Computationalism*

Stuart C. Shapiro
Department of Computer Science
and Center for Cognitive Science
State University of New York at Buffalo
226 Bell Hall
Buffalo, NY 14260-2000
U.S.A
shapiro@cs.buffalo.edu

March 9, 1995

Abstract

Computationalism, the notion that cognition is computation, is a working hypothesis of many AI researchers and Cognitive Scientists. Although it has not been proved, neither has it been disproved. In this paper, I give some refutations to some well-known alleged refutations of computationalism. My arguments have two themes: people are more limited than is often recognized in these debates; computer systems are more complicated than is often recognized in these debates. To underline the latter point, I sketch the design and abilities of a possible embodied computer system.

1 Artificial Intelligence and Computationalism

There are several disparate goals pursued by Artificial Intelligence (AI) researchers: computational psychology, computational philosophy, and advanced Computer Science [Shapiro, 1992]. In this paper, I will concentrate on computational philosophy, which could also be called "the computational study of cognition." AI is often thought of in the popular press as a technology. The term "AI program" or "AI technique" is often used. More properly, however, AI is the name of a discipline—the scientific discipline devoted to the investigation of whether cognition may adequately be modeled computationally. Thus, the computational study of cognition—the computational study of how to produce behaviors which we are willing to call intelligent. Intelligence is not something

*This is a preliminary version of Stuart C. Shapiro, Computationalism, *Minds and Machines, 5* 4 (November, 1995), 517–524. All quotes should be from, and all citations should be to the published version.

predefined for purposes of the study, but will be defined as a result of the study. Another way of describing the study is as an attempt to answer the question, "Is intelligence a computable function?" We know that there functions that are not computable [Biermann, 1990], and we know lots of functions that we can compute. There is much in between, including, at the present time, cognition. AI researchers are investigating how much of cognition can be moved from the inbetween class into the known-computational class. They do this by building programs, testing them, and seeing where they succeed and where they fail. Thus, the stance that intelligence is computation—that cognition is computation—can be seen as a working hypothesis of the AI researcher. We assume it as a working hypothesis, and proceed to investigate its boundaries. It always amazes me that some people undertake to claim *a priori* that it's impossible for cognition to be computation. The hubris, it seems to me, is borne by those people who claim that it's already known, or already clear that cognition is not computation, not by the AI researcher who is trying to find out.

Like AI researchers, the goal of Cognitive Scientists is also to understand cognition, with more stress on *human* cognition. The debate on whether cognition is computation, or cognition is appropriately modeled by computation, is, in Cognitive Science, often phrased as whether the mental level is an appropriate level at which to model human cognitive behavior.

2 Refutations to "Refutations"

In this section, I will give some refutations to some alleged refutiations to computationalism.

2.1 Lucas's Argument

In 1931, Gödel presented his now famous Incompleteness Theorem [Gödel, 1931], which essentially says that any formal system powerful enough to represent arithmetic is either inconsistent or incomplete. That is, if a powerful enough formal system is consistent, there is a statement, expressible in the system that is true, but that cannot be proved within the formal system. This theorem has been used repeatedly [Lucas, 1961, Penrose, 1989] to argue that the human mind could not be a formal system, *i.e.,* that cognition could not be computation. The argument (sometimes, and henceforth in this paper, called "Lucas's Argument") essentially goes as follows: Given any formal system, we (human minds) can know that there is a sentence in that system that is true, yet can't be proved in that system. Therefore the human mind is more powerful than any formal system. Therefore, the human mind cannot be duplicated by any formal system.

The first problem with this argument is that Gödel's theorem says that there is a statement, G, expressible in a given formal system $\mathcal{F}$ that is true, but that cannot be proved within $\mathcal{F}$. It does not say that there cannot be another formal system, $\mathcal{F}'$, in which it can be be proved that G is true but not provable in

$\mathcal{F}$. For example, let us sketch such a formal system. We will use a version of Fitch-style Natural Deduction [Fitch, 1952], with its usual syntax, semantics, and rules of inference, augmented with the following:

$\mathcal{F}$: Some formal system powerful enough to represent arithmetic.

G: The Gödel statement expressed in $\mathcal{F}$.

***Provable*:** A unary predicate, whose interpretation is the set of wffs expressable in and provable in $\mathcal{F}$.

***True*:** A unary predicate, whose interpretation is the set of wffs expressable in and true in $\mathcal{F}$.

Axiom 1: $G = \neg \mathit{Provable}(G)$. This equality, which shows what G says can actually be derived, but we will take it in this sketch to be an axiom.

Axiom 2: $\forall p(\mathit{Provable}(p) \Rightarrow \mathit{True}(p))$. This is a version of the assumption that $\mathcal{F}$ is consistent.

Axiom 3: $\forall p(p \Leftrightarrow \mathit{True}(p))$. This just says that asserting something is the same as asserting that it is true.

$= E$: The rule of inference, $\mathcal{A}(\mathcal{P}), (\mathcal{P} = \mathcal{Q}) \vdash \mathcal{A}(\mathcal{Q})$, of substitutability of equals for equals.

The proof, in this formal system that G is true but unprovable is:

$$\begin{array}{rl|lll}
1. & & \mathit{Provable}(G) & \mathrm{H}yp & \\ \cline{2-4}
2. & & \forall p(\mathit{Provable}(p) \Rightarrow \mathit{True}(p)) & \mathrm{A}xiom2 & \\
3. & & \mathit{Provable}(G) \Rightarrow \mathit{True}(G) & 2, \forall E & \\
4. & & \mathit{True}(G) & 1, 3, \Rightarrow E & \\
5. & & G = \neg \mathit{Provable}(G) & \mathrm{Axiom1} & \\
6. & & \mathit{True}(\neg \mathit{Provable}(G)) & 4, 5, = E & \\
7. & & \forall p(p \Leftrightarrow \mathit{True}(p)) & \mathrm{Axiom3} & \\
8. & & \neg \mathit{Provable}(G) \Leftrightarrow \mathit{True}(\neg \mathit{Provable}(G)) & 7, \forall E & \\
9. & & \neg \mathit{Provable}(G) & 6, 8, \Leftrightarrow E & \\
\multicolumn{1}{r}{10.} & \multicolumn{3}{l}{\neg \mathit{Provable}(G)} & 1, 9, \neg I \\
\multicolumn{1}{r}{11.} & \multicolumn{3}{l}{G = \neg \mathit{Provable}(G)} & \mathrm{Axiom1} \\
\multicolumn{1}{r}{12.} & \multicolumn{3}{l}{G} & 10, 11, = E \\
\multicolumn{1}{r}{13.} & \multicolumn{3}{l}{\forall p(p \Leftrightarrow \mathit{True}(p))} & \mathrm{Axiom3} \\
\multicolumn{1}{r}{14.} & \multicolumn{3}{l}{G \Leftrightarrow \mathit{True}(G)} & 13, \forall E \\
\multicolumn{1}{r}{15.} & \multicolumn{3}{l}{\mathit{True}(G)} & 12, 14, \Leftrightarrow E \\
\multicolumn{1}{r}{16.} & \multicolumn{3}{l}{\mathit{True}(G) \wedge \neg \mathit{Provable}(G)} & 10, 15, \wedge I
\end{array}$$

So in this formal system, we have proved that G is true but unprovable in $\mathcal{F}$ as long as $\mathcal{F}$ is consistent. Thus, it is perfectly consistent with Gödel's theorem for human minds to be formal systems within which it can be proved that there are other formal systems that are incomplete.

If Gödel's theorem cannot be used to show that minds are not formal systems on the basis that minds can do something formal systems cannot, *viz.,* show the incompleteness of formal systems, perhaps Gödel's theorem still shows that minds cannot be formal systems because formal systems are limited (by being incomplete) in ways that minds are not. This objection was already answered by [Turing, 1950]:

> The short answer to this argument is that although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect . . . We too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines. [Feigenbaum and Feldman, 1963, p. 22]

That is, people too are incomplete—there are truths they do not know. In fact, people are also inconsistent—they make mistakes.

To illustrate this point more concretely, I have formulated a Gödel statement for people. You should think about this, and try to decide for yourself, whether or not it is true:

I CANNOT KNOW THAT THIS STATEMENT IS TRUE.

Notice that this is the same kind of modification of the Liar's Paradox ("This statement is false.") that the Gödel statement ("This statement is unprovable.") is. If you believe that the statement is true, well then, it says that you can't know it, so it's a true statement that you can't know is true. If you believe it is false, then that means that you *do* know that it's true. But then it's a false statement that you know is true, and by the definition of knowledge as justified *true* belief, you can't *know* something that is false. So it can't be false because then you would know something that's false. So it must be true. Now, if you come to that conclusion, then you might say that, well, you do know that it's true, but of course that leads to an inconsistency. So the only way to know that this is true is to be inconsistent and the Gödel theorem is that consistency implies incompleteness—you're either inconsistent or incomplete. Now, just as with the Gödel statement, it's possible to prove in one formal system that something is true but not provable in another formal system. So you can realize about someone else that "I know that the statement, when 'I' means you, is true, but you can't know it's true."

There have been other "Gödel sentences for humans" suggested in the literature. Casti [Casti, 1989] suggests "Lucas cannot consistently assert this sentence." However, this sentence could only work for Lucas, and, anyway, someone can refrain from asserting a sentence they know to be true. Smullyan [Smullyan, 1978, Smullyan, 1986] gives many Gödel-type sentences, but they all involve Knights (who never lie), Knaves (who never tell the truth), and logicians (who, at least always reason logically), so it might be felt that these sentences have nothing to do with the normal human mind.

There are, surely, other truths that we can't know are true: how the universe began; the solution to chess; even, how we work at the cognitive level. I believe we could have a model of our own cognition that always seems to give correct predictions, but that we could never *know* whether it was, in fact, accurate.

2.2 Searle's Chinese Room

Another famous alleged refutation to computationalism is Searle's Chinese Room argument [Searle, 1980]. Briefly, the argument is as follows. Assume that I (Searle) am in a room, and people pass cards containing Chinese messages to me through a slot in the door. I do not understand Chinese, but there is a big instruction book, written in English, which I do understand, in the room with me. Using this book, and looking at the incomprehensible Chinese message, I arrive at an equally incomprehensible Chinese message which I write on another card, which I pass back through the door slot. Assume further that whenever this happens, the message I pass out is, to the people outside, a perfectly good Chinese response to the message they had passed in. I still don't understand Chinese, and neither does the book, the room, or the collection of book, room, and me. The symbol manipulation instructions in the book are enough to let me generate the answering messages, but they are not enough to generate understanding, or any other cognitive state.

This argument is adequately refuted in [Rapaport, 1988] and elsewhere. I just want to add that I think one reason people find Searle's argument so convincing is that they imagine themselves in the room manipulating the cards, and they know that they don't understand Chinese. One thing that's easy to miss is what a hard job the guy in the room has. Those of us who have done any programming whatsoever know it's very easy to write a program that's virtually impossible to trace. Programs get their power from doing a lot of work based on lots of local decisions that programmers tediously work out. But a program quickly gets so complicated that not even the programmer can trace it through in any reasonable amount of time to come up with the same answer the program does. All of us who have written programs have had the experience of our programs surprising us, producing results that we didn't predict.

You can't take the fact that a person cannot reliably carry out the instructions contained in a long program as a reason that computationalism must be wrong, because you probably also can't trace all your neural connections. The point is that the ability of the guy in the room to manipulate the cards so that the people outside percieve a competent Chinese discussion is an assumption of Searle's thought experiment, but it is a poor thought experiment, because you are invited to imagine yourself doing something that you just couldn't do.

2.3 Embodiedness

Another popular alleged refutation to computationalism is the assumption that cognition requires embodiedness, and the further assumption that computers are not embodied. I will discuss the second part of this in the next section, but for the first part, it's worthwhile thinking about what your opinion of handicapped people is. When does a person lose enough of his or her abilities, mental or physical, that you are willing to say that that person no longer has cognitive states? If someone is blind, and can no longer see the world; if they're invalided, and are no longer able to have interactions with the world; if they have severe cerebral palsy, and no longer have connections through their body with the outside world; if they have mental illnesses, loss of affect. Take every objection to computers having cognitive states because of deficits in embodiedness, and see if there are, in fact, people who have similar problems, and judge whether or not you would be willing to say that, therefore, that person does not have cognitive states.

3 A Sketch of a Computational Cognitive Agent

In this section, I will sketch a computational model of cognition, a possible computer system that would count as a cognitive agent. Everything I'm going to mention either has been done, to at least some extent, or is clearly on the horizon, either in my own laboratory or in other AI laboratories. My purpose is to suggest a more complicated view of computer programs than you might have in mind.

This computer agent will be a robot with a body, legs for walking around, arms and hands with shoulders, elbows, wrists, and fingers for manipulating things, and a head mounted on a flexible neck, which will be mounted on the body. On the head will be two cameras for eyes.

Each camera eye will have a fovea supporting high-resolution central vision, and a periphery supporting lower-resolution peripheral vision. The two eyes will be used together to provide stereo vision, and the robot will be able to move the head around to get a better view of things.

The robot will have microphones for the input of human speech and other sounds. These will be mounted on the head so that the multiple microphones, supported by head movement, can be used for localization of the sources of sounds.

There will be touch sensors distributed around the robot's body so it can detect when it is in contact with things. These will be especially dense on the fingers and hands, so it can tell when it is holding something. There may be additional sensors, such as infra-red receivers and/or radar.

Each joint will contain sensors so the robot will be able to determine the position of its limbs. It will also have sensors for various internal needs such as the level of its batteries and the lubrication of its joints.

The robot will have hand-eye coordination. For example, it will be able to focus its eyes on a spot within its reach, and then put a hand there. This will be used for catching, grasping, and manipulating objects. It will also have body-eye coordination. For example, to follow something, it will move its head and eyes to focus on it, move its body to orient it with its head, and then walk so as to keep the object in front of it and at the same distance from it.

People will be able to input information to the robot using a reasonably large subset of English, and will be able to command it to do things with the same language. They will also be able to explain new words and phrases to it, and to explain how to do new tasks.

The robot will be able to output information, via a speaker, using a subset of its input language.

The robot will be able to store information about a wide variety of objects and people, real, imaginary, and fictional, and reason with and about that information. In particular, the robot will be able to store information about the people it interacts with, and use that information when generating its output. For example, it would formulate referring expressions using what it has stored about the beliefs of its adressee about the referent.

The robot will also have information stored about itself, including what it is doing and what it has done and said. When generating English information about itself, it will refer to itself using the first person pronoun. So it might say "My batteries are low," "My right hand needs more lubrication," or "I am getting a hammer for John." Since it stores what it said and did, it can use this information when referring to things. For example, it might refer to "the wrench I gave you yesterday."

We could explicitly give the robot rules about how to behave and it will be able to use these rules, discuss them, talk about them, and use them to actually act. The robot will have goals, in the sense of things to do. It will be able to reason about what to do next, so one could give it rules about competing goals, and what's more important to do when. It might have a goal and have some ideas of how to accomplish it, but it might keep putting it off for more important goals. Moreover, it will be able to discuss those decisions with you. Since it can remember what it did, and what it thought about, you could discuss with it what it did yesterday or why it did what it did. Since it will remember what it did and remember what it thought about why it did it, it could perhaps, later get new information that would have been useful in deciding what to do yesterday if it only had it in time, and so it could talk about what it would have done, had it known better.

4 Conclusions

Computationalism, the notion that cognition is computation, is a working hypothesis of many AI researchers and Cognitive Scientists. Although it has not been proved, neither has it been disproved. In this paper, I gave some refutations to some well-known alleged refutations of computationalism: Lucas's Argument; Searle's Chinese Room argument; and the general requirement of embodiedness. My arguments had two themes. First, people are more limited than is often recognized in these debates. In particular, they are subject to the limitations that all computer systems are. Second, computer systems are more complicated than is often recognized in these debates. To underline this point, I sketched the design and abilities of a computer robot that is embodied, senses its body and its internal states as well as objects in its environment, and has beliefs about itself as well as other agents and other objects.

To conclude, I see the process that AI and Cognitive Science are engaged in as another chapter in humanity learning its place in the universe. In astronomy, the sun centered solar system was resisted for a long time because humanity felt that it was in the center of the universe. Eventually we accepted that our place was just a corner of the universe. Most of the time we don't feel too bad about that knowledge. Evolution showed that we weren't special in terms of our place in the biological world. People resisted for a long time the notion that they evolved from other creatures, but eventually most accepted it, and don't feel too bad about that either. I think that we are now in the process of learning about our role in the cognitive world—that cognition is not some special thing that arises from our being humans, or even from being biological organisms, but is a natural part of any complicated enough information processing system. I think a lot of the resistance to this can be seen to be of the same class as the previous two resistances. But increased knowledge in this area will, I believe, be just as beneficial as it was in astronomy and biology.

References

[Anderson, 1964] Anderson, A. R., editor (1964). *Minds and Machines.* Prentice Hall, Englewood Cliffs, NJ.

[Biermann, 1990] Biermann, A. W. (1990). Noncomputability. In *Great Ideas in Computer Science*, chapter 13, pages 351–373. MIT Press, Cambridge, MA.

[Casti, 1989] Casti, J. L. (1989). *Paradigms Lost.* William Morrow, New York.

[Feigenbaum and Feldman, 1963] Feigenbaum, E. A. and Feldman, J. (1963). *Computers and Thought.* McGraw-Hill Book Company, New York.

[Fitch, 1952] Fitch, F. B. (1952). *Symbolic Logic: An Introduction.* Ronald Press, New York.

[Gödel, 1931] Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38:173–198. Reprinted in [Anderson, 1964].

[Lucas, 1961] Lucas, J. R. (1961). Minds, machines and Gödel. *Philosophy*, 36:120–124.

[Penrose, 1989] Penrose, R. (1989). *The Emporer's New Mind.* Oxford University Press, New York.

[Rapaport, 1988] Rapaport, W. J. (1988). Syntactic semantics: Foundations of computational natural-language understanding. In Fetzer, J. H., editor, *Aspects of Artificial Intelligence*, pages 81–131. Kluwer, Holland.

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

[Shapiro, 1992] Shapiro, S. C. (1992). Artificial intelligence. In Shapiro, S. C., editor, *Encyclopedia of Artificial Intelligence*, pages 54–57. John Wiley & Sons, New York, second edition.

[Smullyan, 1978] Smullyan, R. M. (1978). *What is the Name of This Book?* Prentice Hall, Englewood Cliffs, NJ.

[Smullyan, 1986] Smullyan, R. M. (1986). Logicians who reason about themselves. In Halpern, J. Y., editor, *Theoretical Aspects of Reasoning about Knowledge*, pages 341–352. Morgan Kaufmann, Los Altos, CA.

[Turing, 1950] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59:433–460. Reprinted in [Feigenbaum and Feldman, 1963, pp. 11–35].