mind and cannot think. We know that the Zhang brain program by construction has a mind and can and does think. We seek some middle ground.

The Zhang brain program is only one of a large number of thinking brain programs that could be constructed (one for every person that is, ever was, will be, or could be, for that matter). Presumably, there are commonalities among these (having to do with brain structure and function, etc.), and these commonalities can be exploited to simplify the program by reducing complex processes to functionally equivalent processes that may be more easily understood.

Accordingly, we may phrase the central question posed by cognitive neuroscience as: what is the highest (most simplified) level at which we can still feel confident or comfortable that a program based on the Zhang brain program has a mind? That is, can mind or even some portion of it be constituted at a level higher than, say, the level of individual neurons (or whatever level we agree it is unambiguously constituted at)? To what extent is it the case that the actual input / output transformation functions of various classes of active brain elements is, up to a point, irrelevant and that active elements with input / output transformation functions more amenable to computer calculation will serve as well? This is a non-trivial question, considering that recent research clearly indicates that simulations of circuits comprising interconnected spiking neurons exhibit vastly richer behavior than earlier continuous-response neuron models. And we have yet to see elaborate spiking models whose behavior reflects multiple types of post-synaptic receptors and neurotransmitter substances. Clearly, our models will have to become even more complex before we can learn how to simplify them appropriately.

Come Out of that Chinese Room with Your Hands Up!

From the construction of the Zhang brain program we know that whether or not John Searle understands Chinese, the instantiation of the Zhang brain program understands Chinese. Even if it is John Searle who is doing all of the calculations necessary to simulate Zhang's brain, and even if he is doing them all in his head, there is no reason to believe that his knowledge or ignorance of anything other than the way to perform the instantiation of the Zhang brain program, has even the smallest bearing on whether that instantiation has a mind and understands Chinese.

---

its own, but it is still a mind. It is difficult to assess (particularly in the analysis of a *Gedanken* program) how susceptible macrobehavior is to this kind of situation. I tend to think such situations are rare—as Freud put it, all behavior is overdetermined. Nonetheless, the argument I am making is that the simulation is implemented at a level of detail where differences of this kind do not affect the functioning of the system qua brain in such a way as to destroy it's acceptability as a normal functioning brain, even if the observed macrobehavior of the system in which the simulated brain operates diverges from the observed macrobehavior of the person whose brain served as the prototype.

I'll say it again. John Searle's intuition that because he does not know Chinese, no program he can possibly instantiate can know Chinese is just plain incorrect. *Punkt*.

So, I'll make a deal, John. I'll agree that

(1) It is not the case that not just any old program that you can instantiate in the Chinese room and that passes the Turing test in any form you like can be said to know Chinese.

That is, the ability to pass the Turing test in Chinese is a necessary but not sufficient condition. And you'll agree that

(2) Your instantiation (or anything else's timely instantiation) of the Zhang brain program has a mind and knows Chinese.

With respect to the question of whether the instantiation of a specific program or class of programs has a mind, you'll agree that

(3) Solipsistic arguments about computations being in the eye of the beholder do not serve a useful purpose in this context and should be abandoned.

Finally, in light of your statement that just as "we obtain support for the surmise that others are conscious from the knowledge that they have a similar biological makeup to our own," you'll agree (or reaffirm) that

(4) The proper basis for assessing the putative consciousness of an instantiation of a computer program is whether, in a meaningful way, at an appropriate level of discussion, it has similar (pseudo) biological makeup to our own.

There is surely enough there to enable generations of philosophers and theoretical cognitive neuroscientists to be gainfully employed for generations to come.

It seems clear that this indicates the direction we have to take in a post Turing test world: Consciousness requires the implementation of particular algorithms and data structures. Our task is to figure out how to characterize them. One would have thought that the demise of pure behaviorism would have taught us this long since.

References

Abolfazlian, A. 1995. "What's Connectionism got to do with IT?" *Proceedings of the Second Swedish Conference on Connectionism (SCC95)*.

Baum, Eric B. 2004. *What is Thought?* The MIT Press: Cambridge, Massachusetts.

Cutrona, Louis J., Jr. 2005. "Intentionality: When Syntax Is Semantics" (in prep.)

Dennett, Daniel C. 1991. *Consciousness Explained*. Boston, Toronto, London: Little, Brown.

Fodor, Jerry A. 1975. *The Language of Thought*. Harvard University Press: Cambridge, Massachusetts.

Gettier, Edmund L., III. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23: 121-123.

Hawkins, Jeff with Blakeslee, Sandra. 2004. *On Intelligence*. New York: Times Books, Henry Holt and Company.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: 417-457.

- - - 1992. *The Rediscovery of the Mind*. The MIT Press: Cambridge, Mass.

- - - 2004. *Mind: A Brief Introduction*. Oxford University Press: New York.

Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind* 59: 433-460.