

## Searle's Chinese Room and its Aftermath

Vincent John Mooney III

Master of Arts, Department of Philosophy, Symbolic Systems Program

Ph.D. candidate, Department of Electrical Engineering, Computer Systems Laboratory

Stanford University, Stanford, CA 94305

June 15, 1997

### Abstract

*This paper attempts a systematic survey of some of the issues raised by Searle and his critics, while including new arguments on the application of the Chinese Room Argument to causal reductionism and showing implausible results of criticisms of Searle, such as consciousness blinking in and out at rapid speed. Also included is Terry Winograd's unpublished reply to Searle's original argument in the Behavioral and Brain Sciences.*

*Two main points recur often in the debate. On the one hand, Searle and his supporters argue that the conclusion that beer cans and strings have intentionality, just by virtue of being appropriately connected, is absurd. On the other hand, critics argue, "What else could be the case?"*

*Clearly, the goal of some Artificial Intelligence researchers is to fully copy human intelligence and wisdom in a computer program or set of computer programs. The feasibility of such a goal is put into deep question as John Searle's argument unfolds in the literature, especially as regards the lack of demonstrated ability of syntactic manipulations to lead to semantics. The current state of the debate seems to center on this issue – can syntax cause semantics – and the status of Searle's claimed absurd beer cans and strings conclusion.*

Note: the picture on the preceding screen is taken from *Abacus* magazine as it appears in [76].

Copyright ©1997 Vincent Mooney

# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduction</b>                                      | <b>1</b>  |
| 1.1      | Organization of Sections . . . . .                       | 1         |
| <b>2</b> | <b>The Claims of Artificial Intelligence</b>             | <b>1</b>  |
| 2.1      | The Turing Test . . . . .                                | 1         |
| 2.2      | The Physical Symbol System Hypothesis . . . . .          | 2         |
| 2.3      | What is functionalism? . . . . .                         | 4         |
| <b>3</b> | <b>The Original Chinese Room</b>                         | <b>5</b>  |
| <b>4</b> | <b>Computation</b>                                       | <b>7</b>  |
| <b>5</b> | <b>The System Reply</b>                                  | <b>8</b>  |
| <b>6</b> | <b>The Robot System Reply</b>                            | <b>9</b>  |
| 6.1      | Simulation and Implementation . . . . .                  | 10        |
| 6.2      | Brain modeling . . . . .                                 | 11        |
| 6.3      | Syntax versus Semantics . . . . .                        | 11        |
| 6.4      | Searle’s Reply . . . . .                                 | 13        |
| 6.5      | A Problem with Robotic Functionalism . . . . .           | 14        |
| <b>7</b> | <b>Dennett’s Objection</b>                               | <b>15</b> |
| 7.1      | Searle’s reply . . . . .                                 | 15        |
| 7.2      | Hofstadter and Dennett’s Further Objection . . . . .     | 16        |
| 7.2.1    | If speed is the issue, give a particular bound . . . . . | 16        |
| 7.2.2    | Silicon brain replacement . . . . .                      | 17        |
| 7.3      | An Exchange . . . . .                                    | 18        |
| <b>8</b> | <b>Combined Robot and System Reply</b>                   | <b>19</b> |
| 8.1      | What is “functional equivalence”? . . . . .              | 19        |
| 8.2      | Connectionism . . . . .                                  | 20        |
| 8.3      | Commentary . . . . .                                     | 21        |
| <b>9</b> | <b>The Robot Reply Refuted</b>                           | <b>23</b> |
| 9.1      | Tea-Carrying . . . . .                                   | 23        |
| 9.2      | He really does understand . . . . .                      | 23        |
| 9.3      | Someone else understands . . . . .                       | 24        |

|  |           |
|--|-----------|
| <b>10 Brain simulation</b>   | <b>25</b> |
| 10.1 Simulation can be Real . . . . .                                      | 25        |
| 10.2 Summary and Status . . . . .  | 26        |
| <b>11 Searle versus Churchland and Churchland</b>                          | <b>26</b> |
| 11.1 Searle’s axiomatized argument . . . . .                               | 26        |
| 11.2 Syntax Can Generate Semantics . . . . .                               | 28        |
| 11.2.1 Two Comments . . . . .  | 29        |
| 11.3 Searle’s Response . . . . .   | 29        |
| <b>12 Thinking Machines and Virtual Persons</b>                            | <b>29</b> |
| 12.1 Causal Powers of CPUs . . . . .                                       | 30        |
| 12.1.1 Commentary . . . . .  | 30        |
| 12.2 Intentionality and Computationalism . . . . .                         | 31        |
| 12.2.1 Result of the Chinese Room: A Strategy on the Turing Test . . . . . | 31        |
| <b>13 Nonreductive Functionalism</b>                                       | <b>32</b> |
| 13.1 Fading Qualia . . . . .   | 32        |
| 13.2 Panpsychism . . . . .   | 33        |
| <b>14 Some New Arguments</b>   | <b>34</b> |
| 14.1 Computer plus chemical change . . . . .                               | 34        |
| 14.1.1 Another problem . . . . .   | 35        |
| <b>15 The Reaction of AI: Why Does This Matter?</b>                        | <b>36</b> |
| 15.1 Reasons to pay attention . . . . .                                    | 37        |
| 15.2 Winograd and Flores . . . . .   | 38        |
| 15.2.1 Comments on Winograd . . . . .                                      | 39        |
| 15.3 What are the goals of AI? . . . . .                                   | 40        |
| 15.4 Strong AI as a Degenerating Research Program? . . . . .               | 41        |
| 15.5 Logical Positivism . . . . .  | 42        |
| <b>16 Conclusion</b>   | <b>43</b> |
| 16.1 Status . . . . .  | 43        |
| 16.2 Philosophy and Science . . . . .                                      | 44        |
| 16.3 Closing comments . . . . .  | 44        |
| <b>A Appendix: Terry Winograd’s unpublished reply to Searle 1980a</b>      | <b>46</b> |



# 1 Introduction

This paper will examine a few threads in the fascinating and many-faceted twists and turns in philosophical arguments surrounding John Searle’s famous Chinese Room Argument. With this argument, Searle has succeeded in generating almost two decades of controversy in the philosophical and scientific literature surrounding Artificial Intelligence (AI).

Since the issues addressed here cross the boundaries of philosophy and computer science, I will proceed in a very careful way, defining the terms and principles used in subsequent discussions and arguments. In this way, I hope to clarify some of the issues raised by Searle and make them more accessible to those outside of philosophy. I also believe that this will clarify Searle’s claims, making them more amenable to exact analysis.

## 1.1 Organization of Sections

The paper is organized as follows. Section 2 briefly examines the claims of actual AI researchers. Section 3 gives the original Chinese room argument as advanced by Searle. Section 4 gives a brief definition of computation, with appropriate references. From Section 5 on out three different levels of writing occur. The first is an exposition of particular lines of criticism against Searle. We try to summarize these and number each one as a distinct **Objection**. The second level occurs with Searle’s replies, each of which we also try to concisely enumerate as a separate **Reply**. (In a few cases, another author also come up with a **Reply**.) Finally, the third level is when I give my own arguments about the issue in question, usually in a separate subsection.

## 2 The Claims of Artificial Intelligence

Before entering the Chinese Room debate, it is appropriate to examine, in their own words, the claims of AI researchers regarding what might be called the “mental capacities” of their programs (note that some AI researchers do not even believe that *we* have true “mental capacities”; nonetheless, they still have to give an account of what in fact goes on in the Chinese room, if they want to refute Searle’s argument). This will help set the stage for Searle’s thought experiment and later debate.

### 2.1 The Turing Test

Alan Turing proposed the famous “Turing Test” in which a human interrogator interviews two others via teletype to try to determine which is a computer and which is a human (Turing 1950)[149]. True to his scientific inclinations, Turing made a prediction regarding this “imitation game”:

I believe that in about fifty years' time it will be possible to program computers, with a storage capacity of about  $10^9$ , to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning. The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted (Turing 1950, p. 57)[149].

One small comment: when Turing says that the question of whether or not machines can think is "too meaningless to deserve discussion," he is asserting a philosophical, not scientific, position. He has nowhere proven scientifically that this question is meaningless – at least, I am not aware of any such claimed proof. Without commenting on the accuracy of his statement, I think it is important nonetheless to note its non-scientific and non-empirical character.

## 2.2 The Physical Symbol System Hypothesis

In 1958 Allen Newell and Herbert Simon stated that "Intuition, insight, and learning are no longer exclusive possessions of human: any large high-speed computer can be programmed to exhibit them also" (Newell and Simon 1958)[129]. About their General Problem Solver, Newell and Simon assert that it is a "program that simulates human thought" (Newell and Simon 1961)[130]. In 1965 Herbert Simon said "Machines will be capable, within 20 years, of doing any work that a man can do." (Simon 1965, p. 96)[144]

In 1975 Allen Newell and Herbert Simon received the prestigious ACM Turing Award for their work in AI. In their award lecture, they outlined various monumental scientific theories of the past, such as the cell doctrine in biology, plate tectonics in geology, and the germ theory of disease; clearly, they mean to place their own theory as potentially on equal footing to these previous breakthroughs (Newell and Simon 1976, p. 115)[131]. They define a *physical symbol system* as having two important features: (1) Such systems obey the laws of physics, and thus are realizable by engineered systems; and (2) the term "symbol" is not restricted to human symbol systems. Then they introduced their own hypothesis, hoping to contribute to the advancement of science, as follows:

*The Physical Symbol System Hypothesis.* A physical symbol system has the necessary and sufficient means for general intelligent action (Newell and Simon 1976, p. 116)[131].

A few pages later they identify a physical symbol system as an instance of a universal machine. The model of a Turing machine shows the limits of what can be computed by a

universal machine (Newell and Simon 1976, p. 117)[131]. This identification is also explicit in Newell's 1980 article where he declares the following: "*Symbol Systems* are the same as *universal machines*" (Newell 1980, p. 154)[132].

This hypothesis, they claim, "represents an important discovery of computer science, which if borne out by the empirical evidence, as in fact appears to be occurring, will have major continuing impact on the field" (Newell and Simon 1976, p. 120)[131]. And the connection between this hypothesis and the human mind is spelled out by Simon in a book he published 5 years later:

The computer is a member of an important family of artifacts called symbol systems, or more explicitly, physical symbol systems. Another important member of the family (some of us think, anthropomorphically, it is the *most* important) is the human mind and brain (Simon 1981, p. 27)[145].

Clearly then, the human mind is a physical symbol system. That is the hypothesis. Marvin Minsky also has the same idea. In his book, *The Society of Mind*, he says:

Most people still believe that no machine could ever be conscious, or feel ambition, jealousy, humor, or have any other mental life-experience. To be sure, we are still far from being able to create machines that do all the things people do. But this only means that we need better theories about how thinking works. This book will show how the tiny machines that we'll call "agents of the mind" could be the long sought "particles" that those theories need (Minsky 1986, p. 19)[127].

This kind of AI, which seeks to identify the mind with "agents" which perform symbolic manipulations via programs, is dubbed "Strong AI" by Searle (Searle 1980a)[86]. Such AI programs can, according to many of their proponents, literally be said to understand. In the introduction to the script approach taken by Schank and Abelson, they claim the following for their programs:

This new stratum of conceptual entities we call the Knowledge Structure (KS) level. It deals with human intentions, dispositions, and relationships. While it is possible computers cannot actually experience such intentions and relationships, they can perfectly well be programmed to have some understanding of their occurrence and significance, thus functioning as smart observers (Schank and Abelson 1977, p. 4)[143].

It is the evaluation of just such claims that Searle intended to address.

A more philosophical statement of this hypothesis, called *functionalism*, has been expounded by Putnam (Putnam 1967)[137]. Simon, Newell, and Putnam all claimed that it is plausible that the human mind is, in fact, implemented in virtue of computations carried out on a universal Turing machine. I will consider a more exact statement of this hypothesis in the following section.

### 2.3 What is functionalism?

Functionalism can be described conveniently using Putnam's terminology (Putnam 1967)[137]. In short, functionalism claims that a mental state is a state of a Turing machine, whose function in the Turing machine determines which mental state it implements. For example, pain is a particular Turing machine state realized in humans, octopuses, etc.

Putnam notes that any Turing machine is describable by a **Machine Table** which specifies the inputs, states, and next states for each input and state. Putnam extends the Turing machine model by adding probabilistic transitions from one state to another. Such a **Machine Table** with probabilistic transitions is called a **Probabilistic Automaton**.

Now a **Description** of system  $S$  says that  $S$  possesses distinct states  $S_1, S_2, \dots, S_n$  related to each other in the **Machine Table**. We identify the **Machine Table** in the **Description** with the **Functional Organization** of  $S$  relative to the **Description**. Finally,  $S_i$  such that  $S$  is in state  $S_i$  at a given time is the **Total State** of  $S$  at that time relative to the **Description**. Note that  $S_i$  is specified implicitly by the **Description** by the set of transition probabilities given in the **Machine Table** (which, if the probabilities are not all 0 and 1, is a **Probabilistic Automaton**). For example, pain could be functional state  $S_i$  of a whole organism.

Clearly this is compatible with other forms of functionalism (e.g. Lewis[121]), since they all posit that a mental state is a disposition to act in certain ways and to have certain mental states, given certain sensory inputs and certain mental states.

Putnam makes an important comment that his theory does not disprove other theories such as behaviorism (deny the existence of mental states) or type-type psycho-physical identity theory (a mental state like pain is exactly identifiable with a physical brain state), but rather is an empirical hypothesis that better fits the data. Also, his functionalism gives functional identity in terms of empirical psychology. In particular, this means that some of the inputs and outputs of a functionalist model can be neurons or other physical descriptions used in scientific psychology.

Given this, Putnam summarizes his proposal as follows:

- (1) All organisms capable of feeling pain are **Probabilistic Automaton**.

- (2) Every organism capable of feeling pain possesses at least one Description of a certain kind (i.e., being capable of feeling pain *is* possessing an appropriate kind of Functional Organization).
- (3) No organism capable of feeling pain possesses a decomposition into parts which separately possess Descriptions of the kind referred to in (2).
- (4) For every Description of the kind referred to in (2), there exists a subset of the sensory inputs such that an organism with that Description is in pain when and only when some of its sensory inputs are in that subset. (Putnam 1967, p. 200)[137]

Keep this definition in mind. It will be important later when we address specific criticisms. Now we should turn to Searle's famous Chinese room.

### 3 The Original Chinese Room

Since this is the basis for the entire debate, it is worth stepping through the original Chinese room argument as it appeared in the *Behavioral and Brain Sciences* [86]. I will intersperse comments to clarify points that come out heavily in the ensuing debate. All page numbers refer to [86]. If you consider yourself already quite familiar with the argument, you may want to skip this section and only use it to refer back.

Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes. (Pages 417-418.)

Notice that Searle "understands" the rules in English. The rules allow him to "correlate" one set of formal symbols with another set of formal symbols. By "formal" Searle apparently means that he attaches no meaning to them, but instead they are identified entirely by their shapes.

Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch “a script,” they call the second batch a “story,” and they call the third batch “questions.” Furthermore, they call the symbols I give them back in response to the third batch “answers to the questions,” and the set of rules in English that they gave me, they call “the program.” (Page 418.)

Now this is a key move on the part of Searle. Notice that he steps out of the first person perspective, namely Searle-in-the-room manipulating symbols, to the third person perspective. With this move he now describes the interpretations given to the symbols by the people who are providing them: namely, the first group of symbols is a “script,” the second a “story,” and the third “answers.” In other words, the symbols are given *semantic* content, but not by Searle-in-the-room.

Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external points of view – that is, from the point of view of somebody outside the room in which I am locked – my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. (Page 418.)

Now Searle fulfills the dreams of many an AI researcher. The program passes the Turing test (Turing 1950)[149]! Notice that “following the instructions” means reading the rule book in English, which clearly involves understanding. The correlating of Chinese symbols, however, remains at the level of shape recognition.

As regards the [claims of strong AI], it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. For the same reasons, Schank’s computer understands nothing of any stories, whether in Chinese, English, or whatever, since in the Chinese case the computer is me, and in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing. (Page 418.)

Before proceeding it is important to clarify what Searle is and is not claiming with this argument. First of all, he is not claiming that the word “understanding” does not have degrees and levels. All he needs for his argument to go through is that there exist clear cases where “understanding” literally applies and clear cases where “understanding” does not literally apply. Then, for the example, he needs the reader to agree that he “understands” English and does not “understand” Chinese, which, Searle claims, is an empirical fact in his own case.

This presents the Chinese room argument in the aspects under which it is most heavily criticized in the subsequent barrage of literature. Searle has a positive thesis as well, namely that the human mind is caused by the brain (a biological system) and nothing else. He repeatedly asserts throughout his articles and books that consciousness is an evolved, natural trait like lactation or photosynthesis. In other words, he holds to a straightforward version of materialism without any detailed argument. This positive thesis will not be the concern of this paper; instead I will focus on the negative argument of the Chinese Room, namely that no digital computer can understand solely by virtue of running a formal program. Note that I will also not enter into the debate about whether or not intentionality and consciousness entail each other; Searle thinks they do, and I agree, so sometimes my wording slips back and forth between the two. The reader may substitute her favorite term – say, intentionality – each time she reads consciousness, or vice versa (replace intentionality where it appears with consciousness).

Searle categorizes and replies to various arguments against his; however, instead of considering his counterarguments here, we will present them in the context of specific authors’ criticisms.

## 4 Computation

Before proceeding to the criticisms, we cover some basic accepted definitions and results regarding the notion of computation.

In computer science the notion of computability – also called definability (Enderton 1972)[118] – has been formalized and examined in great detail. For a problem to be computable is the same as for it to be definable in an algorithmic language (writing down the definition gives us the explicit steps to compute the answer). In other words, a function is computable if we can write down the exact steps such that for any input, there exists a finite  $n$  such that after  $n$  executions of our exact steps, the output of the function has been calculated. Of course, the “steps” specified must themselves be finite, such as move a tape head one position to the left or right, write a symbol on the tape, or change state (among a finite choice of possible states). In fact, the finite steps just specified are, informally, those of a Turing machine – for a more formal definition see (Lewis and Papadimitriou 1981, p. 170)[122].

Turing showed that any attempt to strengthen his initial machine, e.g. by adding extra states or multiple tapes, does not increase the number of functions computable on the machine. If a function is computable on any “strengthened” machine, then it is also computable on his original machine, i.e. there exists an equivalent Turing machine that computes the same function (hence the expression “Turing equivalent” or “Turing equivalent function”). Now, the Church-Turing Thesis states that *any* computational procedure can be carried out by a Turing machine (Church 1936)[154]. This is a thesis, not a theorem, because it is not a mathematical result; it simply asserts that our informal concept of computability corresponds to Turing’s formal model. It could be empirically disproven if, given a particular function, an alternative model of computation that fulfills the requirement of finite labor at each step in the computation were always able to calculate the output to the function in a finite number of total steps, **and**, for the same function, we could prove that no Turing machine could always compute the output in a finite number of total steps. However, “no one considers this likely,” certainly not any prominent computer scientists (Lewis and Papadimitriou 1981, p. 223)[122].

As an example of a non-Turing equivalent procedure, consider the function that, given the number 2, returns its complete square root in base ten. Since the number of digits used to represent the square root of 2 in base ten is infinite, no procedure can be specified that determines the full sequence of digits of  $\sqrt{2}$  in a finite number of steps. (However, for practical computer systems, there is always a limit to the precision of any number, and hence computations can be carried out involving irrational numbers, although of course the results of such calculations are subject to round-off error.)

We now turn to a consideration of the chief lines of criticism directed at Searle in the literature. I will, in each case, first describe the objection to Searle (summarized as an **Objection**) and then outline Searle’s response (concisely captured as a **Reply**). From time to time I will also interpolate my own comments.

## 5 The System Reply

The first and most obvious reply is to say that while Searle does not understand Chinese, the entire system – Searle, rulebook, bits of papers, and perhaps the room – does in fact understand Chinese. In fact, observes Wilensky, if we *ask* the system if it understands Chinese, it will tell us yes (Wilensky 1980)[107]. Searle is just a subcomponent to the system. Just because a part of the whole – Searle is just a part of the Chinese understanding system – does not understand Chinese, it does not follow that the whole does not understand Chinese.

**Objection A** *Searle-in-the-room is just a subcomponent of a Chinese understanding system. Just because Searle-in-the-room does not understand Chinese, it does not follow that the system*

*does not understand Chinese.*

Searle has a simple reply to this:

[L]et the individual internalize all of these elements of the system. He memorizes the rules in the ledger and the data banks of Chinese symbols, and he does all the calculations in his head. The individual then incorporates the entire system. There isn't anything at all to the system that he does not encompass. We can even get rid of the room and suppose he works outdoors. All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system that isn't in him.

Searle's move is clear and decisive: he **is** the system, but he only performs computations on symbols. He is not a subcomponent of the system anymore. Yet, it remains an empirical fact, asserts Searle, that he does not understand Chinese.

**Reply a** *Let Searle-in-the-room memorize the rulebook and the bits of paper so that he is the system. He can execute any program and still not understand Chinese. Hence the system does not understand Chinese, because Searle does not understand Chinese.*

Interestingly, of the six members of Computer Science departments among the 27 commentators to Searle original *BBS* article, five responded with some version of the system reply [49, 60, 63, 80, 107], while only a few of the remaining 22 commentators endorses the system reply. Among Computer Science faculty, only Schank[85] gives a different reply, the robot reply, which we consider next.

## 6 The Robot System Reply

By far the most popular reply is the Robot System Reply, and the most eloquent proponent is Stevan Harnad. In an insightful article (Harnad 1989)[42], Harnad, who “abstemiously umpired the debate” in *Behavioral and Brain Sciences* (Searle's original article was accompanied by 27 commentaries, and later issues contained 8 more commentaries), makes many useful distinctions and launches the most convincing version of the argument in the literature to date, in my opinion. Note, however, that a similar but less developed argument was given by Bynum four years before Harnad (Bynum 1985)[9].

## 6.1 Simulation and Implementation

Suppose we prototype the flying of a new aircraft and test our design with simulation. Subsequently, on the maiden voyage the plane flies successfully. This empirical success, says Harnad, shows that “from the standpoint of our functional understanding of the causal *mechanism* involved, the two are theoretically equivalent: They both contain the relevant theoretical information, the relevant causal principles.” A mechanism is a physical system operating according to causal, physical laws. Understanding a mechanism, Harnad says, is knowing its relevant causal properties. In this case, the causal properties relevant to flight are necessary.

What about a critic who argues that the success is a fluke? Although Harnad does not explicitly address this point, he does point out that the program must simulate with a great deal of technical knowledge – aerodynamic factors like lift and drag, material strengths, etc. The likelihood of an accidental success, it would seem, would be extremely small. Presumably repeated successes with other planes, using the same relevant causal principles, would further decrease the possibility of a fluke.

Harnad next distinguishes two kinds of implementations and two corresponding kinds of functionalism. A *c-implementation* is running software on a computer chip, i.e. the implementation of the flight simulation using formal computer code (presumably assembly code of some type). A *p-implementation*, on the other hand, is the actual physical device built on the basis of causal principles formally encoded and tested in computer simulation. Note especially the fact that the actual physical device has causal connections to the world. This distinction between c- and p-implementation mirrors that made in Smythe’s reply to Searle (Smythe 1980)[101].

Now, Harnad labels a “hardware fallacy” the belief that the brain-in-a-vat (Dennett 1981)[24] is a c-implementation instead of a p-implementation or hybrid p/c-implementation. (Dennett’s brain-in-a-vat is a brain physically separated from the rest of the body yet in complete contact via radio transmitters, thus allowing the brain to experience vision, etc., as if the brain were still in the body.) In other words, the brain-in-a-vat is not software only; it is either all hardware or mixed hardware-software. If one falls into the hardware fallacy, then one believes a special kind of functionalism:

The form of functionalism underlying the hardware fallacy is better described as *symbolic functionalism*: the belief that mental function is really just symbolic (e.g. verbal, inferential, computational) function; that the mind manipulates symbols the way a Turing machine does, and hence that the brain just supports the hardware for doing computation – just a c-implementation (Harnad 1989, p. 8)[42].

Harnad footnotes Fodor (Fodor 1981)[39] and Pylyshyn (Pylyshyn 1985)[72] as two examples of researchers guilty of the hardware fallacy. The mark of *symbolic functionalism* is the belief that

the mental function is purely symbolic and Turing equivalent (see Section 4 for a discussion of Turing equivalent functions). *Robotic functionalism*, on the other hand, includes nonsymbolic functions, for example sensors and actuators (mechanisms for movement or control).

Harnad’s point is that Searle’s test, the verbal Turing Test, discredits symbolic functionalism, but not robotic functionalism, which requires the “Total Turing Test” – being able to do all that we human beings can do. This is roughly the same point made by Smythe(Smythe 1980)[101] and Russow(Russow 1984)[82], namely that Searle’s example does not work against robotic functionalism.

## 6.2 Brain modeling

Next consider brain theory. The difference between simulation and implementation is the same as the difference between simulation and synthesis. In the case of the airplane, the simulation was the computer model running, which showed successful flight. The implementation or synthesis was the actual plane which flew in the air.

Now consider robotic functionalist theory. A *simulation* of such a theory will not have a mind, but the *p-implementation* will, Harnad argues.

**Objection B** *Even if a simulation does not have a mind, its p-implementation (physical implementation with causal connections to the world) can have a mind.*

## 6.3 Syntax versus Semantics

Two serious problems arise when we try to ground symbols in a semantic use.

- (1) With symbols defined only in terms of other symbols, there is no way to get out of the loop.
- (2) How symbols are *experienced* as *meaningful* (Nagel 1974)[65].

These problems, Harnad acknowledges, have not been satisfactorily answered by anyone at present. So what is the status of a robotic functionalist p-implementation? If this robot is regarded as merely syntactic, “then surely *we* are only syntactic devices too”!(Harnad 1989, p. 15)[42]. This last response I call the “what-else-could-it-be?” argument (Dreyfus uses the same label[116] as does van Gelder[151]). Lycan (Lycan 1980)[55], Jacquette (Jacquette 1989)[51], and Newell (Newell 1990)[134] also give the “what-else-could-it-be?” argument. Newell expresses the view as follows:

[A]lthough a small chance exists that we will see a new paradigm emerge for the mind, it seems unlikely to me. Basically, there do not seem to be any viable alternatives. This position is not surprising. In lots of sciences we end up where there

are no major alternatives around to the particular theories we have. Then, all the interesting kinds of scientific action occur inside the major view. It seems to me that we are getting rather close to that situation with respect to the computational theory of mind (Newell 1990, pg. 5)[134].

While Newell argues for his physical symbol system hypothesis, which is a c-implementation, Harnad instead observes that our brains have causal connections to the world. Thus, if neural interactions can be fully described by a formal model (a c-implementation) combined with inputs from the outside (yielding a mixed c/p-implementation), then, since neurons constitute the brain, our mixed c/p-implementation has the causal powers of the brain. By Searle's own logic, it would follow that such a robotic functionalist c/p-implementation (or just a p-implementation – recall that any software program can be translated to pure hardware) has a mind, since the implementation captures the causal properties of the brain.

Now, Searle is right to point out that we have no detailed account of how semantics arises in virtue of syntax. Yet what is the alternative? Robot system design is the best option currently available. Thus, philosophically, a robotic functionalist explanation is the most reasonable explanation we have, even though a lot remains to be worked out.

**Objection C** *Yes, we do not know how semantics can rise from machine syntax. But what else could it be? Machine syntax must be able to give rise to semantics, we just don't know how yet.*

Pylyshyn registers a similar objection by saying that rather than declare semantic determination by syntax to be impossible, as Searle does, a better question is to ask what can fix the semantic interpretation of a functional state (Pylyshyn 1985)[72]. To his credit, Pylyshyn notes the extreme difficulties that have to be overcome in order to answer his question. In particular, observes Pylyshyn, the Lowenheim-Skolem theorem shows that the output of a digital computer can always be coherently interpreted as referring to integers and arithmetic relations over the integers. Thus, it would seem that at least one semantic interpretation of any digital computer will always be an arithmetic relation. Nevertheless, he holds out hope that the relevant interpretations can be constrained enough to yield definite semantic content. Then we can judge the results as with any other scientific endeavor (Pylyshyn 1985)[72].

Cognitive Science, Harnad believes, will not tell us what consciousness and meaning *are*, but will show us “what mechanistic principles will generate their outward manifestations.” Whether or not a robotic functionalist system can successfully model the human person, including intentional states, is an open question, says Harnad.

## 6.4 Searle's Reply

Searle had already encountered this sort of reply in his conversations with Yale researchers. Thus he replies to it in the original 1980 article in *BBS*. Searle first notes that the reply concedes that symbolic functionalism is false, since the robot requires causal connections to the world (hence no methodological solipsism). Next he says the following:

But the answer to the robot reply is that the addition of such “perceptual” and “motor” capacities adds nothing by way of understanding, in particular, or intentionality, in general, to Schank’s original program [or any other AI program]. To see this, notice that the same thought experiment applies to the robot case. Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case, you give me more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose, unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot’s legs or arms.

This is a key move, note it well: Searle-in-the-robot receives sensory input from the television camera and appendages but only in the form of Chinese symbols. To return to Searle:

It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these other facts. I am receiving “information” from the robot’s “perceptual” apparatus, and I am giving out “instructions” to its motor apparatus without knowing either of these facts. I am the robot’s homunculus, but unlike the traditional homunculus, I don’t know what’s going on. I don’t understand anything except the rules for symbol manipulation. Now in this case I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program I have no intentional states of the relevant type. All I do is follow formal instructions about manipulating formal symbols. (Page 420.)

The key move here is to isolate Searle-in-the-robot from directly experiencing the sensory input from the television camera and appendages. This preserves the conditions of the original Chinese room: nothing more than the shapes of the Chinese symbols are recognized. The robot movement is automatically directed by the output Chinese symbols, without any meaning attached to the symbols for movement. Searle-in-the-room does not understand Chinese; thus the Chinese symbols output cannot be said to be due to intentionality. It follows, Searle

argues, that the robot then has no intentional states at all, since Searle-in-the-robot(room) is the robot's homunculus and Searle-in-the-robot understands nothing about what the robot is doing.

To reply to Harnad, Searle would clearly claim that his robot thought experiment is **not** just a *simulation* but is an actual *implementation* – specifically, a mixed c/p-implementation. Given any feasible program to control a robot, Searle-in-the-robot can execute it and still not understand. Hence, the robot lacks intentional states, and robotic functionalist hypothesis fails too.

**Reply b** *Given any feasible program to control a robot with transducers, Searle-in-the-robot can execute the program and still not understand. Therefore, the robot lacks intentional states.*

## 6.5 A Problem with Robotic Functionalism

I will attempt in this section to show that the robotic functionalist approach of Harnad leads to an absurd consequence. To start, let's assume that the robotic functionalist hypothesis is correct. Namely, we have a robot called Rover whose mixed c/p-implementation achieves intentionality, but Rover's simulation in a c-implementation does not have intentionality.

Now suppose Rover is sent to Mars to search for certain types of rocks on the surface. Rover learns a great deal, sufficient to cause intentionality and consciousness. Scientists on earth converse with Rover and learn many new facts about Mars. Now upon returning to earth the scientists disconnect Rover's external limbs and sensors (T.V. camera, etc.), connecting instead a standard computer terminal and keyboard in order to exchange language symbols with Rover. The scientists have many more questions about Mars and so the conversation continues just as before. However, in this case there are no sensory/transducer connections to the world. Hence, according to the robotic functionalist hypothesis, Rover is now a c-implementation and lacks intentionality.

Suppose further that the scientists connect Rover's limbs and sensors back, reverting the exact situation as when Rover had just returned from Mars. Suddenly, then, Rover achieves intentionality again. The conversation about Mars continues uninterrupted and without any noticeable change. In fact, as an experiment, the scientists continue connecting/disconnecting Rover's limbs and sensors, thus causing a continuous sequence of intentional  $\leftrightarrow$  nonintentional "switches," without any corresponding changes in observable behavior. This consequence seems quite absurd – the only difference is a few physical connections, yet intentionality pops in and out!

**Reply c** *Robotic functionalism allow intentionality to “switch” in and out infinitely often based on a few physical connections, without any noticable behavioral changes. This is an absurd consequence.*

We next consider a different tack in objecting: questioning Searle’s intuition.

## 7 Dennett’s Objection

Daniel Dennett is among the original 27 responders in *BBS*. He holds nothing back, labeling Searle’s argument “sophistry” (Dennett 1980)[25]. The basic problem, says Dennett, is that Searle’s story is based on an *intuition pump*, namely that since he does not understand Chinese, neither does the program, and then neither does the robot. While Dennett agrees with the intuition in the original Chinese room case, he claims it is a trick to extend the same intuition to handle the robot reply. Since Searle knew that this was an important objection to his *gedanken* experiment, why not present the robot as the original case?

Imagine you **are** the robot, Dennett says. You walk around, talk to friends, and perhaps even participate in natural selection. Do you lack understanding? Now it is not so clear. (A similar point is made by Edelson (Edelson 1982)[31].) In fact, says Dennett, the article “Where Am I?” pumps exactly the opposite intuition (Dennett 1980)[24].

**Objection D** *Searle’s example is just an intuition pump. Equally valid but opposite intuitions can be gathered from other examples.*

Furthermore, consider a robot which is an exact behavioral duplicate of us at the beginning of the human evolutionary chain. We would have evolved exactly the same, according to Searle, but would have lacked understanding. How does Searle account for this? Perhaps some kind of *élan vital*? We know that Searle is a materialist, so this cannot be the case. So the problem is that Searle “is looking *too deep*.” It would be just as mysterious to peer into the synapse-filled jungle of the human brain. (This last comment is a repeat of Objection C.)

**Objection E** *Where is the supposed difference due to “intentionality” and “semantics” if we would have evolved the same anyway? Semantics does not really exist but is a prescientific notion.*

### 7.1 Searle’s reply

Fortunately for us, Searle had a chance to reply to Dennett in the very same issue of *BBS*, and two years later when reviewing a book by Dennett. The problem with Dennett’s objection, says Searle, is that it is underdescribed, since in imagining us to be the robot, we are never

told what is going on inside the robot's mind (Searle 1980b)[87]. With the level of detail in Searle's original robot reply, we see that nowhere is any semantic content attached to the Chinese symbols. Even if Searle's own body were the robot, his brain would still be unable to learn Chinese without semantic content available. Thus Dennett does not generate any counterintuitions since his version does not describe all the relevant facts.

**Reply d** *Dennett's Objections Dand E and related examples underdescribe the robot's mind. With the extra detail added, we see from Reply b that no semantic content arises. Thus, Dennett's objection does not generate counterintuitions.*

## 7.2 Hofstadter and Dennett's Further Objection

The next round occurred the following year with the publication of *The Mind's I* by Hofstadter and Dennett. In the book they republish Searle's Chinese room argument with their own commentary. They assert: "Searle has committed a serious and fundamental misrepresentation by giving the impression that it makes any sense to think that a human being" could execute by hand the sequence of steps of an AI program that understands Chinese. In other words, Searle's example is an impossibly unrealistic concept of the relation between intelligence and symbolic manipulation. The Chinese room argument is invalid because the difference in complexity Searle assumes overcome is not possible (Hofstadter and Dennett 1981, p. 373)[48].

Certainly Hofstadter and Dennett are right in pointing out that Searle's example ignores real-time issues. An AI Chinese language understanding program would consist of millions and millions of lines of code executing in parallel. It would probably take Searle-in-the-room a year to answer one question, and this is being generous.

**Objection F** *Searle's example is impossible. Specifically, a human could never execute the instructions of an AI program in a reasonable amount of time.*

### 7.2.1 If speed is the issue, give a particular bound

Note, however, that Hofstadter and Dennett do not deny that Searle can execute the *elementary* machine instructions of a computer in constant time. Thus the *computational complexity* of Searle and of the computer execution will be identical, only differing by a constant factor. Hofstadter and Dennett's objection is weak and only serves to point out that if time bounds are placed on input-output speed, then Searle's experiment obviously might not be able to meet them. However, if speed really is the issue, then Hofstadter and Dennett need to explain why any particular choice of such time bounds would give rise to intentionality and understanding in the faster implementation and not in the slower [126].

However, suppose Hofstadter and Dennett were willing to give a particular time bound, e.g.  $10^{12}$ . But then a computer which works  $10^{12} - 1$  times as fast as Searle-in-the-room would not understand, whereas suddenly a little bit faster and it would understand. This certainly sounds strange. Hofstadter and Dennett would have to come up with a different reply.

**Reply e** *If speed is really the issue between real understanding and Searle's Chinese Room implementation, then state a particular time bound which has to be achieved to reach understanding. Any such bound results in a possible situation where a computer program runs just above and below the bound, flipping in and out of understanding, with no behavioral difference. This is a reductio ad absurdum.*

### 7.2.2 Silicon brain replacement

Hofstadter and Dennett next present a second line of attack following the example of Zenon Pylyshyn. Let's give Pylyshyn's original reply:

From [Searle's] point of view it would be extremely unlikely that any system not made of protoplasm - or something essentially identical to protoplasm - can have intentionality. Thus if more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output *function* of each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now except that you would eventually stop *meaning* anything by it (Pylyshyn 1980)[71].

Now, to be fair to Searle, he says quite clearly that Pylyshyn's example is an empirical possibility; Searle only requires that such silicon replacement maintain the causal powers of the brain (Searle 1980b)[87]. Searle's claim **is not** that silicon will never realize the causal powers of the brain; this may be possible, Searle says. Instead, Searle's claim **is** that silicon will not realize the causal powers of the brain *in virtue of instantiating a formal program*. His further empirical claim is that all sort of substances in the world, such as water pipes and toilet paper, lack such causal powers.

Finally, Hofstadter and Dennett claim that Searle's reply, as it stands, gives us no way to tell the difference between genuine meaning and the genuine "you" from artificial meaning or an artificial "you." For Searle to profess that we can hear the same sound two different times, and on one occasion the sound has a meaning, while on another it does not, is to have his cake and eat it too. Their complaint lends further support to Objection E, in that they claim that Searle's account gives us no way to distinguish true intentionality from pseudo-intentionality. Hence, perhaps "intentionality" and "semantics" are prescientific notions.

### 7.3 An Exchange

In 1982, Searle had the opportunity to review Hofstadter and Dennett's book in the *New York Review of Books*. This generated an interesting exchange between Searle and Dennett. For starters, Searle grants that his thought experiment is unrealistic in that it is too slow. Nevertheless, he insists that the system has no way to attach meaning to the uninterpreted Chinese symbols. Semantics is still lacking. To drive the point home, consider the following scenario:

So let us imagine a thirst-simulating program running on a computer made entirely of old beer cans, millions (or billions) of old beer cans that are rigged up to levers and powered by windmills. We can imagine that the program simulates the neuron firing at the synapses by having beer cans bang into each other, thus achieving a strict correspondence between neuron firings and beer-can bangings. And at the end of the sequence a beer can pops up on which is written "I am thirsty." Now, to repeat the question, does anyone suppose that this Rube Goldberg apparatus is literally thirsty in the sense in which you and I are?

Notice that the thesis of Hofstadter and Dennett is not that *for all we know* the collection of beer cans might be thirsty but rather that if it has the right program with the right input and output it *must be* thirsty (Searle 1982b)[89].

Dennett objects that Searle's "causal powers of the brain" remain totally mysterious. Furthermore, Dennett notes, Searle has not responded to the accusation that his theory provides no way to distinguish meaningful speech from meaningless speech (Dennett 1982)[26].

Searle responds by granting Dennett's conclusion that Searle's theory does not allow one to isolate meaningful speech; however, Searle says, it is just like saying that a steam locomotive and an electric train can have exactly the same output while working on completely different internal principles. Thus it is not so out of the ordinary to say that a robot could duplicate our speech and behaviour while not understanding (thus lacking intentionality).

Searle also notes that Dennett does not reject Searle's beer can example as a logical consequence of Dennett's view. Thus Dennett accepts the *reductio ad absurdum* (of course Dennett would not characterize the consequence as such).

**Reply f** *Under strong AI, we are forced to grant that a collection of beer cans and strings implementing the right program **must be** thirsty. This is a **reductio ad absurdum**.*

Next we will consider the reply by an eminent philosopher, Georges Rey.

## 8 Combined Robot and System Reply

Georges Rey agrees with Searle on one important point: the inadequacy of the Turing Test (Rey 1986)[79]. The Turing Test, Rey points out, is behavioral, whereas strong AI is a version of functionalism. So his first objection is quite simple:

**Objection G** *Functionalism is not committed to the Turing Test. Showing the Turing Test to be false does not disprove functionalism.*

Functionalism, Rey continues, posits that in order to be the same, two symbol systems must have the right sorts of internal states, regardless of how similar the behavior is. He introduces two terms into the debate as follows:

If a system's inputs and outputs are not mediated by the *right sorts* of internal states – in the case of strong AI, the *right sort* of program – then the system will not be regarded as satisfying some mental predicate, no matter how much its behavior may resemble the behavior of a system that does. Let us call the stronger condition, whereby a system not only behaves like, but follows the “same program” as an existing mental being, “AI-functionally equivalence,” the view that this condition is sufficient for mental states, “AI-Functionalism.” (Rey 1986, pp. 170-1)[79]

To be AI-functionally equivalent, Rey declares, we have to have the same *inside* the room as well. In other words, according to strong AI, a mental state is a functional state, not a behavior. Thus, behavior will not be enough to conclusively determine what functional state (mental state) a thinking being is in. Clearly, then, a system which passes the Turing Test may yet lack mentality and not be AI-functionally equivalent, since the behavioral equivalence does not necessarily imply functional equivalence.

### 8.1 What is “functional equivalence”?

Unfortunately Rey does not go into more detail about what constitutes “functional equivalence.” So, I will borrow some terminology from Section 2.3 to make his position more specific. Otherwise it would be difficult to analyze Rey's arguments.

Let system  $A$  be described by a Machine Table which specifies the inputs, states, and next states for each input and state. Now the Description of  $A$  consists of distinct states  $A_1, A_2, \dots, A_m$  related to each other in the Machine Table. The Machine Table gives us the Functional Organization of  $A$  relative to the Description. In an exactly parallel manner, we have system  $B$  with Description  $B_1, B_2, \dots, B_n$  and a unique Machine Table specifying the Functional Organization of  $B$ .

Now let  $X$  represent a sequence of inputs of arbitrary length. We assume that system  $A$  is in state  $A_i$ ,  $i \in \{1, 2, \dots, m\}$ , and system  $B$  is in state  $B_j$ ,  $j \in \{1, 2, \dots, n\}$ . Then state  $A_i$  is *functionally equivalent* to state  $B_j$  if and only if the output of  $A$  is indistinguishable from the output of  $B$  for every possible input sequence  $X$ . This is the standard definition of state equivalence in Electrical Engineering and Computer Science (Roth 1985, p. 353)[141].

However, this does not seem to capture what Rey wants. Certainly he would like to allow for different behaviors for people in the same mental state (functional state), even given the same input sequence? In other words, suppose you and I are both in pain. A hospital nurse offers us both some Advil. Under the definition above, we would have to both either accept the Advil or reject the Advil in order for us to be in pain. Yet certainly Rey does not intend this consequence. So this attempt to define “functional equivalence” fails. Thus, it is unclear what Rey means by “functional equivalence” or “AI-functional equivalence.”

**Reply g** *“AI-functional equivalence” is not well defined and does not appear to withstand scrutiny and hence must be clarified before being used philosophically.*

Now let’s take a closer look at Rey’s argument. He asserts the following:

The question for strong AI is rather: is what is happening *inside* the room *AI-functionally equivalent* to what is happening inside a normal Chinese speaker? (Rey 1986, p. 171)[79]

Behavior is not enough. Functionalism is different from behaviorism. We can summarize this objection as follows:

**Objection H** *AI-functional equivalence is not just behavioral equivalence. The same internal states have to occur as well.*

This really does not qualify for the status of a formal objection to Searle precisely because, as noted above, the terms used in the objection are not clearly defined. The objection derives its appearance of validity from its vague appeal to “functional equivalence.” Nevertheless, Rey could possibly shore up Objection H with a precise definition.

## 8.2 Connectionism

Furthermore, Rey declares, Searle provides no rules for the other programs that the AI functionalist needs to have. Programs for perception, depth ordering, belief fixation, and problem solving are not specified in Searle’s scenario. So we need to modify Searle’s example and have many CPUs running in parallel in order to really run a full AI system. Hence, we need a

connectionist model. The nested CPUs execute the various interacting programs required for different subsystems. In addition, we need methods for communication between the different subsystems, e.g. perception and belief fixation. In short:

**Objection I** *The Chinese room does not model a connectionist machine.*

In fact, Rey argues, functionalism is not committed to the autonomy of the language subsystem. Perhaps a teletype hooked up to a computer can never be AI-functionally equivalent to the language subsystem of a normal Chinese speaker. If the other subsystems (depth ordering, perception, etc.) are needed for the language subsystem to function properly, then it is no surprise that Searle's Chinese room is unable to understand language.

**Objection J** *Functionalism is not committed to language autonomy. Failing at the level of language does not prove that functionalism is false.*

Finally, functionalism is not committed to imputing to the agent's CPU what is properly only a property of the agent. For Searle's robot reply to work, he has to maintain not that *we do not know* how semantics arises from syntax, but that *it is impossible* to derive semantics from syntax. This is similar to a point made by Harvey (Harvey 1985)[44]. Instead, Rey avers, we should ask *how* can syntax and causal connections to the world yield semantics? Perhaps logic, sensory input and decision theory could attach meaning to the Chinese squiggles.

Now the beer can analogy (Reply f) does cause people to feel uneasy, just as did Ned Block's original version of the analogy involving the nation of China or the Chilean economy (Block 1978)[110]. Yet, says Rey, it seems difficult to see how to avoid it. What is needed is some principled way to distinguish the right stuff from the wrong. No one has a solution to this yet.

This brings us to a final point. What is involved in understanding ordinary Chinese? What *are* the necessary and sufficient conditions for understanding? Rey proposes the following: that understanding consists of rationally selecting inputs ("hypotheses") that turn out to be reliably true. To claim that only human neurons and causally equivalent entities can have mentality is to commit specious speciesism. The system reply argues that Searle-in-the-room plus the rest of the system understands, even if Searle-in-the-room does not. Without going into the drawbacks of this reply, a combined robot and system reply seems to hold the best promise for fulfilling AI-functional equivalence.

### 8.3 Commentary

Rey covers a lot of ground that represents common themes throughout the literature. Objection G is quite common – in fact, I think it is fair to say that most philosophers who have considered the issue reject the adequacy of the Turing Test. (Informally, in my experience,

probably most computer scientists accept the adequacy of the Turing Test. However, most computer scientists are not familiar with the philosophical problems with behaviorism.) In any case, if we abandon the Turing Test, then the claims of strong-AI lack testability, and hence become non-scientific claims. (Unless, of course, someone proposes an alternative to the Turing Test.)<sup>n</sup>

Objection I is also quite common. The basic problem I have with this objection is that Searle nowhere limits the size or nature of his rule book, other than that it be a set of formal rules in a Turing equivalent sense. Thus, all of the extra programs mentioned by Rey are covered in the original Chinese room argument. Rey nowhere explains why, other than speed, multiple CPUs are *necessary* to execute the many AI programs needed to model different subsystems. Certainly he cannot point to any formal reason of functions computable in the connectionist model which are not computable with a single CPU.

**Reply h** *A result of the Church-Turing Thesis is that a Turing machine can implement any program or set of programs implemented on a connectionist machine. Hence, Objection I is simply not true.*

In a similar vein, it is unclear why Objection J cannot be dealt with by adding the extra programs necessary to support language autonomy. The unstated premise necessary for Objection J to hold is that the rulebook in the Chinese Room Argument cannot contain any other program beyond that of language understanding. However, this premise is clearly false ( unless Rey says that a c-implementation is not enough for the set of AI programs needed, in which case he would have to argue along the lines of Harnad).

**Reply i** *The Chinese Room Argument nowhere states that the rulebook cannot contain a other programs in addition to a language understanding program. Hence, the Chinese Room Argument shows that **any** collection of programs will lack understanding, not just the language program, and so Objection J fails.*

I commend Rey for bringing up Reply f. His response that “it is not at all clear, however, how it is to be avoided” (Rey 1986, p. 181)[79] is incompatible with his conclusion “that Searle’s example is not an example that ought to cast any doubt on Strong AI” (Rey 1986, p. 175)[79]. Certainly the problem is difficult to avoid if one is committed to strong AI; otherwise it is fairly simple to avoid. But at least Rey mentions it; most critics of Searle just sweep it under the rug.

Overall, in my judgment, Rey’s battery of objections offers one of the most sophisticated versions of Objection C.

## 9 The Robot Reply Refuted

Maloney categorizes three standard replies to Searle's robot reply. None of them, he claims, succeed (Maloney 1987)[57].

### 9.1 Tea-Carrying

The first standard reply is that Searle misrepresents the power of the AI Chinese-understanding program. Strong AI would not only require Chinese understanding but also consistent conversations and behavior. For example, the robot would prepare tea if asked in Chinese. However, even if we *include* tea-carrying, it does not follow that the robot *understands* what he is doing. For example, let Marco (Maloney's name for robot-Searle) memorize all the rules that the robot follows. Marco can follow instructions about how to get a cup, put hot water in it, put bag #8 in the water, etc., without *understanding* that he is making tea. Thus, Searle's robot reply can be extended to work against a stronger form of the Turing Test. Basically, this is a fuller version of Reply b.

### 9.2 He really does understand

The second standard reply is exemplified by Dennett and Cole (Cole 1984)[13]. The strategy here is to say that Marco *does* understand Chinese, despite his protestations that he does not.

**Objection K** *Despite his protestations to the contrary, Searle-in-the-room really does understand Chinese.*

According to Maloney, the problem with this reply is that then strong AI includes the claim that semantic content is not necessary for language comprehension. Marco cannot translate Chinese to English, and he does not understand, semantically, Chinese words. The ability to translate, at least to some degree, is necessary for language comprehension. If this is granted, then the strategy (saying Marco really does understand Chinese) fails. Thus we conclude that Marco's inability to translate overrides his answers, in Chinese, that he does understand Chinese. In other words, translation is a necessary consequence of understanding.

Russow has a reply to Maloney: a correct program *would* translate Chinese to English. The problem with this reply is that to require rules that allow translation claims that to fully instantiate Chinese one needs to be able to translate to English too. But clearly this is false. And to say that Marco's understanding of English is simply mastery of a formal program, and thus the two formal programs (English understanding and Chinese understanding) must be able to interact, is to beg the question of Searle's general argument. Namely, Searle argues that

he-as-robot *understands* English and only *follows rules* for Chinese at the level of shapes, with no understanding.

Even further, suppose Marco masters rules for speaking and translating between Japanese, Sanskrit, and Tibetan. Marco **still** does not understand any of them, even though he can translate between them.

### 9.3 Someone else understands

A third and final standard reply says that although Marco does not understand, Polo, who is distinct, does. The question remains: who is Polo? Polo could be (a) the entire physical system or (b) a cognitive agent inside of Marco. These two options correspond, respectively, to the system reply(objection) of Objection A and the following objection:

**Objection L** *A subsystem of Marco (Marco is Searle-in-the-room) does in fact understand Chinese.*

In Maloney’s setup, Marco has a set of cards that specify the rules for translating each language. Given this assumption, a problem with (a) can be seen by first observing that the “entire physical system” includes the cards. Now allow Marco to shift cards to another set that have exactly the same information written on them. Polo must still be there! If we say that Polo is Marco plus all possible sets of cards, then suppose Polo plays poker. Polo is now Marco plus Chinese cards plus poker cards. Now let Polo play pool. Since (a) specifies the entire system, Polo is now Marco plus Chinese cards plus the pool table and balls. Only someone in the grip of an ideology, says Maloney, would not see the absurdity of these consequences.

**Reply j** *If the system of Searle-in-the-room(Marco) plus language translation cards is a separate understanding entity, then we end up with the logical consequence of the entity being poker cards, a pool table & balls, and other objects used by Searle-in-the-room. This is a **reductio ad absurdum**.*

Now, assuming that Marco and Polo are distinct, the problem with (b) arises from the fact that they cannot both be identical to the body they share. Neither can either alone be identical to the body, since arguments favoring one can be easily recast to favor the other. To remain consistent with materialism, we must construe Marco and Polo as identical with, or realized in, different parts of Marco’s body. They could share resources (nervous systems, etc.) to some extent. In any case, now strong AI denies personal unity.

## 10 Brain simulation

A final objection registered by several commentators observes that a simulation of the brain's neuronal pattern could surely produce intentional states. Ringle sternly objects to Searle's "belief that the physical properties of neuronal systems are such that they cannot *in principle* be simulated by a nonprotoplasmic computer system" (Ringle 1980)[80]. What are Searle's mysterious "causal powers"? They could be a direct linkage of the nervous system to the external world. But then Searle has to make the case that an organic rod or cone in a human retina captures information which, in principle, a photoelectric cell could not. The other option for Searle's "causal powers" is the capacity of protoplasmic neurons to produce phenomenal states – sensations, pains, etc. But, says Ringle, this is just like the ancient claim that only organic creatures like birds or insects can fly.

**Objection M** *A computer simulation of the brain at the neuronal level would produce intentionality.*

Searle responds by saying that he posits no such limit on simulation of neuronal systems. His only claim is that a simulation is not the same thing as the real thing, just as simulated hurricanes do not destroy towns and simulated digestion does not digest food (Searle 1980b)[87].

**Reply k** *A simulation of neural interaction does not have the same causal powers of the real thing (real neural interactions), just like a simulation of a hurricane does not destroy buildings.*

### 10.1 Simulation can be Real

Dyer has an interesting reply to Searle's simulation argument (Dyer 1994, pp. 190-1). Suppose we replace each molecule with a binary representation of all aspects of its three-dimensional topology and chemistry. Assume our scientific knowledge is complete enough for us to specify all interactions at the molecular level. Now simulate the interactions.

Note that there is a systematic correspondence between the program and the actual molecules. If life is a systems phenomena, emergent from the neurobiology, then such a system really is alive, says Dyer.

Thus, Dyer gives a more specific example which fleshes out Ringle's claim. Of course, Ringle and Dyer would say, we do not yet have the capability to simulate fully at the molecular level. If we could, however, we could duplicate the causal powers of the brain. Presumably, then, we could write a program that achieves intentionality and consciousness.

In my judgment, Dyer's argument is yet another version of Objection C.

## 10.2 Summary and Status

The next section will cover the most direct and extensive debate to date between Searle and some of his critics. But first, let us take a step back and see where we are.

We have seen that the Chinese room generates strong intuitions about understanding. In particular, all commentators so far have agreed to the basic empirical facts Searle claims for the scenario: Searle-in-the-room understands English and does not understand Chinese; Searle-in-the-room is able to step through the rules of a computer program. Harnad's robotic-functionalism reply is a valiant attempt that, among other things, formulates a common objection, namely Objection C. Dennett questions whether semantics is precise enough of a notion to use it in analysis (Objection E), but Dennett does not contest Reply f as a logical consequence of his views. Rey formulates Objection G which seems to be widely agreed to by philosophers on all sides[150]. Maloney has shown how the system reply leaves one open to Reply j. Finally, Searle gives a weak response to Objection M and says that he places no limit on the ability of non-biological materials to simulate causal properties of biological materials.

## 11 Searle versus Churchland and Churchland

To start off the 1990's, *Scientific American* decided to organize a written debate between John Searle on one side and Paul & Patricia Churchland on the other. Searle repeated his Chinese room argument with a more explicit set of premises and conclusions than previously published, which we will begin by restating here (Searle 1990)[93].

### 11.1 Searle's axiomatized argument

Searle starts by clarifying what he is **not** arguing.

**Definition 11.1** *A machine is a physical system capable of performing certain functions.*

Given this definition, says Searle, yes machines can think, because humans can think. A human is nothing more than a physical system. This assertion of materialism is unargued and is not necessary for his subsequent negative argument. However, it is important to see that Searle thinks that his reasons are independent of the metaphysical debate about the existence of immaterial substances.

The question Searle addresses is the following: "Can a machine think just by virtue of implementing a computer program?" In other words, his target is symbolic functionalism (at least at this point). It remains to be seen whether or not he extends the argument to work against robotic functionalism.

**Axiom i** *Computer programs are formal (syntactic).*

**Axiom ii** *Human minds have mental contents (semantics).*

**Axiom iii** *Syntax by itself is neither constitutive of nor sufficient for semantics.*

**Conclusion I** *Programs are neither constitutive of nor sufficient for minds.*

At this point Searle announces that his argument “has nothing to do with common sense.” As for connectionism, he asserts that any function which can be calculated on a parallel machine can be calculated on a serial machine. This assertion is supported by Dyer, a researcher in connectionism, who says that any neural network – the basis of connectionism – can be simulated on a Turing Machine (Dyer 1994, p. 189)[29]. (That neural networks are the basis of connectionism can be seen in (McClelland and Rumelhart 1986)[124] and (Smolensky 1988)[148].) Therefore, asserts Searle, connectionism poses no new challenges to the Chinese room argument, answering Objection I (this reply was listed earlier as Reply h).

**Axiom iv** *Brains cause minds. (Materialist axiom.)*

**Conclusion II** *Any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains.*

**Conclusion III** *Any artifact that produced mental phenomena, any artificial brain, would have to be able to duplicate the specific causal powers of brains, and it could not do that just by running a formal program.*

**Conclusion IV** *The way that human brains actually produce mental phenomena cannot be solely by virtue of running a computer program.*

Searle then lists the common objections, which we have already been discussing. Specifically, he lists Objections A, B, E, K, L, and M. Searle swiftly deals with all of these objections by arguing that none of them comes to grips with the core of the Chinese room argument: the distinction between syntax and semantics, i.e. formal symbol manipulation and the attachment of meaning to symbols.

We can learn as much from the objections Searle does not list as from those he does. In particular, Searle does not list Objections D, G, H, J, or C. Presumably he feels that he has answered Objection D already in his replies to Dennett: first of all, it is just plain wrong, since we *do* have the intuition that there is no understanding in the Chinese room; second of all, if you deny the intuition, you end up with absurd consequences like attributing mentality to

a collection of beer cans and strings. As for Objections G and J, these are in fact significant concessions that many AI researchers are unwilling to concede anyway. Finally, Searle does not address Objections H and C. He could claim that the first objection is a more detailed condition than strong AI postulates. As for the second, “what else could it be?” is hardly an argument, and it does not deal with the absurd consequences that follow if we grant the functionalist story.

Now for Churchland and Churchland’s novel objection to Searle raised in the *Scientific American* article.

## 11.2 Syntax Can Generate Semantics

Churchland and Churchland directly take on Searle’s Axiom iii. Without it, none of his conclusions follow. They base their argument on an analogy with electromagnetic forces. Suppose we lived in 1864 when Maxwell suggested that light and electromagnetic waves are identical. To oppose him, we construct the following argument.

**Axiom v** *Electricity and magnetism are forces.*

**Axiom vi** *The essential property of light is luminance.*

**Axiom vii** *Forces by themselves are neither constitutive of nor sufficient for luminance.*

**Conclusion V** *Electricity and magnetism are neither constitutive of nor sufficient for light.*

These are meant to exactly mirror Axioms i, ii, iii and Conclusion I. Now consider that, together with the above argument, we tell the following story:

Consider a dark room containing a man holding a bar magnet or charged object. If the man pumps the magnet up and down, then, according to Maxwell’s theory of artificial luminance (AL), it will initiate a spreading circle of electromagnetic waves and will thus be luminous. But as all of us who have toyed with magnet or charged balls well know, their forces (or any other forces for that matter), even when set in motion, produce no luminance just by moving forces around!

Maxwell might respond by saying that the frequency of oscillation by the man is too low to have any effect visible to us. He might even bite the bullet and say that the room is, nonetheless, bathed in a very small amount of light, even though we don’t see it.

Similarly, argue Churchland and Churchland, the Chinese room “looks” semantically dark. But that does not *demonstrate* that it is so. We may have an uninformed commonsense understanding of semantic and cognitive phenomena. Searle’s Chinese room exploits our lack of understanding of these phenomena. In reality, they say, “Searle is once more mistaking the limits of his (or the reader’s) current imagination for the limits on objective reality.”

### 11.2.1 Two Comments

Two interesting comments before proceeding. Churchland and Churchland echo Georges Ray in Objection G and reject the validity of the Turing Test. This seems to be a trend, as noted previously (see Section 10.2).

Secondly, Churchland and Churchland bring up the same issue as Hofstadter and Dennett did. However, then the question of Section 7.2.1 still applies: what is the time bound? Churchland and Churchland appear to say that, in fact, there is no time bound: the Chinese room “looks” like it does not understand, but in fact it does. It is just that the understanding is exceedingly slow.

Unfortunately for Churchland and Churchland, this leaves them with Replies f and j, neither of which they answer.

### 11.3 Searle’s Response

In his response to Churchland and Churchland, Searle begins by noting that arguments from analogy are usually quite weak because it is difficult to ensure that the two cases are analogous. In this particular case, asserts Searle, the analogy fails because formal symbols have no physical, causal powers. Now the Churchlands declare that Searle is begging the question when Searle asserts that formal symbols are not equivalent to mental contents. One way to test for an obvious truth, says Searle, is to assume the converse and see what follows. So let’s assume that mental contents *are* formal symbols.

Now suppose instead of translating Chinese, Searle-in-the-room starts interpreting the Chinese symbols as moves in a chess game and begins to play. Or he correlates the symbols to the stock market and plays the market. (Admittedly he would not do very well.) Now which semantics do the Chinese symbols represent? Chess? Chinese? Stocks?

Recall, says Searle, that strong AI rests on *formal programs*. There is a dilemma for the Churchlands’ analogy between syntax and electromagnetism: either syntax is in terms of purely formal mathematical properties or not. If it is, then there’s no analogy, because syntax has no physical powers and no causal powers. If syntax is not purely formal, then there is indeed an analogy, but not one applicable to strong AI.

## 12 Thinking Machines and Virtual Persons

In 1994 a book was published which declared itself dedicated to showing that the Chinese room argument is wrong. “Without [Searle], this book wouldn’t be necessary,” proclaims the editor (Dietrich 1994a, p. 5)[22]. The book, titled **Thinking Machines & Virtual Persons**,

contains many interesting articles. It has three sections. The first criticizes the notion of intentionality; Daniel Dennett has an article here. The second deals with the issues we have been dealing with so far in this paper – the nature of causal powers of computers and so forth. The third sections deals with psychology and role of representation. I will try to cover some select arguments from the second section (Dyer 1994, who is in the second section, has already been mentioned earlier).

## 12.1 Causal Powers of CPUs

The first article of the second section argues that Shakey the Robot, who once inhabited the hallways of SRI research labs in Menlo Park, CA, achieved a certain level of consciousness (Cole 1994)[16]. The basic problem, says Cole, is that *Searle's* mind grapples with a rule book and bits of paper, but the resulting Chinese understanding is *not* Searle's. So really Searle can only conclude that *he* does not understand Chinese. Therefore, Searle does not refute the idea that a suitably programmed computer can really understand a language.

Cole asserts that when he understands English, *he* is the one that understands, and not his body nor his head. It is a *person* who understands. It may be the case, says Cole, that a single physical system implements two distinct persons – indeed, a virtual person, in fact! Thus, Cole proclaims, his answer is *not* the system reply nor the robot reply.

Now Shakey the Robot could stop, see objects, bump into things, get feedback from the environment, and formulate goals. Thus, claims Cole, Shakey has a limited level of consciousness, a subjective quality.

### 12.1.1 Commentary

This article is one in a series of articles by Cole that plays on the notion of personal identity[13, 14, 15]. He rests on the plausibility of a continuous notion of personal identity, much like (Parfit 1971)[136]. In other words, you are not a single united person from conception through death; you can be many people. Thus, in the case of the Chinese room, there are two or more people implemented in Searle's body *without* having to divide sections of the body implementing each person. Cole uses this move to discredit Maloney's arguments (Section 9), specifically Reply j: Marco and Polo both *are* implemented at the same time by the same body without any parallel processing or time sharing necessary (Cole 1991b, p. 408)[15]. Polo is a "Virtual Person" implemented in Marco's body. This move also allows Cole to claim that Shakey is conscious. Cole nowhere addresses Reply f.

The problem as I see it with Cole's claim of consciousness for Shakey is that it would also extend to the Chinese speaker in the Chinese room scenario. Cole admits this, and sees no

problem with it. However, then perhaps Terry Winograd's SHRDLU is also conscious. Even my calculator and perhaps the thermostat in my apartment is conscious, since they all implement goal-oriented computations. These possibilities are just a logical consequence of the fact that Cole never deals with Reply f explicitly in his writings. Since I consider these to be absurd consequences, I consider Cole to be refuted.

In addition, I must confess I find Cole's premise on the nature of personal identity to be implausible and somewhat ludicrous. If a relative of Cole suffered bodily harm from someone who claimed to be implementing a program for someone else, I doubt he would drop all charges and try to sue/arrest the "other" person.

## 12.2 Intentionality and Computationalism

Many commentators express surprise at Searle's claim that a computer simulation could behave exactly like an actual human person yet not have intentional states. Dyer is no exception, and he questions Searle on this issue as well (Dyer 1994)[29]. Imagine, Dyer says, person  $P$  answering all sorts of questions in English about himself. Then, let  $CP$ , who is implemented in virtue of  $P$  and a rulebook, give wonderful answers in Chinese, assert that he –  $CP$  – exists, and so forth. Why should we deny that  $CP$  exists?

### 12.2.1 Result of the Chinese Room: A Strategy on the Turing Test

This question is a common theme, and I think it deserves an answer. If Searle is right, then a general strategy would be to ask questions that require intentionality and consciousness. For example, consider the following conversation (T is the tester, C is the computer):

T: I'm not very happy today. Are you?

C: What makes you think that you are not happy?

T: It's just a feeling. Tell me about your feelings.

C: I am feeling quite happy.

T: Why? You shouldn't be. Am I acting nice toward you?

Notice the question about *how* the tester is acting toward the computer. If formal computer programs lack intentionality, then any computer programmer is going to have a very hard time simulating correct behavior with these kinds of questions. Eliza, Parry, and SHRDLU gain their effectiveness from the limited domain in which they operate. (Eliza simulated the questions of a Rogerian psychologist; Parry simulated the responses of a paranoid mental patient; and SHRDLU lived in a simplified blocks world where he could move objects and answer questions

such as who owned a particular block. For a good overview of these programs and others, see (Copeland 1993)[17].)

So my answer to Dyer is the following: *CP* would have a very difficult time passing the Turing Test. In particular, if a series of questions follows what I call an *intentional strategy*, that is, asking questions that require a feeling *about* the world and concepts, then the chances of *CP* passing the Turing Test would be very slim indeed.

We now consider some aspects of the most recent round of salvos on the issues raised in the Chinese room.

## 13 Nonreductive Functionalism

In a recent book, Chalmers argues for what he calls *nonreductive functionalism*. Just as one may believe that consciousness arises out of a physical brain without being a physical state, so also can one believe that consciousness arises out of a functional organization without being a functional state. This is nonreductive functionalism – “a way of combining functionalism and property dualism” (Chalmers 1996, p. 249)[12]. Property dualism holds that certain high level descriptions – pain or wetness, for example – are not reducible to lower level physical descriptions.

### 13.1 Fading Qualia

The most interesting new argument added to the debate by Chalmers, in my opinion, is what he calls the “fading qualia” argument. We start with two systems: on the one hand, a robot which, since it is implemented by computer chips and transducers, lacks intentionality and consciousness; on the other hand, we have a person who presumably has both intentionality and consciousness.

Now consider very small changes in both. To the robot, we add a few neurons. To the person, we replace some neurons with a silicon chip. The input/output behavior is maintained exactly the same with each replacement. This idea is not new; it was first proposed by Pylyshyn (see Section 7.2.2). We continue in this way, constructing a whole collections of intermediate cases. Eventually we get to the point where the robot is completely made of flesh and bones; the person is completely made of silicon chips and transducers. Now Chalmers asks the following question: *What is it like to be the person?* Given that in the final result intentionality and consciousness have switched, one of two things must have happened along the way. Either (1) consciousness gradually faded away (Chalmers calls this case *fading qualia*) or (2) there was some point where consciousness suddenly flicked out.

Consider option (2), which Chalmers dubs *suddenly disappearing qualia*. That at some specific point the replacement of a neuron or two causes a complete blinking out of consciousness, with no corresponding change in behavior, is difficult to swallow. Furthermore, at what point does this occur? At 25% of neuron replacement, or 50%? The point would be entirely arbitrary. Furthermore, we can rerun the experiment, says Chalmers, at a finer grain within the neuron, “so that ultimately the replacement of a few molecules causes a whole field of experience to vanish” (Chalmers 1996, p. 255)[12]. This seems quite bizarre. This option, suddenly disappearing qualia, is not disproved, but it appears quite implausible.

Searle answers that if this scenario were possible (he thinks it is empirically not the case that silicon can duplicate the causal powers of neurons), then (1) would be the most likely explanation (Searle 1992, p. 66)[94]. But consider the person, let’s call him Joe, who becomes the robot. Joe certainly thinks he has conscious experiences at the start, for he does. But as the replacements continue, he slowly loses consciousness, *and does not realize it at all*. Remember, we hypothesize no change at all in external behavior. So, for example, if Joe is viewing something red, he slowly loses the qualitative raw feel (qualia) of seeing red. Presumably the red stops being *bright* on the continuous spectrum to losing consciousness and hence qualia. So Joe experiences seeing pink, or some such shading. The crucial feature here, says Chalmers, is that Joe is systematically *wrong* about all that he is experiencing (Chalmers 1996, p. 256)[12]. Even worse, on a functional construal of belief, Joe will *believe* that he is completely conscious. The account seems to run into many problems here. Joe will have rational processes and thinking, in fact Joe is conscious, but he is utterly wrong about his own conscious experiences. Yet his behavior is not impaired in any way at all, and he (and we) have no way to judge the change. So option (1) seems quite bizarre too.

## 13.2 Panpsychism

Chalmers openly embraces panpsychism – the idea that consciousness is everywhere. For example, in discussing the differences of consciousness between thermostats and rocks, Chalmers notes, “It may be better to say that a rock *contains* systems that are conscious” (Chalmers 1996, p. 297). This kind of discussion is a logical consequence of biting the bullet on Reply f. Instead of tracing back his theory to see which axioms lead to the absurd consequence, Chalmers embraces it. In this regard, Searle makes the following comment: “Of all the absurd results in Chalmers’s [sic] book, panpsychism is the most absurd and provides us with a clue that something is radically wrong with the thesis that implies it” (Searle 1997, p. 48)[95].

**Reply 1** *Claiming that rocks and beer cans can have consciousness leads to panpsychism. This is a reductio ad absurdum.*

## 14 Some New Arguments

In this section I'll try to add a new argument to extend Searle's argument against his own postulate of causal reductionism. By causal reductionism I mean Searle's repeated assertions that intentionality is reducible to the causal powers of the brain. Hilary Putnam has offered various arguments against reductionism [139]; what follows is my attempt to add to the list. Note that the argument advanced here does not argue against the Chinese Room Argument in any way; on the contrary, it tries to extend it.

### 14.1 Computer plus chemical change

First, we need some definitions.

**Definition 14.1** *A computer system is a combination of a formal computer program plus connections to the world.*

In Searle's robot reply he allows the input from a television screen to be passed into the room in the form of additional symbols. So presumably the Chinese room argument still works against a computer system.

Now for my own *gedanken* experiment. A similar argument is given by Cuda (Cuda 1985)[19].

Suppose we have a machine called a *Chemical Change Machine*, or CCM for short. This machine can alter chemical bonds and atoms even at the subatomic level. I take the possible existence of CCM as an empirical issue. I add a CCM to a computer system as a causal connection to the world. In particular, given a particular collection of molecules, CCM can alter chemical bonds, move electrons, and recombine atoms. It may possibly even cause fusion or fission, if absolutely necessary. The computer system, given a particular symbolic input representing the current state of a collection of molecules, can calculate the intermediate states necessary to arrive at a particular final state, where each intermediate state comprises only a single change in a chemical bond or atom. For each intermediate state, CCM can then take the symbolic output of the computer and perform the alteration specified.

Now I claim as a matter of empirical scientific fact that this computer system, appropriately programmed, can digest pizza (this is one of Searle's examples). Digestion is, after all, just a change in the chemical bonds of food particles.

Now consider the changes that occur in the brain. Of course we are limited in this consideration since currently we do not know all of the brain's biochemical workings. Nevertheless, assume we have a model which accurately characterizes the biochemical causal properties of the brain. Now, we implement a computer system composed of the computer model of the brain

connected to CCM. According to Searle's Conclusion II (see Section 11.1), we now have the causal powers to create a mind. Then it follows that when we run the program, the computer system will understand and have intentionality.

But this contradicts Reply b of Section 6.4 (and Section 9.1), which extends Conclusions III and IV of Section 11.1 to apply to a robot system with transducers. How can this be avoided?

One way is to deny the possibility of creating CCM. However, this would pose serious limits on science, which is not an attractive alternative.

Another way out is to say that the causal powers of the brain are due to more than atomic, subatomic, and biochemical changes. Again, this seems unattractive. (Remember, Searle **nowhere** addresses the issue of what types of causal powers the brains have – he views it as an empirical issue.) In particular, to claim that the causal powers of the brain are due to more than biochemical changes is a strong prediction about the nature of brain science which I am not sure anyone is prepared to make. For starters, what other kinds of changes are there in the brain?

Perhaps there are some other ways to escape from the contradiction with Reply b, but I am not sure what they would be. In any case, there seems to be another problem elucidated by this example.

#### 14.1.1 Another problem

Suppose we grant that the computer system causes a mind in virtue of the causal powers of CCM. However, then the question still remains: where is the semantics? I would assert the following:

**Axiom viii** *A change in a chemical bond or atom does not cause semantics.*

Note that Hofstadter makes the same point in his original reply to Searle when Hofstadter says, "There is no intentionality at the level of particles" (Hofstadter 1980)[49]. However, perhaps this goes against Searle's Axiom iv. In any case, Searle would probably not accept Axiom viii. So let's apply Searle's own method and assume that it is not true. What follows? Say a neuroscientist finds which chemical bond change or atomic change gives rise to semantics. He or she isolates the molecule or collection of molecules where the change occurs. Using CCM, presto!, the simulated brain now has semantics and intentionality! This arrives at a similar point as the *suddenly disappearing qualia* of Chalmers (see Section 13.1). Namely, a molecular change causes a whole field of qualitative raw experience to suddenly appear. So, as before, while we do not disprove this possibility, we nevertheless observe that it appears quite implausible.

So if we accept Axiom iv and Axiom i of Section 11.1, then the addition of CCM to the computer system does not add semantics. Recall that at each step in the computer program, we either have a formal symbol manipulation or a change in a chemical bond or atom.

**Conclusion VI** *A computer system composed of a formal computer program and a causal connection to the world which can arbitrarily cause a change in a chemical bond or atom does not cause semantics.*

Then, by Axiom iii, we arrive at the following:

**Conclusion VII** *The computer system of Conclusion VI is neither constitutive of nor sufficient for minds.*

But remember that the computer system has the causal powers of the brain, and hence by Searle's Conclusion II it is sufficient to cause a mind.

Searle's only way out of this seems to be to deny that the computer system captures the causal powers of the brain. This would mean that something more than atomic, subatomic, and biochemical changes are needed. **This option posits a fundamental limit on the nature of brain science: some causal powers more than those associated with atomic, subatomic, and biochemical changes must exist.** This option seems quite unattractive.

## 15 The Reaction of AI: Why Does This Matter?

From the AI perspective of science, why does this matter? Why would a researcher in hardware-software codesign (namely, the author of this paper) spend the time to read the philosophical literature instead of proposing new explanations and experiments? Isn't breath spent on the Chinese room example a waste of time?

A succinct version of this kind of objection is expressed by Nils Nilsson in a recent article in *AI Magazine*. He begins by lamenting the switch in focus in AI over the past 20 years to expert systems and high performance programs. After noting possible technical reasons for the switch, he states the following:

Finally, the arguments of those who say it can't be done might have had some effect. People who know insufficient computer science but consider themselves qualified to pronounce on what is possible and what is not have been free with their opinions (Penrose 1994, 1989; Dreyfus and Dreyfus 1985; Searle 1980).[135]

The implication is clear: of course it is possible for computer science to model the mind, and those outside of the field have insufficient knowledge to say what is not possible. No effort is expended in trying to address issues such as Reply f.

This type of reply was also the approach taken by Marvin Minsky. He says the following:

In this essay, Searle asserts without argument: “The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don’t. If you get a theory that denies this . . . the theory is false.”

No. The study of the mind is not the study of belief; it is the attempt to discover powerful concepts – be they old or new – that help explain why some machines or animals can do so many things others cannot. I will argue that traditional, everyday, precomputational concepts like believing and understanding are neither powerful nor robust enough for developing or discussing the subject (Minsky 1980)[63].

Now this is an amazing claim. Consider the possibilities: either both humans and thermostats have beliefs, or one has beliefs while the other does not, or none have beliefs. Minsky does not want to enter into this discussion; he holds that the concept “belief” is prescientific and will be replaced in the future.

Unfortunately Allen Newell apparently never took the time to reply to Searle’s argument, although he strongly disagreed with it. Many authors note that Searle’s argument is most clearly seen as a refutation of Newell’s “Physical Symbol System” hypothesis. His 1990 book, which grew out of the William James lectures he gave at Harvard University in 1987, never addresses Searle at all (Newell 1990)[134].

## 15.1 Reasons to pay attention

On purely technical grounds, Nilsson and Minsky have a point: “understanding” is not a precise, developed scientific term in the way that other terms are, e.g. mass. Does this mean that all the philosophical arguments are pointless? Well, on the one hand, there is something to be said for conceptual clarity. What exactly do we mean when we say a machine “understands”? All scientific enterprises start with some prescientific ideas. Debate at such an early stage may not be conclusive, but it still adds value in that it clarifies the exact goals and methods of the project. In this way we can specify which experiments will demonstrate what concepts, and if the experiments fail which hypotheses are disproved. Otherwise the scientific method – first hypothesize, then make experiments, and finally form theories based on solid results – cannot work, because we won’t know which experimental failures disprove which hypotheses. In short, science investigates phenomena; and certainly the phenomena of “understanding” is worth studying.

Following is the defense Terry Winograd makes:

My own interests in the debate lie along pragmatic lines. As active participants in an intellectual field, we are always faced with the question of “What is worth doing?”

Many of my colleagues in computer science and engineering are skeptical about the value of philosophical debate in answering such questions. They do not see such discourse producing the kinds of hard-edged answers that would give them definite direction and specific guidance. I believe that they are wrong, both in rejecting the value of such discussions, and in expecting answers that meet the criteria of mathematical and scientific argumentation. (Winograd 1997)[50]

A further philosophical question is raised by the reply of Nilsson and Minsky. Namely, it brings up the relationship between philosophy and science. But I'll leave comments on this until the conclusion.

## 15.2 Winograd and Flores

The philosophical approach of the paper so far, and the reactions of AI researchers, falls squarely within the analytic tradition. Suppose we were to take a different approach, that of phenomenology. One such attempt at this approach can be seen in an unpublished reply to Searle by Terry Winograd (included in this paper as Appendix A), as well as in the book by Winograd and Flores, *Understanding Computers and Cognition*[152]. In the book, according to one commentator, “Winograd himself has radically scaled back his estimate of the [AI] field’s potential.” (Smith 1991, p. 251)[147].

In Winograd’s reply, he says that Searle has a naive view of “understand.” Searle claims that there is a sense in which X understands and Y does not. To “understand”, according to Winograd, is not a simple object-predicate relationship, but is in fact situation-dependent. For example, one can say, “I read his dissertation but I didn’t understand it.” What we really need is a speaker-hearer interaction. A rough paraphrase of what it means to understand is given by Winograd as follows:

**Definition 15.1** *X understands Y if, having heard (or read) Y, X’s potential for future action is changed in the appropriate ways.*

Furthermore, says Winograd, to say a machine understands is to adopt a view of the machine as an autonomous agent. Our hesitation to do this stems from our social attitudes and is not a matter of being right or wrong – but **is** a matter of adopting attitudes that can be dehumanizing.

About SAM, Winograd says

Most people would feel that the understanding of a story must include more than the ability to answer simple questions about whether a hamburger was eaten. Therefore, Schank’s program does not undergo the “appropriate” changes, and does not understand.

Nevertheless, Winograd says, people are free to use whatever terms they want, as long as they are willing to pay the consequences. “We live within the world we create with our language,” asserts Winograd.

In conversation, and in the introduction to a recent journal dedicated to the issue of is the mind a computer program[50], Winograd has maintained that “understand” is not a precise term. We can always describe things as “understanding” in some sense. For example, we can hit a slab of slate and say, “See, it understands that it is supposed to stay together in one piece.” Thus, we cannot say that “understand” never applies to a rock.

### 15.2.1 Comments on Winograd

One small observation is that Definition 15.1 requires another definition of what it means to be “changed in the appropriate ways.” Assuming that Definition 15.1 is filled out, isn’t the definition good enough to conclude that Searle understands English (in most cases), that a thermostat *never* understands English, and that Searle *never* understands Chinese in the Chinese room scenario? In particular, in the Chinese room, Searle does has not “heard (or read)” Chinese, since presumably to read Chinese means to interpret the Chinese shapes, not just notice them as shapes.

We do create our own world to an extent, so in this Winograd is right. But there are limits. Searle places one of these limits at the claim that the strong AI research program can succeed in producing actual intentionality. If Winograd were to accept that humans understand and thermostats don’t, then it would be unclear why Winograd thinks that Searle’s arguments are invalid. However, in conversation Winograd has said that he really thinks Searle cannot apply a binary zero to the case of the thermostat, i.e. we cannot say that the thermostats really does not understand. In this case, however, then it is unclear how Winograd avoid the panpsychist consequences. If we can say that a rock can “understand” in some use of the term “understand,” then are not we left with panpsychism? In conversation Winograd has clarified that he does not think “understand” can always be applied in a clear way. For this reason he declares in the introduction to a recent issue of *Informatica*[50] that the questions, “Can a computer understand?” and “Can machines be intelligent?”, are incoherent. Similarly, to ask whether or not consciousness is everywhere (panpsychism) is an incoherent question. It is not that Winograd’s position endorses or opposes panpsychism; it is just that the question is incoherent, so obviously it has no answer.

The problem I have with this can be seen in the following way. First, as many authors have already done in the literature, I distinguish between “understand” in a data-processing sense – understand<sub>d</sub> – and “understand” in the same sense that you and I understand – understand<sub>h</sub>. So I would refine the question: can we say that a rock does not understand<sub>h</sub>? To ask this question

does not presuppose that we have differentiated *all* possible senses in which one can use the word “understand”; the question only presupposes that  $\text{understand}_h$  is sufficiently different from the other senses we may use “understand” (e.g.  $\text{understand}_d$ ) such that it can be distinguished from these other senses. Now, if we say yes, a rock cannot  $\text{understand}_h$ , then we avoid panpsychism. If we say no, a rock could possibly  $\text{understand}_h$ , then we allow panpsychism. Now I don’t see how any of Winograd’s arguments would disallow me from accepting his and Flores’ framework, yet say that a rock cannot  $\text{understand}_h$ .

At the very least, Winograd’s accusation of incoherence leaves him in the following position: he cannot say that panpsychism is **not** true.

### 15.3 What are the goals of AI?

As things stand now the AI field is split on its goals. The web page of the Human-Level AI seminars at Stanford contains the following quote:

The founders of AI research all had human-level AI as a goal. However, as AI research split up into many subfields and these split further, most research limited its ambitions. When students enter the field of AI, they almost always become attached to some ongoing activity with limited goals.

It is time for thinking about how AI gets to human level from where it is now. It is especially important that some students and other young people think about how this is to be accomplished.[119]

J. Feldman of the International Computer Science Institute in Berkeley, CA, notes the many different angles from which AI, as a philosophy of mind, is under attack. Even Allen Newell[134] talks about the likelihood of alternative formulations to rule-based AI. After decades of research and controversy, says Feldman, “No one has yet come close to formalizing human perception and thought, and there are now serious questions about the extent to which this is possible in principle.”(Feldman 1991)[32] For example, says Feldman, if Smith in the Chinese Room receives tiles saying he won the lottery, no output of tiles would even come close to duplicating his reaction if he were to be told about his lottery winnings in English.

There are two standard replies to the Chinese Room, says Feldman. The first is brain simulation. As a practical research direction, however, this idea is not worth very much as the number of molecules in the brain is on the order of  $10^{23}$ . The second reply begins with observing that human knowledge involves a great deal of social interaction and body language – shared experience. Common sense is based in this and thus seems difficult to express in rules. A highly contextual schema might provide understanding in computers. However, observes Feldman, the

chances of intelligence emerging from an unstructured neural network is nil. So the challenge is “to express the insights of AI and other cognitive sciences in a formalism that has appropriate computational and biological properties.” (Feldman 1991)[32]

Hence the technical challenge.

## 15.4 Strong AI as a Degenerating Research Program?

Hubert Dreyfus, a longtime critic of AI, holds nothing back in the introduction to his book, *What Computers Still Can't Do* (Dreyfus 1992)[115]:

[This] book now offers not a controversial position in an ongoing debate but a view of a bygone period of history. For now that the twentieth century is drawing to a close, it is becoming clear that one of the great dreams of the century is ending too. Almost half a century ago computer pioneer Alan Turing suggested that a high-speed digital computer, programmed with rules and facts, might exhibit intelligent behavior. Thus was born the field later called artificial intelligence (AI). After fifty years of effort, however, it is now clear to all but a few diehards that this attempt to produce general intelligence has failed . . . [Strong AI] is a paradigm case of what philosophers of science call a degenerating research program.[115]

Dreyfus declares that strong AI matches the degenerating research program as defined by Imre Lakatos: it defines a new approach which produces positive results in a few limited areas; then as researchers try to extend it more generally, it stagnates, leaving the door open for other new approaches which will gather the new generation of researchers.

Indeed, there seems to be support for Dreyfus' claim. Van Gelder comments:

[O]ver the last decade, or more, the computational vision has lost much of its lustre. Although work within it continues, a variety of difficulties and limitations have become increasingly apparent, and researchers throughout cognitive science have been casting about for other ways to understand cognitions. As a results, under the broad umbrella of cognitive science, there are now many research programs which, one way or another, stand opposed to the traditional computational approach; these include connectionism, neurocomputational approaches, ecological psychology, situated robotics, and artificial life.(van Gelder 1997)[151]

Similarly, Dyer talks about “recent disputes, between connectionism and traditional AI” (Dyer 1994, p. 189)[29].

Others are less pessimistic. B.J. Copeland says, “AI is still in its infancy. At the present stage computer intelligence is a field wide open for success . . . or failure. The future holds all the verdicts.” (Copeland 1993, p. 120)[17]

## 15.5 Logical Positivism

I would like to make one last point, regarding the naive view of logical positivism apparently held by various AI researchers and graduate students. The textbook used this year in Computer Science 221: Introduction to Artificial Intelligence was *Artificial Intelligence: A Modern Approach* by Russell and Norvig. In the section on philosophy, the only philosophers of this century mentioned are Bertrand Russell, Rudolf Carnap, and Carl Hempel[142]. The logical positivism of Carnap and Hempel is mentioned as if it is a viable and flourishing research program (otherwise why give them such prominence). In my conversations with other AI researchers, I was surprised to learn that they in fact thought that this is so. When I informed them that there were quite serious problems with logical positivism, and that as a result there are no philosophers today (as far as I know) trying to complete Carnap and Hempel's original program, they were quite astonished.

In brief, logical positivists claim that the meaning of a word consists in its method of verification. For example, how do I know that  $p$ ? To answer this one must first answer, what does  $p$  mean? The logical positivists claimed that the meaning of  $p$  consists in **nothing else** than the way in which one would come about to know that  $p$ . This is known as the *principle of verification*. For example, how do I know that John is in pain? Well, by observing his behavior. The paradigm of verification is that found in science – measure such and such and mix it with chemical  $x$ , and you will observe  $y$ .

Clearly, for this system to work, we need a base of *observation statements*. Observation statements are supposed to be obviously true, such as that John is jumping up and down or I see a red patch over there. The system also requires a notion of logical truth, so that some statements are considered analytic or true in virtue of the meanings of the words. For example, “a bachelor is an unmarried man” is an analytic truth.

Now anything that is not an analytic truth and cannot be verified, e.g. the assertion that God exists, is meaningless. The group of scientists and philosophers who first promoted logical positivism were known as the Vienna Circle.

If we take logical positivism seriously, then talk of the mind is meaningless, because we can't verify claims about mental states. Strong AI has no meaningful claims to make. And the distinction between strong and weak AI is also meaningless. In reality, the study of the mind reduces to talk of behavior.

One of the basic problems with logical positivism, however, lies in its observation statements. How do you verify an individual experience only seen once by one person? Even more importantly, basic scientific terms like mass seem impervious to definition as an observation. No logical positivist was able to give a coherent theory of observation statements that included

theoretical terms like mass and was consistent with the rest of logical positivism.

Even worse, the principle of verification itself came under attack. In particular, what are the steps to verify the principle of verification? It seems that it is not, in fact, verifiable in its own strict sense. But then it is meaningless.

As a result, logical positivism “is dead, or as dead as a philosophical movement ever becomes” (Passmore 1972)[156]. In fact, to discuss the “verifiability theory of meaning” in philosophical circles is like “flogging a dead horse” because “there are no longer any logical positivists left” (Wisdom 1963)[157]. A. J. Ayer himself, one of the leading protagonists of logical positivism in the English speaking world, when asked to describe the main problems with logical positivism, said, “I suppose the most important...was that nearly all of it was false” (Ayer 1978)[153].

For this reason, Paul Thagard, of the Cognitive Science Laboratory at Princeton University, after explaining their AI project, says: “There is no proposal of computationally renewing the positivist/behaviorist dream of reducing all meaning to external stimuli” (Thagard 1986, p. 143)[104].

For a more complete understanding of the movement I recommend an article by Church (Church 1949)[154] and a book by Hanfling (Hanfling 1981)[155].

In conclusion, the editors of *Artificial Intelligence: A Modern Approach*[142] did a poor job in letting the few pages on philosophy in the book give the false impression of logical positivism as a viable research program. All of this, of course, in no way denigrates the technical quality of Russell and Norvig’s textbook or any of my AI researcher friends!

## 16 Conclusion

### 16.1 Status

After seventeen years, the Chinese room argument seems alive and kicking. Clearly it is a classic. Eric Dietrich comments that “it is not difficult to say what’s wrong with the argument, at least to oneself; rather it’s hard to get one’s colleagues to agree with one’s particular diagnosis” (Dietrich 1994b)[23]. *This identifies a common intellectual commitment to computational reductionism without wanting to accept the consequences*, in my opinion. Dietrich says “since we are currently without a candidate mechanistic theory explaining all, or even most of, cognition, these intuitions remain personal and as much ruled by aesthetics as reason” (Dietrich 1994b)[23]. This is the strongest argument given by the opponents of Searle, the “what-else-could-it-be?” argument, or Objection C. The most compelling argument given by Searle, in my opinion, is the beer can *reductio*, Reply f. If I have to choose between denying Objection C versus accepting the result of Reply f as valid – namely, that beer cans and strings appropri-

ately connected do have intentionality – I will deny Objection C. This puts me in leagues with Winograd and Searle. I accept Searle’s negative argument against strong AI as carrying the day.

However, I think the mixture of Searle’s negative argument and positive account does not work. I find Searle’s hypothesis that intentionality is in the same league as lactation or photosynthesis to be as false as strong AI. In particular, I find Chalmers’ problem (Section 13.1) and my own arguments (Section 14) to be quite convincing, in addition to other philosophical arguments beyond the scope of this paper.

## 16.2 Philosophy and Science

I think a very serious issue is posed in trying to come to terms with the reaction of many AI researchers. In particular, they say that people outside of the field with insufficient knowledge feel able to proclaim what is and is not possible within AI. This brings up the age old question of the relationship between philosophy and science. Searle clearly holds the view that in some cases philosophy can tell ahead of time, before the results are in, what is and is not possible technically. His justification consists on basic axioms, such as that thermostats do not understand, and their logical consequences.

The question here is, will “understanding” ever be a completely scientific notion? Furthermore, is all experience reducible to science? Another philosopher who has thought deeply about this issue is Hilary Putnam. In a series of books and articles, he has pointed out the pseudo-scientific – and slightly problematic – character of reductionism and also of strong AI (Putnam 1991)[138]. He, along with Nussbaum, raises the possibility of an Aristotelean approach which might yield more salient philosophical and scientific results (Putnam 1994)[139]. But to treat these claims properly would probably require a Ph.D. thesis or more.

## 16.3 Closing comments

In closing, I would like to comment that many times, while reading through the various arguments, I would come up with what I thought was an innovative reply, only to find out later that someone had already made the argument. Also, what seemed to interest most writers was Searle’s dual assertions that a computer cannot cause a mind and that brains cause minds (materialism).

Finally, please note that the arguments presented here are a subset of the arguments presented by the respective authors. For the full set of arguments, please read the references. I only claim to have attempted to use my best judgment as to which were the most common and plausible arguments to consider.

# Acknowledgments

This paper benefited greatly from comments from Denis Phillips, Terry Winograd, Jose Meseguer, and Patrick Scotto Di Luzio.

# A Appendix: Terry Winograd's unpublished reply to Searle 1980a

Terry Winograd - Notes on Searle's Notes on Artificial Intelligence - version of May 26, 1979 4:43PM

Searle argues against two claims he sees implicit in work on Artificial Intelligence: That AI can “explain” human cognitive ability, and that a computer can be said literally to “understand” and have other cognitive states. I find myself in accord with his skeptical attitude towards both claims, but for reasons that do not correspond to his arguments. Searle fails to deal with the real phenomena at issue: the act of using AI-based explanations in discourse about human abilities, and the act of saying “X understands Y”. By assuming that the questions are “does the machine really understand”, and “is an AI model really an explanation”, Searle obscures what I see as the central problems.

I will first discuss the use of words like “understand” in describing the behavior of computer programs. In the light of this discussion I will review Searle's responses to the counter-arguments, and then reexamine the basic claims and explain my own concerns as to their validity.

## **“Understand” as an objective predicate**

Throughout the paper, Searle adopts the naive view that “understand” can be understood as a straightforward two-place predicate – that there is some objective sense in which it “really is the case” that X understands Y or doesn't. He makes liberal use of loaded modifiers like “obvious”, “perfectly”, and “certainly” in giving examples: “it is quite obvious in the example that I don't understand a word of the Chinese material, whereas I do understand the English material perfectly” (p. 4). “It is obvious that I do not understand the Chinese” (4), “The man certainly doesn't understand Chinese, and neither do the water pipes...” (10).

Searle's “just-plain-old-obvious-common-sense” posture is often a useful antidote to sophistry, but in this case it conceals the important issues. His interpretations of the counter-arguments all suffer from his unexamined adherence to the assumption that “understand” should be treated as a straightforward predicate. He explicitly states (13) “In ‘cognitive sciences’ one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects.” I believe that we need to understand “understand” in a different way (which Searle may choose not to call ‘cognitive science’).

## **“Understand” as a situation-dependent distinction**

Even a casual examination of the use of “understand” in everyday discourse reveals that it is not “obvious” when it is applicable. For the moment, let us deal only with cases of the

form “X understands (or understood) Y” where X is a person and Y is a linguistic text or utterance. Many interesting issues are raised by utterances like “Only Einstein really understood relativity”, “My analyst doesn’t understand me,” “I just can’t understand what happened at Jonestown,” and “Do you understand Serbian?”, but they aren’t central to this discussion. Even in the case of understanding something linguistic, there does not appear to be a clearly definable boundary between “understanding” and “not understanding”. We can say things like “I read his dissertation but I didn’t understand it.” “Do you think he understood your allusion to the hiring situation?”, and “It’s the kind of book a high-school-student can understand.” There are cases where from one perspective we would say that someone does understand something, while from another we would regard the same person as not understanding. As a native English speaker, I (in some obvious sense) understand a newspaper editorial. But if someone later points out some political undertones, I may later say “Oh, I really didn’t understand what he was saying.”

This kind of phenomenon is not a rarity – much of our basic vocabulary has the same property. If I ask “Was the crowd big?”, the answer does not depend simply on the number of people, but on a background of comparison with some anticipation (particular to speaker, hearer, and situation) of what size would be expected. Similarly, “understand” carries with it an unspoken consensus between speaker and hearer. (This is hinted at by the fact that we can apply intensifiers like “more” and “less” or “better” and “worse” in talking about understanding). As a rough paraphrase, we might say:

*X understands Y if, having heard (or read) Y, X’s potential for future action is changed in the appropriate ways.*

This paraphrase puts the weight of the speaker-hearer consensus on the word “appropriate”. In some contexts, it is appropriate for someone who has heard a fragmentary account of an event in a restaurant to answer the question “Did John eat the hamburger” in a certain way. In other cases (such as understanding a warning), there may be clear immediate actions that we expect, while in others (such as understanding a poem), the appropriate changes may be impossible to specify precisely. However, if a person reads a poem and his future potential for action is not in some way changed by it, we feel comfortable in saying “He didn’t really understand it.” Of course, for practical situations, there is no way of testing whether the appropriate changes have happened. One can make a finite number of specific tests (make observations, ask questions, etc.) but more could always be generated, and it is possible they would have different results. But before discussing the issue of just how much “coverage” is required, it is necessary to look at a different aspect of what is happening when someone says “X understands Y.”

### **“Understand” as an orientation**

In applying a predicate to an entity, one is implicitly committed to the belief that the entity

is the kind of thing to which the predicate properly applies. In a simple formal sense, this is couched in terms of “selection restrictions”. In saying that an idea is “green”, we are implicitly also saying that it is the kind of thing to which color predicates apply. Since it is not, the result is semantically ill-formed. If the world fell neatly into distinct categories, this formal notion of “restriction” would be adequate. But, as many people have pointed out, much of the use of language falls somewhere between this kind of simplistic object-property categorization, and something that might be called metaphor. If I say that a person or a bee or a volcano is “angry”, I am slipping from a case in which it is clear that the predicate is appropriate to others in which I am using it to convey not only the angriness, but also my orientation towards the object being characterized.

Much of the gut-level force of Searle’s argument comes from the unstated recognition that in uttering a sentence containing mental terms (“understand”, “perceive,” “learn”), we are adopting an orientation towards the thing referred to by the subject of the sentence as an autonomous agent. The issue is not whether it really is autonomous – the question of free will has been debated for centuries and work in AI has provided no new solutions. The issue is that in using mental terms, we make it an autonomous agent. In using the word “action”, rather than “behavior” in the above paraphrase for “understand”, this autonomy was implicit. Only an autonomous agent can understand; only an autonomous agent can act.

There are many reasons why one can feel uncomfortable with the tendency to adopt the same orientation towards people (who are the prototype for autonomous beings), and towards machines (or organizations). It isn’t that in doing so someone is right or wrong, accurate or inaccurate, but that they are accepting (often unwittingly) attitudes and role relations that can be dehumanizing and destructive of the social structure.

### **The counter-arguments**

If we accept the analysis of “understand” given above, there is a more interesting set of issues raised by the arguments than Searle admits. Taking them in his order:

1) The systems reply (Berkeley). “While it is true that the person who is locked in the room does not understand the story...he is merely part of a whole system and the system does understand the story.”

This person is arguing that although one cannot adopt an orientation towards the rule-following homunculus as an autonomous agent, one can adopt that orientation to this whole mysterious arrangement – a group to whom one passes stories and questions in Chinese, and that passes back answers after some inscrutable process of cogitation. The “system” does understand, in the same sense that one might discuss whether a review committee understood a funding proposal. As mentioned above, there may be reasons not to want to take this orientation to impersonal “systems” whether made up of people or otherwise, but the Berkeley

point is a valid one. To use the word “understand” one has to decide the boundaries of the system that will be treated as autonomous.

There is an interesting reflection of this point in the way we discuss our own minds. I can say “The rational side of me understands, but my feelings are....” In doing so, I am moving away from the more usual orientation in which I am a single autonomous entity, towards one in which I view myself as made up of a number of independent “subpersonalities”, each acting of its own accord.

2) The robot reply (Yale). “...a [physically embodied] robot would ... have genuine understanding.”

Once again, the respondent is elaborating what he or she sees as a condition for treating something as autonomous – its physical embodiment. Furthermore, the argument addresses the question of what it means for the “potential for future action to be changed in appropriate ways.” The measure of appropriateness is closely tied to the ways in which the possible actions are copies of the corresponding human actions. We would expect people in general to be more likely to adopt an orientation of autonomy towards an android than towards a less superficially human-like machine. This does not mean that they would have better philosophical justification for doing so, but that their whole background of perception and action towards the android would be affected by its physical characteristics regardless of their conscious recognition that it was a machine.

3) The brain simulator reply (Berkeley and M.I.T.) “Suppose...[the program] simulates the actual sequence of synapses...in the brain of a native Chinese speaker...Surely...if the machine understood the stories, and if we refuse to say that, wouldn’t we also have to deny that native Chinese speakers understood the stories.”

Searle finds this reply odd, since it depends on a notion of simulation that does not usually find its way into AI arguments. I think he misses the point that these people were responding to him (an admitted philosopher), in what they saw as a philosophical, not technical vein. The issue he raised is whether any computer could ever be described as “understanding”, and they were trying to pick a case that was most likely to convince him that such a computer could exist. This is not at all incompatible with other (technical) discussions in which they might argue that other ways of building the machine would be more effective or appropriate.

In responding to their arguments, he distinguishes between the “formal properties” of the brain and the “causal properties”. Lurking in this is an interesting and debatable form of dualism. He would agree that the brain operates according to physical causal principles, and that the computer (or water pipes or whatever) may well operate with analogs or representations of those same principles. He seems to argue that “mental causality” exists in some other domain, whose connections to physical causality are as mysterious as the mind-body problem has always

been. I will argue below a somewhat different version of this distinction (one that relates all versions of causality to the act of observation and interpretation). It is probably not one that Searle would accept, and without some clarification of this issue, Searle's argument does indeed leave him with the dilemma that a person (at least to the degree that a person is a physical being) doesn't understand either.

4) The strong brain simulator reply (M.I.T.) "Suppose we had the programs operate 'in parallel'...with..billions of people...each corresponding to one neuron...though we wouldn't say of any particular individual that he understood Chinese, couldn't we say it of the whole team?"

Searle states that this is the same mistake as the "system reply". I agree that the parallelism makes no essential difference. But in saying "it doesn't make any sense to make such collective ascriptions to the team" Searle is making the mistake of refusing to acknowledge that people can choose to adopt an orientation towards groups (or teams) as unitary autonomous agents.

5) The other minds reply (Yale). "[if] a computer can pass the behavioral tests as well as [people] can...if you are going to attribute cognition to other people you must in principle also attribute it to computers."

Searle discounts this argument by insisting that cognitive science cannot be based purely on behavioral description. "In 'cognitive sciences' one presupposes the reality and knowability of the mental." Given his dualistic view mentioned earlier, this implies that the cognitive scientist assumes without argument the reality of "the mental" for other people, but that for machines there can be "the computational processes and their output without the cognitive state." Since I don't understand his notion of "cognitive state" it is hard to address this.

I find quite different grounds for rejecting the "other minds" argument. Solipsism is a philosophical choice in some abstract intellectual sense, but not a real choice for anyone who lives in a society. By the very act of entering into conversation with other people (even if that conversation be a defense of solipsism) one is adopting an orientation that assumes their autonomy as individuals. This is not the case when one interacts with machines (or animals, or even all the people one deals with, as histories of slavery and genocide have demonstrated). There are alternatives for how we understand our relationship to machines, and although an individual does not have an unencumbered choice (being already enmeshed in the tradition provided by his language), the tradition as a whole can move in different directions. It is certainly not an a priori conclusion that we must adopt the same orientation towards people and machines, no matter how well they mimic each other in behavioral tests.

6) The many mansions reply (Berkeley). "...eventually we will be able to build devices that have these [mental] causal processes."

Searle claims an openness to this possibility, commenting that it seems to trivialize the project of artificial intelligence. Based on his earlier comments, I do not understand how he

can say “I see no reason in principle why we couldn’t give a machine the capacity to understand...since in an important sense our bodies and brains are precisely such machines.” My interpretation of his response to the “brain simulation” response denies this. If someone were to build an exact replica of a human being (down to the chemical details) it would seem that his earlier objections still apply – that the replica operates according to physical causal principles, and that there is no understanding in the molecules or in the neurons or in the constructed brain matter, any more than the water pipes of his example. A claim that this replica is in some essential sense different from a person must be grounded either in a strong form of mind-body dualism or in an argument based on differences of orientation as I have discussed above.

### **When is it appropriate to say that someone (or something) understands?**

Returning to an examination of the validity of the claims Searle sees in AI, I disagree with Searle’s assumption that there is some objective sense in which words like “understand” are correctly applied to some objects and not to others. The question is much more a social one – when is it appropriate (or to borrow a word from speech act theory, felicitous) to characterize a situation as “understanding”.

I find myself (along with most people I know) using mental terms for animals and machines often. I am quite likely to say “This program only understands two kinds of commands...” and to find this way of talking effective in explaining the behavior of computer programs to other people. In many contexts, it is perfectly clear to speaker and hearer that for some situation, the range of “appropriate changes of potential for future action” is quite limited and clear. In the sentence mentioned above, “understand a command” means to perform those operations that I as a programmer intend to invoke in giving the command. It is clear to me and to the person I address that other changes (such as getting impatient, or noting that I seem to give those kinds of commands often, and therefore treating me differently) are not appropriate. In this case, “understand” is being applied as “literally” as it ever could be. Any attempt to claim that it is being used “only metaphorically” or “incorrectly” flies against the facts of ordinary language use. By those standards, most of what we say is either metaphor or wrong.

When we look at discussions of AI, the situation is more problematic, since there is usually not a sufficient background of mutual agreement between speaker and hearer (or reader) about the range of appropriate change. Most people would feel that the understanding of a story must include more than the ability to answer simple questions about whether a hamburger was eaten. Therefore, Schank’s program does not undergo the “appropriate” changes, and does not understand. In general, since AI claims are couched in terms of “doing what a person does”, the natural assumption about the range of appropriate changes is that they include the full extent of what one would normally expect in an adult human native speaker of the language. In this sense, it is clear that no existing AI program understands, and that it is misleading to say

that it does, except in specialized technical conversation where the background of expectations is not based on full human abilities.

Having said this, I am not quite comfortable with the implication that the use of mental terms is fully appropriate even in those cases where the context makes the expectations clear. There can be a conflict between the two dimensions I listed in characterizing “understand” – the paraphrase as “appropriate change of potential for future actions”, and the orientation inherent in using mental terms. I am irritated when I hear people using mental terms in all of their discussions of machines (their car, refrigerator, etc.), and often when these terms are used for computers as well. The problem lies in the other consequences of adopting an orientation towards machines as autonomous. We live within the world we create with our language, and any consistent use of words leads to subtle but potentially important changes in our understanding of our world and how we fit into it. It seems overzealous to say that one should categorically avoid using mental terms in speaking of machines, but I am sympathetic with the view that there are dangers in doing so, and feel that we should be aware of the price we are paying whenever we do it.

### **Can AI explain human understanding?**

Quite aside from the question of whether one can appropriately apply a term like “understand” to a machine, there is the basic question of the nature of explanation of cognitive phenomena, and the role played by machine simulations. The discussion is often confused (as Searle points out) by the fact that AI programs do not purport to be detailed physical descriptions of the brain. I think there has been a persistent failure in AI discussions (see [1,2,3]) to distinguish between the domain of mechanism and the domains of behavior and intention.

All AI programs (by virtue of being computer programs that could be run on physical computers) are in the domain of mechanism. The claim is that for the relevant purposes, the differences in mechanisms (i.e. between a particular architecture of neurons and a particular digital computer) are irrelevant. The terminology and discussion, however, tends to place things in the domain of intention. In saying that a “goal” is represented by a certain formal structure, an AI researcher is doing two things: stating that there is some mechanism in the brain that is in some essential way equivalent to the formal structure; and stating that there is a straightforward correlation between the exercise of this mechanism in a brain, and the appropriateness of making the corresponding intentional attribution to the person whose brain it is.

There is an important insight that is central in computer science: that mechanisms which are quite different in detail can be equivalent when seen in terms of specific distinctions concerning their actual or potential behavior. It seems a very real question as to whether there are appropriate levels of abstraction at which brain mechanisms can be seen as equivalent to other

(potentially more understandable) mechanisms for which we have clear formal descriptions. What is critical is the recognition that in doing so, we are staying within the domain of mechanism. The fact that we have a more manageable abstract description of brain mechanisms does not solve the mind-body problem. AI programs are still body-descriptions.

With some fear of treading on deep philosophical problems outside my area of competence, I will contrast three different attitudes towards these issues.

1) Mental terms are reducible to mechanisms (Typical AI view). We can provide formal mechanisms with elements and operations labelled “goal,” “understand,” etc. A statement about a person that uses mental terms can be translated into an equivalent statement in terms of the operations of these mechanisms. The mechanisms are characterized at a level of abstraction such that they can be applied to both brains and computers, but ultimately they are reducible to complex classes of physical objects and events.

2) Mental terms describe a different causal domain (Searle). My reading of Searle is that he sees there being different domains of causality – the physical causality of the neurons or transistors or waterpipes, and the mental or intentional causality at work in human cognition. He says that “...no purely formal model will ever be by itself sufficient for intentionality because the formal properties have by themselves no causal powers except the power to produce the next step in the formalism when the machine is running.” As I have said several times above, I find this form of dualism confusing. If someone can correctly predict the future physical events in my brain and body based on a formal model of physical causality, and can also predict my future actions on the basis of intentional causality, there must be some magical connection that guarantees they will be consistent.

3) All causal explanations are interpretations, and we can have interpretations in separate domains (Winograd, influenced by Maturana, Flores, etc.). I can understand Searle’s dualism in the context of a larger shift of orientation towards causality as a mode of interpretation. In trying to understand the regularities in our experience, we interpret it as being made up of events connected by causal relationships. We can have interpretations in several domains (e.g. physical and mental) which have different sets of distinctions by which phenomena are categorized. An explanation is a description of a class of phenomena that in some way enables one to anticipate future experience. To the extent that this “generative” capacity is relevant to a particular domain, we can say that the explanation is valid in that domain. Searle’s view that no formal model will be sufficient can be interpreted as the statement that there is no understandable correlation between the domain of phenomena that we think of as intentional and those that we think of as formal definitions of states of physical objects. If we were to substitute the word “systematic”, or some such for “understandable”, we would be led into some kind of non-physicalist view. My view is that the key is in “understandability”. The

question is not whether there is ultimately some physical/mental correspondence, but whether the nature of that correspondence is such that it can be used for explanation, where explanation is a phenomenon of human discourse.

From this point of view, the question as to whether AI can ever explain cognition becomes both empirical and relative to purpose. To the extent that one can create computational abstractions that characterize the mechanisms of the brain, and to the extent that the operation of these mechanisms can be used for explanation in characterizing mental events, the AI work is “explanatory”. No AI explanation will ever “completely” capture what goes on the mind. No explanation ever “completely” describes anything, except with respect to a fixed set of purposes and background of assumptions about what is relevant.

As with the use of mental terms for machines, the most important issue here lies in the phenomena of the use of language. The question is not “Can AI explain cognition”, but “What is a person doing in using an AI explanation for cognition.” Explanation is not a relationship between formal objects and the world, but a language act by a speaker who is explaining something to a hearer (who, as usual, may be himself). Once again, we must take into account the orientation that is carried by a mode of speaking. Once again, the orientation towards someone as an autonomous agent is at stake. In our everyday discourse we have two very different modes of explanation. For people, we use terms like “choice”, “desire”, and “responsibility”, which go along with a web of attitudes and beliefs that are consistent with respecting the autonomy and independence of the individual. In talking of simple physical objects, we use terms that emphasize the predictable causal relationships among physical events.

In offering a mechanistic explanation of human action (just as in applying mental terms to machines), we are crossing from one orientation to the other. This is not true of AI alone, but of any form of psychology that provides models for mechanisms that determine behavior. To the degree we explain someone’s actions in terms of mechanisms (whether they be behavioristic, computational, or even Freudian) we are adopting a stance towards that person as a mechanism. We treat the person not as an autonomous agent, but as a device whose functioning is determined in an inevitable way by its state and inputs.

As I emphasized above, the issue is not whether a person “really is” an autonomous agent or a deterministic machine. In choosing to apply a mechanistic explanation (be it from AI or elsewhere), we are making it so. As with the use of mental terms for machines, I have ambivalent feelings about making use of this kind of explanation. The ultimate social and moral consequences of treating human beings primarily as mechanisms are repugnant. But this extreme does not follow inexorably from the use of mechanisms as one domain of interpretation for human action. There are many contexts in which, for given purposes in an appropriately shared background, it is useful. In psychiatry, for example, there is an ongoing tension between a

mode of understanding in which the therapist views the patient as a mechanism to be explained causally and modified to change behavior, and a mode in which there is a conversation between therapist and patient which is oriented towards mutual recognition of the patient's autonomy.

It would be futile to argue against all uses of the mechanistic mode of understanding for people. Not only would it be counter to a large part of our tradition, but it would impoverish our ability to understand, by eliminating one potentially relevant way of thinking. What is important is to be aware of the ways in which we make use of different kinds of explanation, and of the effects it can have. The trend in our modern society is to emphasize the mechanistic domain of explanation, and it is important to take responsibility for the ways in which our own uses of language (particularly as public interpreters of cognition) affect that direction.

[1] Winograd, Terry (1980), "What does it mean to understand language?," *Cognitive Science* 4:3 (July-Sept 1980) 209-242. Reprinted in D. Norman (ed.), **Perspectives on Cognitive Science**, Ablex and Erlbaum Associates, 1981, 231-264.

[2] Winograd, Terry (1987), "Cognition, attunement and modularity," *Mind and Language*, 1987, 97-103.

[3] Winograd, Terry and Flores, Fernando, **Understanding Computers and Cognition: A New Foundation for Design**, (220 pp.) Norwood, NJ: Ablex, 1986. Paperback issued by Addison-Wesley, 1987.

## B Appendix: Top Ten List

I tried to think of what a curious reader might like to ask me after all the effort I have gone through to read so many articles. One of the questions could be, “Which articles were the best, in your opinion?” In the tradition of David Letterman, I present the following:

### “Top ten list of best Chinese room publications”

- Harnad 1989
- Churchland and Churchland 1990 (read together with Searle 1990)
- Rey 1986
- Maloney 1987
- Copeland 1993
- Haugeland 1980 (specifically, the silicon for neurons argument)
- Searle 1982b (read together with Dennet 1982)
- Jacquette 1989
- Block 1995
- Dyer 1994

## References

- [1] R. P. Abelson, (1980) “Searle’s argument is just a set of Chinese symbols”, *Behavioral and Brain Sciences*, **3**: 424-425.
- [2] D. Anderson, (1987) “Is the Chinese Room the Real Thing?” *Philosophy*, **62**: 389-393.
- [3] K. Baldner, (1990) “Transcendental Idealism from the Chinese Room: Does God Speak Chinese?” *Proceedings of the Heraclitean Society*, **15**: 9-15.
- [4] H. Ben-Yami, (1995) “A Note on the Chinese Room”, *Synthese*, **95**: 169-172.
- [5] N. Block, (1980) “What intuitions about homunculi don’t show”, *Behavioral and Brain Sciences*, **3**: 425-426.
- [6] N. Block, (1995) “The Mind as the Software of the Brain”, in E. E. Smith and D. N. Osherson, editors, (1995) **Thinking: An Invitation to Cognitive Science**, 2nd edition, Volume 3, The MIT Press, Cambridge, MA, pp. 377-425.
- [7] M. Boden, (1990) **The Philosophy of Artificial Intelligence**, Oxford University Press, Oxford, UK.
- [8] B. Bridgeman, (1980) “Brains + programs = minds”, *Behavioral and Brain Sciences*, **3**: 427-428.
- [9] T. W. Bynum, (1985) “Artificial Intelligence, Biology, and Intentional States”, *Metaphilosophy*, **16**: 354-377.
- [10] P. Cam, (1990) “Searle on Strong AI”, *Australasian Journal of Philosophy*, **68**: 103-108.
- [11] L. R. Carlton, (1984) “Programs, Language Understanding, and Searle”, *Synthese*, **59**: 219-230.
- [12] D. J. Chalmers, (1996) **The Conscious Mind: In Search of A Fundamental Theory**, Oxford University Press, Oxford, UK.
- [13] D. Cole, (1984) “Thought and Thought Experiments”, *Philosophical Studies*, **45**: 431-444.
- [14] D. Cole, (1991a) “Artificial Minds: Cam on Searle”, *Australasian Journal of Philosophy*, **69**: 329-333.
- [15] D. Cole, (1991b) “Artificial Intelligence and Personal Identity”, *Synthese*, **88**: 399-417.
- [16] D. Cole, (1994) “The Causal Powers of CPUs”, in [22], pp. 139-156.
- [17] B. J. Copeland, (1993) **Artificial Intelligence: An Introduction**, Blackwell Publishers, Inc., Oxford, UK. Chapter 6, “The curious case of the Chinese room,” is dedicated to Searle’s argument.
- [18] B. J. Copeland, (1995) “The Curious Case of the Chinese Gym”, *Synthese*, **95**: 173-186.
- [19] T. Cuda, (1985) “Against Neural Chauvinism”, *Philosophical Studies*, **48**: 111-127.
- [20] R. Cummins, (1983) **The Nature of Psychological Explanation**, The MIT Press, Cambridge, MA.
- [21] A. C. Danto, (1980) “The use and mention of terms and the simulation of linguistic understanding”, *Behavioral and Brain Sciences*, **3**: 427-428.
- [22] E. Dietrich, editor, (1994a) **Thinking Computers and Virtual Persons**, Academic Press, Inc., San Diego, CA.
- [23] E. Dietrich, (1994b) “Thinking Computers and the Problem of Intentionality”, in [22], pp. 3-34.

- [24] D. C. Dennett, (1981) “Where am I?” in [48], pp. 217-229.
- [25] D. C. Dennett, (1980) “The milk of human intentionality”, *Behavioral and Brain Sciences*, **3**: 428-430.
- [26] D. C. Dennett, (1982) “The myth of the computer: An exchange”, *New York Review of Books*, Vol. 29, No. 11, pg. 56.
- [27] R. Double, (1983) “Searle, Programs, and Functionalism,” *Nature and System*, **5**: 107-114.
- [28] R. Double, (1984) “Reply to Fields,” *Nature and System*, **6**: 55-57.
- [29] M. Dyer, (1994) “Intentionality and Computationalism: *Minds, Machines, Searle, and Harnad*”, in [22], pp. 173-195.
- [30] J. C. Eccles, (1980) “A dualist-interactionist perspective”, *Behavioral and Brain Sciences*, **3**: 430-431.
- [31] T. Edelson, (1982) “Simulated understanding: Making the example fit the question”, *Behavioral and Brain Sciences*, **5**: 338-339.
- [32] J. A. Feldman, “Robots with Common Sense?”, in [54], pp. 65-72.
- [33] J. H. Fetzer, editor, (1988) **Aspects of Artificial Intelligence**, Kluwer Academic Publishers, Norwell, MA.
- [34] C. A. Fields, (1984) “Double on Searle’s Chinese Room,” *Nature and System*, **6**: 51-54.
- [35] J. A. Fisher, (1988) “The Wrong Stuff: Chinese Rooms and the Nature of Understanding,” *Philosophical Investigations*, **11**: 279-299.
- [36] O. Flanagan, (1991) **The Science of the Mind**, second edition, The MIT Press, Cambridge, MA.
- [37] J. A. Fodor, (1980a) “Methodological solipsism considered as a research strategy in cognitive psychology”, *Behavioral and Brain Sciences*, **3**: 63-110.
- [38] J. A. Fodor, (1980b) “Searle on what only brains can do”, *Behavioral and Brain Sciences*, **3**: 431-432.
- [39] J. A. Fodor, (1981) **RePresentations**, The MIT Press, Cambridge, MA.
- [40] G. G. Globus, (1991) “Deconstructing the Chinese Room”, *Journal of Mind and Behavior*, **12**: 377-392.
- [41] P. Hanna, (1985) “Causal Powers and Cognition”, *Mind*, **94**: 53-63.
- [42] S. Harnad, (1989) “Minds, Machines and Searle”, *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 1, No. 1, pp. 5-25, January.
- [43] S. Harnad, (1990) “The symbol grounding problem”, *Physica D*, **42**: 335-346.
- [44] R. J. Harvey, (1985) “On the nature of programs, simulations, and organisms”, *Behavioral and Brain Sciences*, **8**: 741-766.
- [45] J. Haugeland, (1980) “Programs, causal powers, and intentionality”, *Behavioral and Brain Sciences*, **3**: 432-433.
- [46] J. Haugeland, editor, (1981) **Mind Design**, The MIT Press, Cambridge, MA.
- [47] J. Haugeland, editor, (1997) **Mind Design II**, The MIT Press, Cambridge, MA.

- [48] D. R. Hofstadter and D. C. Dennett, editors, (1981) **The Mind's I: Fantasies and Reflections on Mind and Soul**, Basic Books, NY.
- [49] D. R. Hofstadter, (1980) "Reductionism and religion", *Behavioral and Brain Sciences*, **3**: 433-434.
- [50] *Informatica*, An International Journal of Computing and Informatics, Volume 19, Number 4, November 1995. Special Issue: Mind – Computer, Were Dreyfus and Winograd right?
- [51] D. Jacquette, (1989) "Adventures in the Chinese Room," *Philosophy and Phenomenological Research*, Vol. 49, No. 4, pp. 605-623, June.
- [52] N. Jähren, (1990) "Can Semantics be Syntactic?" *Synthese*, **82**: 309-329.
- [53] B. Libet, (1980) "Mental phenomena and behavior", *Behavioral and Brain Sciences*, **3**: 434.
- [54] V. Lifschitz, editor, (1991) **Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy**, Academic Press, Inc., San Diego, CA.
- [55] W. G. Lycan, (1980) "The functionalist reply (Ohio State)", *Behavioral and Brain Sciences*, **3**: 434-435.
- [56] W. G. Lycan, (1987) **Consciousness**, The MIT Press, Cambridge, MA.
- [57] J. C. Maloney, (1987) "The Right Stuff", *Synthese*, **70**: 349-372.
- [58] J. C. Marshall, (1980) "Artificial Intelligence – the real thing?", *Behavioral and Brain Sciences*, **3**: 435-437.
- [59] G. Maxwell, (1980) "Intentionality: Hardware, not software", *Behavioral and Brain Sciences*, **3**: 437-438.
- [60] J. McCarthy, (1980) "Beliefs, machines, and theories", *Behavioral and Brain Sciences*, **3**: 435.
- [61] D. McDermott, (1982) "Minds, brains, programs, and persons", *Behavioral and Brain Sciences*, **5**: 339-341.
- [62] E. W. Menzel Jr., (1980) "Is the pen mightier than the computer?", *Behavioral and Brain Sciences*, **3**: 438-439.
- [63] M. Minsky, (1980) "Decentralized minds", *Behavioral and Brain Sciences*, **3**: 439-440.
- [64] J. H. Moor, (1988) "The Pseudorealization Fallacy and the Chinese Room Argument", in [33], pp. 35-53.
- [65] T. Nagel, (1974) "What is it like to be a bat?", *Philosophical Review*, **83**: 435-451.
- [66] T. Natsoulas, (1980) "The primary source of intentionality", *Behavioral and Brain Sciences*, **3**: 440-441.
- [67] N. Newton, (1988) "Machine Understanding and the Chinese Room", *Philosophical Psychology*, **1**: 207-215.
- [68] K. Obermeier, (1983) "Wittgenstein on Language and Artificial Intelligence: The Chinese-Room Thought Experiment Revisited", *Synthese*, **56**: 339-349.
- [69] K. Pfeifer, (1992) "Searle, Strong AI, and Two Ways of Sorting Cucumbers", *Journal of Philosophical Research*, **17**: 347-350.
- [70] R. Puccetti, (1980) "The chess room: further demythologizing of strong AI", *Behavioral and Brain Sciences*, **3**: 441-442.

- [71] Z. W. Pylyshyn, (1980) “The ‘causal power’ of machines”, *Behavioral and Brain Sciences*, **3**: 442-444.
- [72] Z. W. Pylyshyn, (1985) **Computation and Cognition**, The MIT Press, Cambridge, MA, pp. 43-45.
- [73] H. Rachlin, (1980) “The behaviorist reply (Stony Brook)”, *Behavioral and Brain Sciences*, **3**: 444.
- [74] H. Rachlin, (1982) “Minds, pains, and performance”, *Behavioral and Brain Sciences*, **5**: 341.
- [75] W. J. Rapaport, (1986a) “Discussion: Searle’s Experiments with Thought”, *Philosophy of Science*, **53**: 271-279.
- [76] W. J. Rapaport, (1986b) “Philosophy, Artificial Intelligence, and the Chinese-Room Argument”, *Abacus*, **3**: 6-17.
- [77] W. J. Rapaport, (1988a) “To Think or Not to Think”, *Noûs*, **22**: 585-609.
- [78] W. J. Rapaport, (1988b) “Syntactic Semantics”, in [33], pp. 81-131 (reprinted in [22], pp. 225-273).
- [79] G. Rey, (1986) “What’s Really Going on in Searle’s ‘Chinese Room’”, *Philosophical Studies*, **50**: 169-185.
- [80] M. Ringle, (1980) “Mysticism as a philosophy of artificial intelligence”, *Behavioral and Brain Sciences*, **3**: 444-445.
- [81] R. Rorty, (1980) “Searle and the special powers of the brain”, *Behavioral and Brain Sciences*, **3**: 445-446.
- [82] L.-M. Russow, (1984) “Unlocking the Chinese Room”, *Nature and System*, **6**: 221-227.
- [83] J. Samet, (1982) “Understanding and integration”, *Behavioral and Brain Sciences*, **5**: 341-342.
- [84] S. F. Savitt, (1982) “Searle’s demon and the brain simulator”, *Behavioral and Brain Sciences*, **5**: 342-343.
- [85] R. C. Schank, (1980) “Understanding Searle”, *Behavioral and Brain Sciences*, **3**: 446-447.
- [86] J. R. Searle, (1980a) “Minds, brains, and programs”, *Behavioral and Brain Sciences*, **3**: 417-424.
- [87] J. R. Searle, (1980b) “Intrinsic Intentionality”, *Behavioral and Brain Sciences*, **3**: 450-457.
- [88] J. R. Searle, (1982a) “The Chinese room revisited”, *Behavioral and Brain Sciences*, **5**: 345-348.
- [89] J. R. Searle, (1982b) “The myth of the computer”, *New York Review of Books*, Vol. 29, No. 7, pp. 3-6.
- [90] J. R. Searle, (1982c) “The myth of the computer: An exchange”, *New York Review of Books*, Vol. 29, No. 11, pp. 56-57.
- [91] J. R. Searle, (1984) *Minds, Brains and Science*, 2nd edition, Harvard University Press, Cambridge, MA.
- [92] J. R. Searle, (1989) “Reply to Jacquette,” *Philosophy and Phenomenological Research*, Vol. 49, No. 4, pp. 701-708, June.
- [93] J. R. Searle, (1990) “Is the Brain’s Mind a Computer Program?”, *Scientific American*, Vol. 262, No. 1, pp. 26-31, January.

- [94] J. R. Searle, (1992) **The Rediscovery of the Mind**, MIT Press, Cambridge, MA.
- [95] J. R. Searle, (1997) "Consciousness and the Philosophers", *New York Review of Books*, Vol. 44, No. 4, pp. 43-48, March 6.
- [96] A. Seidal, (1989) "Chinese Rooms, A, B, and C", *Pacific Philosophical Quarterly*, **70**: 167-173.
- [97] R. Sharvy, (1983) "It Ain't the Meat, It's the Motion", *Inquiry*, **26**: 125-131.
- [98] R. Sharvy, (1985) "Searle on Programs and Intentionality", *Canadian Journal of Philosophy, Supplementary Volume II*, pp. 39-54.
- [99] A. Sloman and M. Croucher, (1980) "How to turn an information processor into an understander", *Behavioral and Brain Sciences*, **3**: 447-448.
- [100] A. Sloman, (1985) "What enables a machine to understand?", *Proceedings of the International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, Inc., Los Altos, CA, pp. 995-1001.
- [101] W. E. Smythe, (1980) "Simulation games", *Behavioral and Brain Sciences*, **3**: 448-449.
- [102] W. E. Smythe, (1982) "Rule following and rule reduction", *Behavioral and Brain Sciences*, **5**: 343-344.
- [103] K. Sterelny, (1990) **The Representational Theory of Mind**, Basil Blackwell Publisher Limited, Oxford, UK.
- [104] P. Thagard, (1986) "The Emergence of Meaning: How to Escape Searle's Chinese Room", *Behaviorism*, **14**: 139-146.
- [105] D. O. Walter, (1980) "The thermostat and the philosophy professor", *Behavioral and Brain Sciences*, **3**: 449.
- [106] T. Weis, (1990) "Closing the Chinese Room", *Ratio*, **III**: 165-181.
- [107] R. Wilensky, (1980) "Computers, cognition and philosophy", *Behavioral and Brain Sciences*, **3**: 449-450.
- [108] Y. Wilks, (1982) "Searle's straw men", *Behavioral and Brain Sciences*, **5**: 344-345.
- NOTE: THE REST OF THE REFERENCES DO NOT CONTAIN ANY EXPLICIT MENTION OF SEARLE'S CHINESE ROOM (THE PREVIOUS REFERENCES ALL DO MENTION THE CHINESE ROOM AT SOME POINT).**
- [109] M. H. Bickhard, (1993) "Representational content in humans and machines", *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 5, No. 4, pp. 285-333, October-December.
- [110] N. Block, (1978) "Troubles with Functionalism", as reprinted in [140], pp. 211-228.
- [111] R. A. Brooks, (1997) "Intelligence without Representation", in [47], pp. 395-420. This paper is a slightly modified version of his 1991 paper by the same title in *Artificial Intelligence* **47**: 139-159.
- [112] A. Church, (1936) "An unsolvable problem of elementary number theory", *American Journal of Mathematics*, **58**: 345-363.
- [113] T. Cormen, C. Leiserson and R. Rivest, (1990) **Introduction to Algorithms**, The MIT Press, Cambridge, MA.
- [114] H. L. Dreyfus, (1986) **What Computers Can't Do**, 2nd edition, Harper & Row, NY.
- [115] H. L. Dreyfus, (1992) **What Computers Still Can't Do**, The MIT Press, Cambridge, MA, pg. ix.

- [116] H. L. Dreyfus, (1992) *ibid.*, pg. xiii.
- [117] H. L. Dreyfus and S. E. Dreyfus, (1986) **Mind Over Machine**, The Free Press, NY.
- [118] H. B. Enderton, (1972) **A Mathematical Introduction to Logic**, Academic Press, Inc., San Diego, CA.
- [119] As appeared on the web page <http://wwwformal.stanford.edu/human-level/> on 24 May 1997.
- [120] D. B. Lenat and E. A. Feigenbaum, (1991) "On the thresholds of knowledge", *Aritificial Intelligence* **47**: 185-250.
- [121] D. Lewis, (1972) "Psychophysical and Theoretical Identifications", as reprinted in [140], pp. 204-210.
- [122] Henry R. Lewis and Christos H. Papadimitriou, (1981) **Elements of The Theory of Computation**, Prentice-Hall Inc., Upper Saddle River, NJ, pg. 223.
- [123] J. McCarthy, (1979) "Ascribing Mental Qualities to Machines," in M. Ringle, editor, (1979) **Philosophical Perspectives in Artificial Intelligence**, Humanities Press, pp. 161-195.
- [124] J. McClelland and D. Rumelhart, editors, (1986) **Parallel Distributed Processing**, The MIT Press, Cambridge, MA.
- [125] H. McGuire, (1992) "How to Read Winograd's and Flores's *Understanding Computers and Cognition*", Center for the Study of Language and Information, Report No. CSLI-92-162, Stanford, CA, pg. 23. The call number at Green Library at Stanford University is P121.P295 M.27
- [126] I owe this observation to Jose Meseguer, who pointed it out to me in a review of an earlier draft of this paper.
- [127] M. Minsky, (1986) **The Society of Mind**, Simon and Schuster, NY.
- [128] J. von Neumann, (1958) **The Computer and the Brain**, Yale University Press, New Haven, CT.
- [129] A. Newell and H.A. Simon, (1958) "Heuristic Problem Solving: the Next Advance in Operations Research", *Operations Research*, **6**: pg. 6.
- [130] A. Newell and H.A. Simon, (1961) "GPS, a Program that Simulates Human Thought", in E.A. Feigenbaum and J. Feldman, editors, (1963) **Computers and Thought**, McGraw-Hill Inc., NY, pp. 279-293.
- [131] A. Newell and H.A. Simon, (1976) "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the ACM*, Vol. 19, No. 2, pp. 113-126, March.
- [132] A. Newell, (1980) "Physical Symbol Systems", *Cognitive Science* **4**: 135-183.
- [133] A. Newell, (1982) "The Knowledge Level", *Aritificial Intelligence* **18**: 87-127.
- [134] A. Newell, (1990) **Unified Theories of Cognition**, Harvard University Press, Cambridge, MA.
- [135] N. J. Nilsson, (1995) "Eye on the Prize", *AI Magazine*, pp. 9-17, Summer.
- [136] D. Parfit, (1971) "Personal Identity", *Philosophical Review*, Vol. 80, No. 1.
- [137] H. Putnam, (1967) "The Nature of Mental States", as reprinted in [140], pp. 197-203.
- [138] H. Putnam, (1991) **Representation and Reality**, The MIT Press, Cambridge, MA.
- [139] H. Putnam, (1994) **Words & Life**, Harvard University Press, Cambridge, MA.

- [140] D. M. Rosenthal, (1991) **The Nature of Mind**, Oxford University Press, Oxford, UK.
- [141] C. H. Roth, (1985) **Fundamentals of Logic Design**, West Publishing Co., St. Paul, MN.
- [142] S. J. Russell and P. Norvig, (1995) **Artificial Intelligence: A Modern Approach**, Prentice-Hall Inc., Upper Saddle River, NJ, pg. 10.
- [143] R. C. Schank and R. P. Abelson, (1977) **Scripts, Plans, Goals and Understanding**, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.
- [144] H.A. Simon, (1965) **The Shape of Automaton: For Men and Management**, Harper & Row, NY.
- [145] H. Simon, (1981) **The Sciences of the Artificial**, 2nd edition, The MIT Press, Cambridge, MA.
- [146] A. Sloman, (1981) “Why robots will have emotions”, *Proceedings of the International Joint Conference on Artificial Intelligence*, American Association for Artificial Intelligence, Menlo Park, CA, pp. 197-202.
- [147] B. C. Smith, (1991) “The owl and the electric encyclopedia”, *Artificial Intelligence* **47**: 251-288.
- [148] P. Smolensky, (1988) “On the proper treatment of connectionism”, *Behavioral and Brain Sciences*, **11**: 1-74.
- [149] A. M. Turing, (1950) “Computing Machinery and Intelligence”, *Mind*, Vol. LIX, No. 236.
- [150] At least, very few philosophy authors among the references in this paper ever say or imply that they think that the Turing Test is adequate. As I recall, all those who do say something (about five or six) speak out against the Turing Test. Perhaps this apparent consensus is due to the demise of behaviorism earlier in this century.
- [151] T. van Gelder, (1997) “Dynamics and Cognition”, in [47], pp. 421-450.
- [152] T. Winograd and F. Flores, (1986) **Understanding Computers and Cognition**, Ablex Publishing Corp., Norwood, NJ. For an excellent report to help in understanding the approach of this book see [125].
- [153] A. J. Ayer, (1978) in an interview with Brian Magee, in Brian Magee, editor, *Men of Ideas*, BBC, p. 131. Later in the interview, Ayer did say that he still believes “in the same general approach.”
- [154] A. Church, (1949) “Review of A. J. Ayer’s *Language, Truth, and Logic*”, *Journal of Symbolic Logic*, **14**: pp. 52-53.
- [155] O. Hanfling, (1981) *Logical Positivism*, Basil Blackwell Publisher Limited, Oxford, UK.
- [156] J. Passmore, (1972) “Logical Positivism”, in P. Edwards, editor, (1972) *The Encyclopedia of Philosophy*.
- [157] J. O. Wisdom, (1963) *Mind*, p. 335.