To the Graduate Council:

I am submitting herewith a thesis written by Kevin Martin Roddy entitled "A Critical Analysis of John Searle's Arguments Against Strong AI." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Arts, with a major in Philosophy.

John E. Nolt

John E. Nolt, Major Professor

We have read this thesis
and recommend its acceptance:

Kathleen Emmitt

Richard E. Aquila

Accepted for the Council:

Carl Minkel

Vice Provost
and Dean of the Graduate School

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master's degree at The University of Tennessee, Knoxville, I agree that the Library shall make it available to borrowers under rules of the Library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of the source is made.

Permission for extensive quotation from or reproduction of this thesis may be granted by my major professor, or in his absence, by the Head of Interlibrary Services when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Signature Kevin M. Rooks

Date 7-5-88

A CRITICAL ANALYSIS OF JOHN SEARLE'S ARGUMENTS AGAINST STRONG AI

A Thesis

Presented for the

Master of Arts

Degree

The University of Tennessee, Knoxville

Kevin Martin Roddy

August 1988

For my friend, E. J. Doherty.

ACKNOWLEDGMENTS

I would like to thank all of the members of my thesis committee, especially Professor John E. Nolt, for their many helpful suggestions and criticisms concerning this thesis. Without their help I would not have been able to successfully complete this thesis.

ABSTRACT

In the article "Minds, Brains, and Programs" John Searle attempts to refute the two claims made by what has come to be known as strong artificial intelligence (strong AI). The first claim is that the "appropriately programmed computer...can be literally said to *understand* and have other cognitive states." The second claim is that "because the computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations."

By the use of a counter-example which has come to be known as the Chinese Room example, and Searle's own criterion for understanding, Searle attempts to show the first claim of strong AI to be false, and the second claim to be implausible.

Chapter I sets out Searle's arguments against the claims of strong AI. Chapters II, III, and IV deal with several critical replies to the Chinese Room example. Chapter V comes to the conclusion that Searle's purported counter-example fails to refute the two claims of strong AI, and that there may be another criterion for understanding than that used by Searle.

TABLE OF CONTENTS

CHAPTER I

SUMMARY OF SEARLE'S ARGUMENTS AGAINST STRONG AI

In his article "Minds, Brains, and Programs", John Searle attempts to refute the two claims of what he calls strong AI (Artificial Intelligence). Searle defines the first claim of strong AI as: A) The belief that "the approriately programmed computer... can be literally said to *understand* and have other cognitive states". Searle goes on to say that those in favor of this first claim of strong AI also make the second claim that: B) "...because the computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations." (Searle, 1980; p. 417) These two claims of strong AI have to be distinguished from what Searle calls weak AI. In this article, Searle does not very clearly say exactly what the claims of weak AI are. He does say that:

> According to weak AI, the principle value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion than before. (Searle, 1980; p. 417)

This is not a very helpful explanation of weak AI. However, Searle does in effect say that the main difference between strong AI and weak AI is that it is only the former that claims that the appropriately programmed computer literally has cognitive states and those programs are the explanations of those cognitive states. Whatever exactly the claims of

weak AI are, they cannot be these two claims since Searle does say that he has "no objections to the claims of weak AI, at least as far as this article is concerned." (Searle, 1980; p. 417)

In his article, he sets himself the task of showing that neither claim A) nor claim B) given above, is justified. He attempts to do this by means of a *Gedankenexperiment* which has come to be known as the Chinese Room example. Searle describes the thought experiment in the following manner: (I quote the *Gedankenexperiment* in full because of its central importance in this thesis.)

> Suppose that I'm locked in a room and suppose that I'm given a large batch of Chinese writing. Suppose furthermore, as is indeed the case, that I know no Chinese either written or spoken, ... 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. 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 I am 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.' To complicate the story a little bit, imagine that these people also give me stories in English which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. 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 point 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 indistinguishable from those of native Chinese speakers. Nobody looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native speaker of English. From the external point of view, from the point of view of someone reading my 'answers,' the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program. (Searle, 1980; pp. 417-418)

This thought experiment is Searle's initial example to try to show that neither of strong AI's claims is correct. Searle believes that in the Chinese room he is in no better and no worse a position than any computer running any similar program. In both cases, Searle concludes, the level of understanding is zero. According to Searle, in both cases all that is happening is the formal manipulation of meaningless symbols. This by itself, continues Searle, is not sufficient in either case to produce understanding. He therefore concludes that strong AI's first claim, (that is, that a computer, or Searle, merely in virtue of running a purely formal program will understand its output) is shown to be false. Its second claim (that the machine and program explain understanding) has not been shown to be false, though it has been shown that the machine's program is not a *sufficient* explanation of understanding (since something else necessary for an understanding of Chinese is apparently missing in the program running the Chinese Room, and in any similar program run by a computer). Searle admits that he has not shown that such a program is not even a *necessary* part of the explanation of human understanding, however, he goes

on to say that the above example at least suggests that "...the computer program is irrelevant to my understanding of the story" or any other cognitive state for that matter (Searle, 1980; p. 418).

Much of the rest of Searle's article is spent revising the initial counter-example to the claims of strong AI, so that it can satisfactorily meet some objections to it. In the next few chapters I will be dealing with some of the most important of these replies, and the revised thought experiments. For now, I just want to say that Searle believes that his conclusions based on the first *Gedankenexperiment*, also apply to the revised forms of it. Searle believes that this is true because no matter how you change the example without changing the basic premise, that is, that all that is happening is the formal manipulation of meaningless symbols, no form of understanding could possibly be produced. To state this in a succinct way:

1. A system that merely involves the formal manipulation of symbols is not capable of understanding.

One such system is Searle shifting bits of paper according to a prescribed set of rules with Chinese symbols written on them. A computer running the appropriate program is another such system.

This is the negative aspect of Searle's thesis, (that is, to show that the claims of strong AI are false). On the postive side, Searle believes that there is at least one special kind of "machine" with certain kinds of

"powers" which make it capable of understanding. The "machine" he is talking about is a human being, and, as stated earlier, Searle does not believe that following rules, or manipulating symbols in a purely formal way is how a human being understands anything. According to Searle, it may very well be the case that the specific biochemistry of the human brain is what is responsible for our having certain kinds of cognitive states such as understanding. It is this biochemistry that the human brain has (which is lacking in our present-day computers) which is causally responsible for producing these cognitive states in us. This is Searle's own solution to the mind-body problem, which he refers to as "biological naturalism" (see Searle, 1984). Searle believes that there really are such things as beliefs, desires, and understanding in this material universe, and that they "cannot be reduced to something else or eliminated by some kind of re-definition." (Searle, 1983; p. 262) (For example, Searle would not want to say that belief X is equivalent to the potential of a neuron, or group of neurons, to fire.) If Searle intends to hold on to these mentalistic terms, which have caused many philosophers numerous problems in their attempts to reconcile them with a belief in a material universe, he will have to be able to explain how they can be related to brains.

Biological naturalism is explained by Searle in the Epilogue to his book, Intentionality:

> The picture that I have been suggesting...is one according to which mental states are both *caused by* the operations of the brain and *realized in* the structure of the brain (and the rest of the central nervous system). (Searle, 1983; p. 268)

In Minds, Brains and Science, Searle tries to explain this causal relationship between the brain, with its specific biochemistry, and mental states, by the use of an analogy. Consider the relationship between H_2O molecules and the liquidity of water:

> The liquidity of water is both *caused by* the behavior of (the H_2O molecules), and at the same time is *realised in* the system that is made up of the (H_2O molecules). There is a cause and effect relationship, but at the same time the (liquidity of water is) ...just (a) higher level (feature) of the very system whose behavior at the micro level causes (that feature). (Searle, 1984; p. 21)

The important point that Searle is trying to make in the article is that since it is probably the specific biochemistry of the human brain that is causally responsible for these cognitive states in us, these cognitive abilities have probably little or nothing to do with either the formal manipulation of symbols, or the purely formal organization of the brain.

Regardless of whatever else these causal powers of the brain produce, they are, according to Searle, in all likelihood responsible for producing intentionality in us. Intentionality is defined by Searle in a footnote to his article as:

> ...that feature of certain mental states by which they are directed at or about objects and states of affairs in the world. Thus beliefs, desires, and intentions are intentional states... (Searle, 1980; p. 424)

According to Searle, we quite naturally ascribe these intentional states to human beings as well as to animals such as monkeys, cats and dogs (see Searle, 1980). The relevant intentional states in the human being--or animal--enable them to, for example, *see that* there is someone standing in front of them, or to *intend* to move their arms or legs, or to *understand* what is being said to them (see Searle, 1980). Without intentionality, without the ability to refer to the things that, for example, the symbols we use are intended to refer to, we wouldn't be able to know the meanings of the symbols that we use, and we wouldn't be able to have any understanding of them. Imagine that the leaves that have fallen from a tree spell out the words: GIVE THANKS TO GOD (I believe that Searle uses a similar example to make the same kind of point). Although it is highly unlikely that something like that would ever happen, if it did happen we would not think that the tree intended for its leaves to fall in that way, or that it knew the meanings of those words, or that it understood them. We would not think that this was anything other than a chance occurance, because we do not ascribe intentional states to trees. If a human being wrote out those words we would normally think that they intended to write them, knew what they meant, and understood them, because we believe that human beings have intentionality and intentional states. Thus Searle believes the following:

2. You can't refer to something without having intentionality.
3. You can't know the meaning of a thought or proposition without first being able to refer to what it refers to.
4. Without meaning and without intentionality you can't have understanding.

Even though Searle in the Chinese Room has all his abilities intact--including his intentional abilities--because he is only given the uninterpreted symbols of the Chinese language to play with, he is not able to use these abilities to come to any understanding of what he is dealing with. According to Searle, he is just like a computer which does not have any form of intentionality to begin with (and therefore no initial ability to understand) since his own intentional abilities apparently cannot change the fact that he is only dealing with the formal manipulation of meaningless symbols. If, on the other hand, a computer was an intentional system very like a human being, and did not merely formally manipulate meaningless symbols, then in all probability, it could understand at least certain things. But intentionality, and therefore the possibility of understanding, will never be possible in a computer, according to Searle, as long as the emphasis is only put on enabling the computer to run a purely formally specified program. As I have already pointed out, Searle believes that intentionality is a "...biological phenomenon and it is likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena." (Searle, 1980; p. 424) It would therefore seem that for a computer to have intentionality, it must first have hardware which exactly replicated the specific biochemistry of the human brain (or at least whatever is repsonsible for intentionality and understanding) so that it could have mental states such as intentionality, or understanding. Otherwise, it should have another kind of biochemistry that had the same causal powers as the human brain (to produce such things as intentionality and understanding). In that case, the computer's ability to understand would

have little or nothing to do with its running a program; rather, its ability to understand would be most importantly due to its brain having the causal power to produce intentionality and understanding, regardless of whether its brain was made from "gray matter", or some other substance with the same causal powers.

As I mentioned earlier, Searle presents his version of biological naturalism as a solution to the mind-body problem, as well as an attempt to explain what is necessary, and perhaps sufficient, for a system to have cognitive states such as understanding. This solution and explanation is of course not the only one which is presently popular: in his article, Searle not only attempts to refute the claims of strong AI, but also tries to refute the claims of functionalism, which has given strong AI much of its conceptual basis.

Functionalism is the latest in a series of philosophical theories of the mind which attempt to explain how we can think, understand, and have other cognitive states, without recourse to immaterial entities such as Cartesian minds. Very basically, functionalists, unlike biological naturalists, believe that it is not the *stuff* that a thing is made of, nor the biochemical structure instantiated in that stuff, which will or will not enable it to understand, but the *functional organization* of that stuff, i.e., how the various parts of that stuff are causally related to each other and to certain other causes and effects. Using Paul Churchland's description of functionalism: "According to *functionalism*, the essential or defining feature of any type of mental state is the set of causal relations

it bears to 1) Enviromental effects on the body, 2) other types of mental states, and 3) bodily behavior." (Churchland, 1984; p. 36) In effect this means that when discussing a mental state such as understanding, you cannot in general define it only as being the ability to produce the right response in the appropriate circumstances. For example, if you asked me the time of day, and I gave you the correct time in response, if I can be truly said to have understood what you had asked me, that would not be all there is to consider. It would only be a case of understanding if my response is related, in the appropriate ways, to other types of mental states, and to other instances of the same type. For example, the response would be related to my *belief* that you really want me to tell you the time, and I *know* that my watch is set at the correct time. Under ordinary conditions it would be very questionable to say that I understood your request if I told you the correct time, and then asked you the same question. In that case it was probably a mere fluke that I gave you the appropriate response, without understanding. Whether it be an angel, a human being, or a computer that we are discussing, the ability to understand will depend on having the right functional organization for that purpose. And that will mean not only having the appropriate bodily responses, but also the appropriate relations to other mental states.

Searle believes that functionalism's solution of the mind-body problem, and its explanation of what is sufficient for understanding and intentionality are wrong (see, for example, Searle, 1984). In the above explanation of functionalism, the biochemical structure of the stuff that the thing is made of really makes no difference, and it is the causal

relations/organization of the thing's parts, and how they function, that is important. As I have already pointed out, Searle believes that intentionality--and consequently, understanding--is dependent on the specific biological structure of the brain "...i.e., (its) chemical and physical structure...". This being the case, just having something, for example, a computer program, that has all the organizational and causal relations relevant to understanding, you do not thereby have what is necessary and perhaps sufficient to really bring about understanding in a computer. If functionalists were to claim that Searle did understand anything in the Chinese Room, then they would have to not only look at the external situation, but also the particular way in which Searle did the job. But according to Searle, all that you would see if you looked inside the room, would be his shuffling batches of symbols according to an already understood set of rules. All of that would not be, concludes Searle, enough to produce understanding of the Chinese symbols. "In strong AI (and in functionalism, as well) what matters are programs, and programs are independent of their realization in machines; indeed, as far as AI is concerned, the same program could be realized by an electronic machine, a Cartesian mental substance, or a Hegelian world spirit." (Searle, 1980; p. 423) Searle believes that if we accept the functionalist thesis, then we may even have to go as far as accepting the appropriately organized system of water pipes as having understanding (see Searle, 1980).

Functionalism and the claims of strong AI are not the only views that Searle tries to discredit in the article. The Turing Test (Turing,

1950), is also supposed to be shown to be wrong by means of the Chinese Room example.

Alan Turing in his famous article "Computing Machinery and Intelligence," provided a test whereby one would have to ascribe as much understanding to a computer as you would to a human being, if you could not distinguish between their responses (Turing, 1950). Turing initially plays what he calls the "Imitation Game" between a man, a woman, and an interrogator. "The object of the game for the interrogator is to determine which of the other two is the man and which is the woman." The game is set up so that the interrogator has no way of seeing or knowing which is the man, and which is the woman. His only means of communication with them is through a teleprinter that types their responses to his questions. The man's job in this game is to try and make the interrogator think that he is the woman. Because of the way the game is set up, the interrogator may not be able to distinguish between the man and the woman based solely on their responses to his questioning. Turing then asks the question: "What will happen when a (computer) takes the part of (the woman) in this game?" In this new situation, Turing concludes, if the interrogator cannot any more successfully distinguish between the responses of the computer and those of the man than he did when playing the game between the man and the woman, then he would have to ascribe as much intelligence/understanding, to the computer as to the man. It should be pointed out that the game is set up so that there is nothing the interrogator could not ask the computer. If the computer, for example, was only able to respond to questions concerning a specific story, then the

interrogator should be able to decide quite quickly which is the computer and which is the man. But if the computer could adequately respond to all sorts of questions, remarks, challenges, wishes and so on, then the interrogator is very much less likely to make the right indentifications; and, on that basis, concludes Turing, the cognitive abilities of the computer must be considered as good as those of the man based on that interrogation.

The Chinese Room is set up so that it is very similar to the situation in the Imitation Game: According to Searle, in the Chinese Room he does not understand the Chinese symbols, but he does understand the English ones. However, following Searle's argument, although there is a difference that he can see, there is no difference that can be detected by the outside observers since his responses in Chinese get so good that they are externally indistinguishable from those of a native speaker. But though they are indistinguishable, that does not guarantee that anything in Chinese is understood by Searle. Appropriate external behavior is not sufficient to guarantee understanding.

These, then, are essentially Searle's positions towards strong AI, functionalism, the Turing Test, and the possibility of understanding and other cognitive states in digital computers. Even though Searle is denying the claims of strong AI, he tries to make it clear that he does not want to say that no kind of machine can understand:

> "Could a machine think?" The answer is, obviously, yes. We are precisely such machines. "Yes, but could an artifact, a man-made

> machine, think?" Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it sufficiently like ours, again the answer to the question seems to be obviously 'yes'. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality and all the rest of it using chemical principles different from those human beings use. It is...an empirical question..."But could something think, understand, etc. *solely* by virtue of being a computer with the right sort of program?"...and the answer..is "no." (Searle, 1980; p. 422)

If some other machine other than a human machine were able to understand, it would not be--could not be, according to Searle--only because it had the right formal organization, and/or external behaviors. Its ability to understand and have other cognitive states would be because either its specific biochemistry exactly duplicated the biochemistry of our own brains (or at least that part of it necessary for intentionality), or because it had another kind of biochemistry that had the same causal powers as our own biochemistry (or at least those powers necessary for intentionality). As I have mentioned earlier, Searle attempts to meet objections to his initial example by revising it, and believes that even with these revisions his position is still sound. Some of these objections and revisions I will be getting to in the next chapter, but for now I want to make one final point. As I stated at the very beginning, these views initially appeared in Searle's article "Minds, Brains, and Programs." That article appeared along with a number of "Peer Commentaries" (some of which I will be discussing in the next chapter). The views expressed in the article have also been incorporated by Searle into his book Minds, Brains and Science (especially in Chapters 2 and 3). In that book, the initial Chinese Room example appears in very much the same form as it did in the

original article, along with the relevant conclusions from it. One important point that Searle does make explicit in the book, but not in the article, is what Searle might refer to as the "as if" nature of computers running programs. In Chapter 3 of the book Searle claims that computers do not in fact follow rules of any kind, though it is "as if" they do:

> To say that I am obeying the rule is to say that the meaning of that rule, that is, its semantic content, plays some kind of causal role in the production of what I actually do...*In the sense in which human beings follow rules...in that sense computers don't follow rules at all. They only act in accord with certain formal procedures*...we can speak *metaphorically* as if this were a matter of following rules. But in the *literal* sense in which human beings follow rules, they only act as if they were following rules. (Searle, 1984; p. 47)

Searle goes on to make the same point concerning any kind of information-processing a human being is likely to be doing, and how that is radically different from the kind of information-processing a computer may do. (This is relevant, since those in cognitive science will often talk as if the only thing people do when they think, or, more specifically, believe, know, wish, etc., is to process information in a certain way):

> To put it at its crudest: When people perform mental operations, they actually think, and thinking characteristically involves processing information of one kind or another. But there is another sense of information-processing in which there are no mental states at all. In these cases, there are processes which are *as if* there were some mental information-processing going on. (Searle, 1984; p. 49)

Searle goes on to claim that those involved in cognitive science have confused these two kinds of information-processing--as well as the

distinction between "as if" rule-following, and real rule-following--and believe that the computers they are discussing do information-processing in the first sense, when they can only do it in the second sense. For a computer to do real information-processing, and real rule-following, it would, using Searle's earlier conclusions, have to have a brain, not necessarily made of the same stuff as ours, but one that gave it the same causal powers as ours (or at least those powers necessary for intentionality) so that it could have cognitive states such as understanding.

CHAPTER II

SOME INITIAL PROBLEMS WITH THE CHINESE ROOM EXAMPLE

In the initial Chinese Room example it becomes clear that Searle is not merely attempting to imagine some kind of story-understander, but that he is also trying to imagine some kind of natural language-understander that will pass the Turing Test. One thing that I did not point out in the last chapter which is important is that Searle gives as an example of a computer with the appropriate program that purportedly does understand, a "story-understander" developed by Roger C. Schank and his colleagues at Yale (see Schank, 1984). According to Searle, this is just one particular example of the kind of a computer with the appropriate program that some in strong AI would claim to have understanding in the literal sense. An example of the kind of story that this computer deals with is the following: "A man went into a restaurant and ordered a hamburger. When the hamburger arrived, it was burnt to a crisp, and the man stormed out of the restaurant angrily without paying for the hamburger or leaving a tip." (Searle, 1980; p. 417) The computer is then asked questions such as: "Did the man eat the hamburger?" The answer that you would obviously give, and the answer that the computer would give, would be: "No, he did not."

The computer running this program is given what Schank calls a script. According to Schank, "Scripts are prepackaged sets of expectations, inferences, and knowledge that are applied in common situations, like a

blueprint for action without the details filled in." (Schank, 1984; p. 114) In the particular program I am describing, the script that would go along with this story would involve, for example, the fact that one does not normally eat food that is burnt--especially when you are paying for it in a restaurant--and so the expected response, given the above story and that kind of script, would be that the man did not eat the hamburger.

It is, therefore, initially this kind of computer "story-understander" that Searle has in mind when he first imagines himself in the Chinese Room. He starts off the example by using the Schankian terminology of "scripts," "story," and "questions (about the story)," but it soon becomes very clear that Searle is imagining not just a story-understander similar to the one in the above example, but something much more complex:

> 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 point of view of somebody outside the room...my answers to the questions are indistinguishable from those of native Chinese speakers. Nobody looking at my answers can tell that I don't speak a word of Chinese. (Searle, 1980; p. 418)

As Hoftstader says in the Peer Commentary published along with Searle's article: "(There) is an incredible jump in complexity--perhaps a millionfold increase if not more." (Hofstader, 1980; p. 433) This incredible jump in complexity is what is needed if the *Gedankenexperiment* is intended as a refutation of the reliability of the Turing Test, since the Turing Test is so devised that the questions and responses that can be a part of the test need not only be about simple stories, but with anything else

that you could discuss with a normal human being. But from the way that Searle had been talking prior to the *Gedankenexperiment*, his only intention at that point was to refute the claims of strong AI by the use of the *Gedankenexperiment*, and not the reliability of the Turing Test. A few pages further on in the article Searle does actually say that passing the Turing Test--and showing its inadequacy--was one of the accomplishments of the Chinese Room. However, it should be made clear that someone who accepts the claims of strong AI does not necessarily also have to accept the reliability of the Turing Test in determining whether or not something really has understanding (see Rey, 1986). Those who defend the claims of strong AI may also believe that they are not trying to create appropriately programmed computers that will pass the Turing Test. Searle, in the Chinese Room example, is supposedly a computer that has a program sufficient for passing the Turing Test, but those in strong AI may not believe that the creation of such a program is possible, regardless of whether or not it could actually have understanding. They may only believe that it is possible to create appropriately programmed computers with understanding at much simpler levels. They may also agree that passing tests for these simpler levels would also not show, without a doubt, that a computer really had understanding, but they may believe that there are other more reliable ways of deciding whether or not a computer really has understanding. Any *Gedankenexperiment* that may show the inadequacy of the Turing Test, does not necessarily imply the falsity of the claims of strong AI (Rey, 1986). Those in strong AI may agree with Searle that getting a computer to pass the Turing Test, if that is indeed possible, is not a sufficient reason for saying that that computer really

has understanding. But even if Searle has a good counter-example to the reliability of the Turing Test, that by itself does not imply that he also has a good counter-example to the claims of strong AI.

Schank himself believes that there are different levels of understanding, and the skills required for them, as well as believing that only the lower levels of understanding are going to be possible in appropriately programmed computers.

In his book Explanation Patterns, Schank describes three types and levels of understanding on what he calls the "spectrum of understanding". A simplified way in which Schank explains the differences between the three levels of understanding is to consider the various ways in which a joke could be understood:

> A computer understander that simply understood the joke, in that it could explain what had happened, would be understanding at the level of MAKING SENSE. A program that actually found the joke funny, to the extent that it could explain what expectations had been violated, what other jokes it knew that were like it that it was reminded of, and so on, would be understanding at the level of COGNITIVE UNDERSTANDING. Finally, a program that belly laughed because of how that joke related to its own particular experiences and really expressed a point a view about life that the program was only now realizing, would have understood at the level of COMPLETE EMPATHY. (Schank, 1986; p. 17)

Schank believes that some present-day computers, with the appropriate programs, have reached the level of making sense. At the making sense level, the computer, if given a news story, may be able to explain what happened, where it happened, and when it happened, using the information

that it has been given, but will not be able to explain *why* it happened in a sense that went beyond listing the facts that it already has, and the script(s) that it is using. An example of a computer run program that had achieved the making sense level, would be the kind that dealt with restuarant stories such as the one briefly described at the beginning of this chapter. At this making sense level, such a computer run program would be able to understand certain kinds of things about the story. The computer run program should be able to understand that people do not normally eat burnt hamburgers, or pay for them, but to only achieve this making sense level, it need not empathetically or creatively understand *why* people don't like to eat burnt hamburgers. (The computer could be simply *told* why people do not like burnt hamburgers, but that is not the kind of understanding that is intended here: The kind that is intended here is, according to Schank, only possible in a computer that had actually tasted burnt hamburgers, or something similar (Schank, 1986).) So although a story about restaurants could be understood in various ways, and at various levels, being able to understand at one level would not necessarily mean understanding at a higher level (since the skills and experiences required for each of them can be very different).

As well as believing that certain present-day appropriately programmed computers have reached the making sense level of understanding, Schank also believes that in the future it will be possible to achieve understanding in a computer at the cognitive level. The cognitive level of understanding means being creative in some way with the material that is given to you as information. At the cognitive level, it is

possible to make "generalizations, correlations and so on" that are not obviously derivable from the information given. Schank gives the following as an example of understanding at this level:

> input: a set of stories about airplane crashes, complete with data about the airplanes and the circumstances
>
> output: a conclusion about what may have caused the crash derived from a theory of the physics involved and an understanding of the design of airplanes, used in conjunction with algorithms that can do creative explanation. (Schank, 1986; p. 12)

If an appropriately programmed computer only with the making sense level of understanding were given the same input, it would probably not be able to do much more than summarise the stories, or perhaps translate them into another language.

It is the highest level of understanding--the complete empathy level--which Schank believes has caused all the fuss when people in AI claim that appropriately programmed computers can understand. Schank does not believe that this level of understanding is going to be possible in computers: "No machine will have undergone enough experiences, and related to them in the same way as you did, to satisfy you that it *really understands* you." (Schank, 1986; p. 15)

For a computer to understand at the complete empathy level, it must itself have undergone, pains, joys, sadness, excitement, etc., to understand these things in the ways that a human being can understand them. No story about plane crashes is going to affect a computer in the way that it can

affect us (the way that "human interest" stories can affect us), though it is claimed that it could possibly understand it at the cognitive level. At that level it would "coldly" correlate and generalize information about airplane crashes, without feeling personally involved in any way.

Although Schank's claim that some appropriately programmed computers can already *understand* at the making sense level is not going to be accepted by everyone--especially Searle--I believe that almost everybody will accept the three levels of understanding as at least a good general outline of the way in which, for example, the very same story could be understood. A simple story about restaurants could be understood at all three levels by the same person. But though understanding it at the complete empathy level would imply being able to understand it at the other two levels, being able to understand it at the making sense level would not imply either of the other two levels. Therefore, if it is possible to create some kind of understanding in a computer at, for example, the making sense level, that wouldn't imply that it could also understand at the complete empathy level or the cognitive level. As I mentioned earlier in this chapter, by the end of Searle's Chinese Room example he is only dealing with a program that is supposed to pass the Turing Test. A computer that is supposed to have understanding only at the making sense level will certainly not pass such a test, and nothing could, on the basis of that test, be said about whether or not it really had understanding. Although human beings normally understand more than one kind of topic, computers with the appropriate programs can be so devised that, if they understand anything at all, they understand one kind of

domain (stories about restuarants, for example). If they are able to make sense about one domain, that does not imply that they can make sense about any other, since they may only have information about one kind of domain. (It may be that the making sense level of understanding can be somehow achieved without being domain specific. But what those in strong AI have, for the most part, attempted to do is to create domain specific programs such as CYRUS which was developed by Schank and his colleagues at Yale (see Schank, 1984). (CYRUS dealt specifically with the professional and personal history of Cyrus Vance.))

Searle is in fact willing to accept this kind of distinction between different kinds of understanding, but he believes that even with this kind of qualification, and even if he stuck to story-telling programs in his example, that would not be sufficient to show that such programs could understand. His objection to much of what I have been saying so far in this chapter could be put in the following way: It really doesn't matter whether or not you qualify understanding by saying that it involves different levels of understanding. To understand anything at any level you must have intentionality and knowledge of the meanings of the symbols involved. None of that exists in either the case of a so-called "story-understander," or in the case of a more complex computer run program that could pass the Turing Test. So it doesn't matter whether I change the details of the example, no understanding is present at any level.

In fact, in the article itself, he makes a similar statement in reply to those who point out this fact that there are different levels of

understanding: "There are clear cases where 'understanding' applies and clear cases where it does not apply: and such cases are all I need for this argument." (Searle, 1980; p. 418) The import that this statement has considering what he has said about meaning and intentionality is that if you can show that something doesn't have intentionality and know the meanings of the symbols that it is using, you have shown that it cannot understand no matter what level of understanding you want to talk about.

On the other hand, if you can show that something does have understanding (or perhaps you only need to show that it is very likely that it has understanding) then you have, according to Searle's own requirements for understanding, also shown that it knows the meanings of the symbols that it is using, and does have intentionality. Searle believes that he has shown that the Chinese Room doesn't have a knowledge of the meanings of the symbols that it is using, and that it doesn't have intentionality. Therefore, it doesn't have understanding. What I want to try and do in the rest of this chapter is to look at things from the viewpoint of first deciding whether or not something has understanding. If it does, then it would follow that it would have intentionality and a knowledge of the meanings of the symbols that it is using. I believe that a basis on which to decide whether or not something has understanding, is to consider what is really involved in passing the Turing Test, and how unlikely it is that anything that didn't understand could pass that test.

As I have mentioned before in this chapter, those in favor of the claims of strong AI may grant Searle--at least for the sake of argument--

one point he is making with the Chinese Room example: that is, that although this system passes the Turing Test, it does not really have the understanding appropriate for the Turing Test. However, though those in strong AI are entitled to do this, I think that it is worthwhile considering what is involved in the Turing Test, before rejecting it on the basis of being unreliable. Before *I* accept the above point--that is, that the Chinese Room could pass the Turing Test without actually having the understanding required--without any further argument, I want to claim the following: anything that is complex enough to pass the Turing Test, would, in all probability really be a system that did understand, unless there was an overriding reason not to think it had understanding. If, in all probability, anything that did pass the Turing Test really did have understanding, then it becomes less credible that the Chinese Room could pass the Turing Test without understanding Chinese; and if it did have the required understanding, it would have intentionality. I want to claim that passing the Turing Test is a very good reason for believing that something really has the understanding required. In that case, even if you accepted it as true that Searle was instantiating a computer program that did not understand Chinese, it is very unlikely, contrary to Searle, that he would therefore really be able to pass the Turing Test.

I do not want to deny Searle's claim that the Turing Test is not a fool-proof way of determining whether or not something is actually an understanding system. The point could be granted that passing the Turing Test does not *necessarily* imply that the system that passed actually has

understanding, but the Turing Test is nonetheless very reliable considering what it can involve:

The particular value of the Turing Test is that it not only concerns simple story understanding, but anything that anyone could possibly know anything about. In Turing's original article (Turing, 1950), the questions that the computer could be asked included the writing of a sonnet, doing simple addition, and playing chess. If Searle supposes that his Chinese Room's responses are "indistinguishable from those of native Chinese speakers," then, if the questioning went on long enough, the questions and successful responses given could not only include the above, but also things like being able to:

1. Detect and respond appropriately to ambiguous words, phrases and sentences.
2. Answer truthfully about its own states.
3. Draw original conclusions about what has been given as input.
4. Give explanations of why a particular response, or responses, was given.
5. Change responses when the same questions are given at different times, and the intervening knowledge-store has changed. (See Schank, 1984, and 1986)

One could try to imagine the following counter-example to the claim that the Turing Test was a certain indicator of a system really having the understanding that has been tested (given the kinds of things that the test could include): Suppose that you have in front of you what can simply

be refered to as a black box which is about five inches square. It has holes on top from which it emits the following responses to various questions:

Questioner: And how are you today?

Black Box: Very well, thank you.

Q: What do you think about the future of Jimmy Swaggart's ministry now that his private life has been exposed?

B.B.: I guess that it is possible that he would never be able to preach on T.V. again, but it is very unlikely that the powers that be will allow this to happen, considering the money that he makes for them.

Q: You really think that he has a chance to put this whole episode behind him?

B.B.: Oh, absolutely. But it will take a great deal of effort on his part to make people believe in him once again.

We can imagine the discussion going on for some time, not only talking about the private life of Jimmy Swaggart, but also dealing with other current events, the Arts, politics, the price of gasoline, what you did at the office party last week, and so on. I believe that everyone will agree that this kind of discussion is intelligent, varied conversation, and would normally make one conclude that whatever the black box was made of, it was an intelligent and understanding system like ourselves. If we later found out that this discussion was nothing but a fluke because the black box was nothing other than a cleverly concealed tape recorder that was voice activated, would we still want to say that the black box was an

intelligent system? I think not. If we found out that it was by sheer chance that we had asked just the right questions, which the black box had been "programmed" to answer, we would not want to say that there really was any understanding or intelligence that the black box had. If we had asked the very same questions in a different sequence--even slightly differently--its responses would have been in the same order, and we would have very quickly realised that this thing had no intelligence or understanding.

The point of this rather improbable example is to show that the Turing Test, or any similar test, cannot with certainty prove that a purported understanding system really understands, and that therefore understanding could not simply be defined in terms of the things that could be included in the test. However, though this black box might pass the Turing Test on one occasion, it is very unlikely that such a box could pass the test a second time considering the kinds of things the test could include. It might, but the probability of the black box passing the test gets lower and lower each time the test is given. There is always the miniscule chance that it could pass it consistently, but such a minute possibility is really not worth talking about, since, as a matter of fact, any test *we* have for understanding is not going to prove with certainty that such and such a system really understands what has been tested. We do need some kind of criterion for deciding whether or not a system understands something, but that criterion, though it may not rule out improbable black boxes, would normally be sufficient for our purposes,

considering what the test can involve. The Turing Test may then be, under most circumstances, such a criterion.

I think that Searle's Chinese Room wouldn't be in a much better a position than the black box, if it really couldn't understand Chinese, and continually tried to pass the Turing Test. Since understanding does not only include responses to input, but also normally includes output that has not previously been asked for or suggested, it seems only fair that when doing the Turing Test, one should expect the system being tested to ask questions, for example, of its own accord. I doubt that either the black box, or Searle's Chinese Room could do this consistently well without having the understanding required. (The black box, for example, might be programmed to ask questions of its own accord at timed intervals, but its timed questions probably wouldn't work on more than one occasion.) Even if the Chinese Room could give uncalled for questions, suggestions, and so on, without understanding, it should also be able to do the other things mentioned to pass the Turing Test, and it is difficult to imagine the Chinese Room doing all of these things without understanding. If it could do things such as telling the truth about its own states, or distinguishing among, for example, the many senses of a simple word such as "set" and being able to use these different senses in the appropriate contexts, it would at least suggest that it had understanding.

Another important element in the Turing Test which could help eliminate those systems that really do not understand, would be the time factor in responding to a particular question or set of questions (see

Hoftstader, 1980). We normally think that understanding, for example, a simple request for the time of day, would not take more than a few seconds if you really did understand the request. The expected period of time allowed would get longer as the information dealt with became more complex, but we would still consider that the kind of conversation that the black box, for example, was involved with, wouldn't take much longer to complete than the time it would take a normal human being to read it. However, as Hofstader (1980) points out, there is no way we could reasonably imagine Searle in the Chinese Room, having even that snippet of conversation, without taking an incredibly long period of time to finish it:

> Imagine a human being, hand simulating a complex AI program, such as a script-based "understanding" program. To digest a full story, to go through the scripts and to produce the response, would probably take a hard eight-hour day for a human being. Actually, this hand-simulated program is supposed to be passing the Turing Test, not just answering a few stereotyped questions about restaurants. So let's jump up to a week per question, since the program would be so complex. (Hoftstader, 1980; p. 433)

Searle wants to claim that the Chinese Room could pass the Turing Test without having any understanding of Chinese, but since there is no reason to suppose that the test could not include time limitations such that if, for example, something did not answer a simple question in less than one minute, we would not consider that it had understood at least that question, and ones like it. Searle appears to assume that those giving the Turing Test to the Chinese Room will gladly wait for its responses no matter how long they take. But this, I believe, incorrectly assumes that

the Turing Test can be passed no matter how long it takes to answer even the simplest of questions. Understanding isn't just giving the right response no matter how long it takes.

The point I have been trying to make in the last few pages is that for something to pass the Turing Test it should be able to: A) pass it consistently, B) do all of the things listed 1 through 5, C) provide information of its own accord, and D) provide responses within a reasonable period of time. Perhaps there is a counter-example that shows that there is something that could do all of these things and still not understand anything, but I don't think that that counter-example is the Chinese Room. It has become more and more evident to me that it could not meet, for example, the demand of responses within a reasonable period of time. (And even if it could on one occasion, it would, I think, be purely a matter of chance, and unreasonable to suppose that it could do it repeatedly.) Perhaps Searle could describe a counter example that met these demands, but I do not believe that it would be the Chinese Room example.

So, in normal circumstances, if the Chinese Room was able to answer all conceivable questions that a normal native speaker of the language could answer (and with the same repsonse rate that a normal native speaker would exhibit), as well as asking the kinds of questions that a normal native speaker could ask, one would, more likely than not, be correct in ascribing understanding to it, unless there was some overriding reason not to. Searle would claim that the very fact that the Chinese Room is only passing the test in virtue of a formal manipulation of symbols is

itself an overriding reason not to think that such a system really had the required understanding, regardless of how well it did on the test.

So far in this chapter I have been claiming that *if* the Chinese Room as a whole could in fact pass the Turing Test, that can be a good reason for supposing that it really does have the required understanding to pass that test. However, in agreement with Searle, I think that it is unreasonable to suppose that *Searle* in the Chinese Room does understand any Chinese. (One might want to claim that he in fact does understand at least some Chinese though he is not aware of it. However, I think that this claim would be hard to defend.) But even granting that Searle is correct in saying that he does not understand any Chinese, that may not be enough to show that this Chinese Room, of which, it can be claimed, he is only a part, does not understand any Chinese either. This point is one that I will be dealing with in the next chapter. On the other hand, there may be something essentially dissimilar between the Chinese Room and a real computer and so it is unlikely that that Chinese Room really is analogous to a computer running a program capable of passing the Turing Test.

Searle might have a counter-example to the claim that a man hand manipulating Chinese symbols could not pass the Turing Test without understanding any Chinese. Searle may have shown that claim to be false, but that doesn't mean that he has also shown that a computer manipulating Chinese symbols and passing the Turing Test, could not understand Chinese. There may be several important dissimilarities between the Chinese Room

and what goes on inside of it, and what goes on inside a computer running a Chinese language program. These dissimilarities may be important enough so that although the man in the Chinese Room doesn't understand Chinese, the computer does understand Chinese. Searle does of course believe that there is nothing in the Chinese Room which is not in a computer attempting to do the same thing, that is, understand Chinese. However, as I will try to make clear in the next chapter, one can view the Chinese Room in at least two different ways, both of which try to show that just because *Searle* does not understand any Chinese, that doesn't mean that a computer running the appropriate program wouldn't or couldn't understand Chinese. So even if the Chinese Room is a counter-example to the adequacy of the Turing Test (and I believe it is a dubious counter-example to that claim), that doesn't necessarily show that it is a counter-example to the claims of strong AI.

CHAPTER III

THE SYSTEMS REPLY

As I mentioned at the end of the last chapter, part of the persuasive force of Searle's Chinese Room example is that it is easy enough to imagine with Searle that *he* does not understand any Chinese in the example. As I also said at the end of the last chapter, one might try to deny this by saying that, even though Searle doesn't believe that he understands any Chinese, he really understands at least some Chinese. He is simply not aware that he understands any Chinese. That is always a possible way in which to argue against Searle. However, I do not think that it would be an easy view to defend. It is possible Searle could understand something that he is not aware of understanding (just as someone taking a test, for example, might believe that he does not know the answer to a question that he actually does know), but it is generally agreed that we are usually aware, or can possibly be aware of our own states of mind, and what we know, understand, etc. A much easier way in which to deal with the Chinese Room is to accept the claim that Searle doesn't understand anything about the Chinese symbols he is shuffling about. It is on that basis that Searle concludes that he has shown the Turing Test to be inadequate--that is, that it can let something pass which does not have the requisite understanding. Searle believes that he has also shown that since the Chinese Room is an instantiation of a computer program, and does not understand Chinese, no computer running the same program could understand Chinese either. However, as I pointed out in

the last chapter, even if he has shown the Turing Test to be inadequate, he may not have shown the claims of strong AI to be false: That Searle doesn't understand any Chinese does not necessarily imply that nothing "in" the Chinese Room example understands the Chinese symbols.

Searle himself may understand no Chinese, but it can be claimed that he is simply begging the question as to whether or not the *entire system* he is describing understands anything about the Chinese symbols that are being manipulated. It is all right for Searle to say things like: "I produce the answers by manipulating uninterpreted formal symbols." But he is not justified on that basis in concluding that the entire system, of which he is only a part, cannot understand any Chinese. The *Gedankenexperiment* starts off as merely a description of what he is supposed to be doing, but since the whole point of the example is to show that such a system could not possibly have any kind understanding, it cannot be initially assumed that *it* doesn't have any kind of understanding; this has to be proved. So the point could be made that Searle has not found an overriding reason to think that even if the Chinese Room could pass the Turing Test it did not in fact understand any Chinese. Searle's only overriding reason for saying that is because he does not understand any Chinese, therefore the Turing Test has been shown to be fallible (and at least the first claim of strong AI has been shown to be false). If this is intended as an honest attempt to provide strong evidence for why such a system could not possibly understand, then the question as to whether or not the symbols are understood by the system--

of which Searle is only a part--has to be left neutral during the initial description.

Searle is of course depending on a strong set of intuitions against the idea that anything in the Chinese Room could understand anything when he first begins to describe the situation. Dennett, however, in his reply to the article makes the point that one can have equally strong counter-intuitions (see Dennett, 1980). In that case, you would not describe the symbols as uninterpreted. So these intuitions may show nothing either way except that they can often be used as a very powerful "pedagogical tool" (Dennett, 1980). So it is questionable whether Searle is justified on the basis of the Chinese Room in coming to the conclusion that no intentionality, for example, is exhibited by this system. Had he left it a neutral question as to whether or not the symbols were uninterpreted by the system, then the question of the existence or nonexistence of intentionality in the system would have still been clearly debateable.

One of the strongest criticisms of Searle's conclusions from the Chinese Room example does essentially depend on this kind of reasoning: that is, that although Searle doesn't understand anything, the entire system, of which he is only a part, does. This is the Systems Reply and in the article Searle states it in the following way:

> While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system and the system does understand the story. The person has a large ledger in front of him in which are written the rules, he has a lot of scratch paper and pencils for doing calculations, he has "data banks" of sets of Chinese symbols. Now,

> understanding is not being ascribed to the mere individual, rather it is being ascribed to this whole system of which he is a part. (Searle, 1980; p. 419)

Several commentators have considered this the strongest of objections to the Chinese Room, though Searle himself considered it the least plausible. Georges Rey, for example, in his article "What's really going on in Searle's "Chinese Room"," describes Searle's activity in the Chinese Room in the following way:

> *That* person (the person doing what Searle does in the Chinese Room) corresponds not to the *whole* of the emulated Chinese speaker, but merely to that person's Central Processing Unit (her "CPU"). For our purposes, this can be abstractly regarded as a species of *Universal Turing Machine*, designed to read and execute the coded descriptions of *other* Turing Machines that are themselves designed to compute specific functions... (Rey, 1986; p. 170)

Rey goes on to say about this CPU that the "AI-Functionalist no more needs or ought to ascribe any understanding of Chinese to (it)...than we need or ought to ascribe the properties of the entire British Empire to Queen Victoria." (Rey, 1986; p. 170) So the effect of this claim is that Searle in the Chinese Room is just a relatively complex but dumb distributer and collator of tasks in a system in which he is only a part (see Dennett, 1978).

John Haugeland gives a detailed description of a Universal Turing Machine which can be used to show why Searle can be considered as acting

as the CPU in the Chinese Room example, and why we need not ascribe any understanding to it (see Haugeland, 1985)

> A Turing machine consists of two parts: a head and a tape.... The tape is just a passive storage medium: it is divided along its length into squares, each of which can hold one token (from some prespecified finite alphabet). We assume that the number of tape-squares available is unlimited but that only a finite segment of them is occupied at any one time; that is, all the rest are empty or contain only a special blank token. The tape is also used for input and output, by writing tokens onto it before the machine is started and then reading what remains after it halts.
>
> The head is the active part of the machine, hopping back and forth along the tape (or pulling the tape back and forth through it) one square at a time, reading and writing tokens as it goes. At any given step, the head is "scanning" some particular square on the tape, and that's the only square it can read from or write to until it steps to another one. Also, at each step, the head itself is in some particular internal state (from a prespecified finite repertoire of states). This state typically changes, from one state to the next; but under certain conditions, the head will enter a special halt state, in which case the machine stops-leaving its output on the tape.
>
> What the head does at any step is fully determined by two factors:
>
> 1. the tokens it finds in the square it is scanning; and
> 2. the internal state it is currently in.
>
> And these two factors determine each of three consequences:
>
> 1. what token to write on the present square (replacing whatever was previously there);
> 2. what square to scan next (the same one, or the one immediately to the right or left); and
> 3. what internal state to be in for the next step (or whether to halt). (Haugeland, 1985; pp. 133-135)

The most important point that can be made about Turing machines is that for any particular program that could be run on a computer, there is a formally equivalent Turing Machine that could "run" a functionally equivalent program. However, there is normally no need to build such

Turing Machines--especially because they are so incredibly slow. We might build a particular Turing machine for a particular program, but there wouldn't be much point. However, within the present context, Turing machines are important because it appears to be clear that what is going on in the Chinese Room example is supposed to be formally equivalent to some kind of Turing machine. Clearly the tasks that Searle peforms in the Chinese Room--shuffling around bits of paper, acting upon rules given to him from outside, and so on--are equivalent in some way to what the head of a Turing machine does. The bits of paper and the symbols that are written on them are equivalent to the Turing machine's tape and the tokens written on them.

It is therefore not surprising that Searle does not understand any Chinese in the Chinese Room example. No one in strong AI expects the CPU to understand anything, and since that is what Searle, that is, Searle as the head of the Turing machine, is in the Chinese Room, no one expects him to understand. What the systems reply is claiming is that it is Searle *and* the other parts of the system with which Searle interacts that together understand Chinese. Searle's own reply to the systems reply is to make him into the whole system and not just the CPU:

> My response to the systems theory is simple: Let 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 which 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 which isn't in

him. If he doesn't understand, then there is no way the system could understand because the system is just a part of him. (Searle, 1980; p. 419)

The first thing to point out about this objection to the systems reply is that it is questionable whether Searle could in fact internalize all that is in the Chinese Room. The systems reply, if it wants to maintain that it is the whole system, and not Searle, which understands, has to maintain that things like, pens, pencils, paper, the symbols themselves, and so on, all have a relevant part to play in understanding Chinese. So just having Searle memorize the rules, etc. may not be the functional equivalent of the orginal Chinese Room. Even if this internalization is the functional equivalent of the entire system of the Chinese Room, Searle as complete system is not the same as Searle as CPU in the original Chinese Room example (see Carleton 1984). In the original Chinese Room example, Searle is only the CPU, and the contention of the systems reply is that as the CPU he doesn't understand Chinese, though the system as a whole understands Chinese, and he is not in a position to know anything about the system's understanding or not understanding. So when Searle internalizes the Chinese Room, Searle is no longer just the CPU: he is now the entire system which the systems reply claims does understand though the CPU by itself does not.

So my point is that if the systems reply is correct, when Searle is in the Chinese Room he is merely Searle-as-CPU-that-does-not-understand, but in its revised form he is Searle-as-complete-system-that-does-understand. Since Searle's identity is not the same in both cases, the

conclusion he comes to in the first, that is, that he as CPU does not understand Chinese, is not the same conclusion he can come to in the second case since he is more than the CPU. I believe that Searle imagines that when he internalizes the Chinese Room, and then says that he--as complete system--does not understand, he is assuming that there is nothing in the Chinese Room which could understand if he does not understand. But as I have already pointed out, one can have counter intuitions to those of Searle which make one believe that it is possible that whatever else is in the Chinese Room could understand Chinese, while Searle does not. I also believe that Searle does not fully realise that in the first case he was only the CPU, and that therefore he would not understand any Chinese. If he had realised that he probably would not have even thought that his not understanding was sufficient to show that the Turing Test had been shown to be fallible, and that at least the first claim of strong AI had been shown to be false.

Although the systems reply is helpful in making clear what exactly Searle's position is in the Chinese Room, its claim that the entire system of pens, papers, ledgers, and Searle together understand Chinese, is rather questionable. Even though the Chinese Room may be abstractly equivalent to a Turing Machine, and that in turn may be somehow equivalent to a real computer running a Chinese language program, there may be something important which is not there in either the Chinese Turing machine, or in the Chinese Room which means that they cannot understand Chinese even though it might be that a real computer could understand Chinese. I have already mentioned that the slowness of this Chinese Room system, makes it

questionable as to whether or not this system really could pass the Turing Test. (This slowness would also be apparent in the Chinese Room's formally equivalent Turing machine.) But a real computer, although it might not be as quick as a human being in all respects, would be able to respond in a more reasonable period of time. So my point is that although the systems reply is correct in saying that Searle cannot understand Chinese since he is not the entire system, it may be wrong in saying that this entire system does because, for example, it is so incredibly slow. And although that system may be somehow formally equivalent to a real computer running a Chinese language program, they are not equivalent in terms of time taken to respond to input and give output and that may be essentially relevant to understanding. So my point is that even if the Chinese Room does not understand Chinese that may not imply that a computer running a Chinese language program couldn't since they are not equivalent in what may be essential to understanding. And so just because the Chinese Room system couldn't understand Chinese, that wouldn't necessarily mean that at least the first claim of strong AI was false.

After Searle imagines trying to internalize all the elements of the Chinese Room example, he objects to the claim that things like pens, papers, and Searle working together as they did in the Chinese Room example, could have an understanding of Chinese. According to Searle, if the systems reply was accepted as sound--that is, that a system of rules, symbols, books, pens, paper, room and Searle can understand Chinese--then we would also have to accept other kinds of bizarre systems as having understanding:

> ...my stomach has a level of description where it does information processing, and it instantiates any number of computer programs, but I take it we do not want to say that it has any understanding. Yet if we accept the systems reply it is hard to see how we avoid saying that stomach, heart, liver, etc. are all understanding subsystems, since there is no principled way to distinguish the motivation for saying the Chinese subsystem understands from saying that the stomach understands. (Searle, 1980; p. 420)

Even if the systems reply is right in saying that pens, paper, and so on, are functionally equivalent to what a computer does and the time taken was of no account, and that both understand, that would not necessarily have to mean that they would end up taking a slippery slope down to the point where they have to say that your stomach or your heart also understand. Those in favor of the systems reply do at least partly base their conclusion that the Chinese Room system does understand on the claim that Searle makes that it passes the Turing Test (and that it is somehow equivalent to a Turing machine). They have already accepted that if it passes it, it doesn't do that in virtue of only having Searle as CPU. The other parts of the system are relevant, and if they are relevant enough so as to enable them and Searle together to pass the Turing Test, their doing so, as I have argued, is good reason to think that they have what it takes for understanding. So claiming that what is in the Chinese Room system has what it takes to understand and pass the Turing Test, doesn't mean that you have to claim the same for your stomach, liver, or other similar organs of the body.

No stomach, no liver, no heart, etc. is going to pass the Turing Test or any similar test for understanding or any other cognitive state, since

they couldn't take the test in the first place. (You could always try to have a stomach, for example, take the Turing Test or any similar test, but you wouldn't get very far before you realised that it would fail.) You could also distinguish between systems that do understand, and those that don't, by considering the relevant differences in their *formal organization*, and that is precisely what functionalism and strong AI do. There is surely a definitely distinguishable and relevant difference between the formal organization of the human brain, or the computer, with the appropriate program, that purportedly understands, and the stomach, for example, that digests food and not much else. Not all formal organizations are the same, and not all of them have the relevant kind of complexity and output in order to have understanding. If you were to question this claim, then there seems to no principled way in which we can distinguish between, for example, the biochemistry of the human brain, and the biochemistry of the stomach, and why, according to Searle, the former is the only one of the two that "produces" intentionality etc..

No one in AI claims that all computers running their programs have to exhibit understanding, and so even if there is a program that your stomach instantiates, that is not sufficient to show that it has understanding, since that program would have nothing to do with understanding anything. It can also be pointed out that what your stomach or liver does, and what you do when you speak Chinese, though they may all be instantiations of computer programs, are not the same kinds of things: Even if all of these things can be described in some way that involves them all in doing information processing, that does not imply that all

information processing necessarily involves understanding (see Carleton, 1984).

If the systems reply were wrong in claiming that the Chinese Room system has what it takes for understanding, that would still not mean that strong AI is wrong (or that it couldn't reply to Searle's slippery slope objection).

There are two claims made by the systems reply. The first is that Searle acting as the CPU does not understand Chinese, and is not expected to. The second claim is that although Searle--acting as the CPU--does not understand Chinese, the entire system does understand Chinese. Those in favor of strong AI may easily enough accept the first claim. However, they may not wish to accept the second claim. Their reason for doing that may be because although they believe the appropriately programmed computer would literally have understanding, they do not believe that the Chinese Room system is relevantly equivalent to a computer running the appropriate program. In that case, the Chinese Room system should not be expected to have an understanding of Chinese or anything else, since it is not relevantly like a computer running the appropriate program. If those in favor of strong AI believe this about the second claim made by the systems reply, then they should also believe that Searle has still to give a counter-example in which a system relevantly like a computer running the appropriate program, does not understand Chinese, or anything else.

CHAPTER IV

FURTHER REPLIES

In the last chapter I mentioned that Searle does not believe that the systems reply is a good reply to his claim that neither he nor the entire system in the Chinese Room understands any Chinese. I believe, on the other hand, that it is, at least in one way, one of the strongest of all the replies that Searle considers in his article. As I pointed out in the last chapter, there are two claims made by the systems reply. The first is that Searle--as the CPU--is not expected to understand anything. The second claim is that the entire system--of which Searle is only a part--does understand Chinese. I believe that it is the first of these two claims which makes the systems reply one of the strongest replies to Searle's Chinese Room example. Being able to show that Searle is in no position to understand Chinese when he is merely acting as the CPU, means that Searle cannot, as he tries to do, conclude from the fact that he does not understand any Chinese, that nothing in the Chinese Room understands Chinese. From Searle's point of view in the Chinese Room, he--as the CPU--is merely formally manipulating symbols, but that doesn't imply anything about the entire system as Searle mistakenly thinks it does.

It is the second claim--that is, that the enitre system understands Chinese--that I find questionable. In admitting that it is questionable, or further, denying that the entire system understands Chinese, does not necessarily grant Searle his conclusion anyway that the first claim of

strong AI is false. I find it questionable, and probably false, that the entire Chinese Room system understands Chinese because I do not believe that pens, paper, symbols and Searle are relevantly like a computer running a program that is able to pass the Turing Test. Part of my reason for saying that is because of the incredible slowness of this system, and slowness is not very often helpful when it comes to understanding anything. (Another possible reason is that Searle actually states that the Chinese Room is just him (see Searle, 1980; p. 418). If Searle really believes this, then the Chinese Room isn't even formally equivalent to a Turing machine, since he is just the formal equivalent of such a machine's head, and it has already been made clear that that by itself cannot understand. It may be that Searle just made a slip when he states that the Chinese Room is nothing other than him, for it appears to be clear that he is trying to describe something formally equivalent to a Turing machine and a Turing machine is not just its head.)

So even if the Chinese Room system does not understand any Chinese, that does not show that the first claim of strong AI is false. If the Chinese Room system isn't relevantly equivalent to a computer running a program, Searle needs another example to show that a system relevantly equivalent to a computer doesn't and couldn't understand Chinese or anything else.

However, the systems reply, unlike the other replies that do not concede that Searle's arguments against strong AI may be correct, tries to deal *specifically* with the original Chinese Room itself in a way that

should be acceptable to those working in the strong AI field. (The other minds reply (see p. 70), for example, although it challenges Searle's conclusions based on the Chinese Room, is not a very strong argument, or one that proponents of strong AI would find very appealing.) All the other replies that do not concede to Searle his conclusions altogether, at least appear to simply accept that Searle is right when he says that nothing "in" the Chinese Room understands Chinese. After they apparently grant that, they try to think of other kinds of systems which they claim would understand, for example, Chinese. Although, as I claimed above, it is probably false that the Chinese Room system does understand any Chinese, it is not clear from the way the other replies are stated, what reasons the other replies have for apparently accepting that same conclusion. If the other replies are trying to defend the claims of strong AI against Searle's views, they cannot simply claim that the Chinese Room system does not understand, and that Searle does not understand, because they are relevantly equivalent to a computer running a strong AI program. If that were the reason, then there would be no way that they could defend the claims of strong AI since they would have already conceded to Searle his argument against them.

For example, suppose that you claim that the Chinese Room does not understand any Chinese *and* that it does instantiate a program similar to one a computer would to pass the Turing Test. It would seem that if you supposed this you are implying that it is not merely because of the simplicity of the system that it couldn't understand, but instead, also because it *only* instantiated a program sufficient to let it pass the

Turing Test which is what Searle is maintaining. If these other replies do not simply claim that the Chinese Room is too unsophisticated to follow any strong AI program (let alone one that would pass the Turing Test which would be an incredible achievement for any strong AI program (see Schank, 1980; p. 446-447), they are probably going to end up having to concede everything to Searle. I believe, on the other hand, that the other replies are simply saying that the example should be changed to a system that has a more reasonable chance of actually instantiating a program that could possibly pass the Turing Test, or pass a test that would be similar but simpler than the Turing Test.

The robot reply is stated by Searle in the following way:

> Suppose we wrote a different kind of program from Schank's program. Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather it would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about....anything you like. The robot would, for example, have a television camera attached to it that enabled it to see, it would have arms and legs that enabled it to act, and all of this would be controlled by its computer brain. Such a robot would, unlike Schank's computer, have genuine understanding and other mental states. (Searle, 1980; p. 420)

I see no reason for the robot reply to make any mention of Schank's programs, and whether or not they understand. As I pointed out in Chapter II, Searle's Chinese Room example goes far beyond attempting to instantiate a simple story understander program. If Searle's Chinese Room instantiates anything, it instantiates a program capable of passing the Turing Test. It would of course be true that if the Chinese Room did instantiate a

computer program capable of passing the Turing Test, and was not able in virtue of it to understand any Chinese, that a Schankian style program couldn't understand either. (Since the Turing Test program could have incorporated into it this kind of program.) However, since I have tried to deny the claim that the Chinese Room does in fact instantiate a Turing Test program, it is still conceivable that a Schankian style program could understand at least at the making sense level.

So my point is that the robot reply really only needs to admit that Searle and the Chinese Room as a whole doesn't understand any Chinese for the kinds of reasons I have already given. It need only admit that a different kind of system from that described in the Chinese Room example, and one that does instantiate a computer program, is needed. All of that would not imply anything about Schankian style programs, and so the robot reply need only concern itself with the inadequacy of the Chinese Room example, and not the purported inadequacy of Schankian style programs.

In his initial response to the robot reply, Searle states that:

> The first thing to notice about the robot reply is that it tacitly concedes that cognition is not solely a matter of formal symbol manipulation, since this reply adds a set of causal relations with the outside world. (Searle, 1980; p. 420)

I do not agree with Searle that the mention of a set of causal relations with the outside world by the robot reply in some way distinguishes it from the claims of strong AI. (As the above response to

the robot reply appears to imply.) The first claim of strong AI was originally stated by Searle as the belief that the "appropriately programmed computer...can be literally said to *understand* and have other cognitive states." (Searle, 1980; p. 418) This way of stating the first claim of strong AI does not have to mean anything more than the belief that only a certain kind of formal symbol manipulation, in a certain kind of machine, is sufficient for understanding. (I thank Prof. John E. Nolt for making this point clear to me.) All that the first claim is saying is that the appropriately programmed computer can be said to understand. It is not saying that just any kind of machine, system, etc., with just any kind of formal symbol manipulation will have understanding. As the first claim is stated, it can be taken as only allowing certain sorts of appropriately programmed formal symbol manipulators, with some relevant causal interactions with the world, to have understanding. In that case, the robot reply would be in keeping with the claims of strong AI. The robot reply could be taken as merely making clear what it believes to be one of the important aspects of the kind of machine required for understanding, and how this kind of formal symbol manipulator needs to interact with the world.

When Searle first described the Chinese Room example, he believed that whatever that formal symbol manipulator did was in keeping with the claims of strong AI. In the Chinese Room example, there were lots of causal interactions with the world going on. So although Searle is claiming that the Chinese Room couldn't understand, even though it apparently was instantiating a very complex program, he can't really claim

that the robot described is doing anything essentially different or that it adds anything else to the system except a greater complexity. Even if the Chinese Room was indeed instantiating a complex program, the difference between that system and the robot is not, as Searle is suggesting, one of kind, but of degree of complexity. Why should the causal interactions going on inside of a computer be essentially any different from those that it might have with the outside world? In both cases it is dealing with information that is formally specified. Searle apparently forgets that *any* computer has some kind of causal relations with the outside world be it as humble as having information printed *through* a computer keyboard. There may be certain kinds of causal interactions with the world that are more appropriate than others for certain kinds or levels of understanding, and this could be all that the robot reply is making clear. If Searle believes that "simpler" computers are nothing but formal symbol manipulators in spite of their limited causal interactions with the world, why should he think anything else is being suggested by the robot reply except a more complex sort of formal symbol manipulator with more complex causal interactions with the world?

So the point I am making about the robot reply--which Searle seems to miss--is that if you did have a robot that did things like perceiving, walking, talking, and so on, it would have many advantages over a computer that only had a keyboard as its means of interaction with the world. Some kind of interaction with the world is needed so that anyone or anything can gain information and understand. The more complex the interaction with

the world is, the more complex the information is also, and, plausibly, the understanding achieved is also more complex and of a higher level.

Searle may in fact grant the point that I have been making here and concede that these kinds of causal interactions with the world don't change in some subtle way the first claim of strong AI. Even if he grants that point, he will still maintain that no understanding is possible in such a system, because he will still maintain that something is missing, namely, intentionality, the knowledge of the meanings of the symbols used, and, of course, understanding and other cognitive states, because it is still only a certain sort of formal symbol manipulator.

In the article itself, after he claims that the robot reply tacitly concedes that there is more to understanding than formal symbol manipulation, he goes on to say that even if it concedes that, it is still not a system that could possibly understand anything. He tries to show this by the use of a revised version of the Chinese Room:

> Suppose that instead of the computer inside the robot, you put me inside the room and you give me again, as in the original Chinese case, 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 serve to make the motors inside the robot move the robot's legs or arms. It is important to remember 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 without knowing either of these facts. (Searle, 1980; p. 420)

Searle is claiming that even with the addition of sensors, such as a television camera, and body parts, such as arms and legs, *he* still does not understand any Chinese. Searle believes that even though the *robot* has these causal relations with the outside world, Searle, inside of the robot, is not aware of them. He would also claim that even if he was aware of those causal relations, it would not give him an understanding of the symbols that he is merely formally manipulating. Even though this thought experiment is somewhat different from the original Chinese Room, there is, according to Searle, no more understanding in this robot system as there was in the original Chinese Room, i.e. there is, according to Searle, no understanding or any other cognitive states since *Searle* still doesn't have a knowledge, for example, of the meanings of the symbols he is manipulating.

Although this is a revised version of the Chinese Room, I believe that it is still susceptible to the systems reply claim that Searle is merely the CPU. Searle wants to maintain that *he* still does not understand any Chinese. But as the systems reply made clear when dealing with the original Chinese Room, the fact that Searle--as the CPU--does not understand the Chinese (and an understanding of Chinese is not expected of the CPU), or what the robot is doing, or why the robot is moving its arms, legs, etc. doesn't imply that the robot as a whole doesn't understand what or why it is doing whatever it is that it does. I should think that those in favor of the systems reply, as presented in the last chapter, would want to not only point out that Searle is merely the CPU, but also claim that this entire robot system--of which Searle is a part--understood

Chinese, or whatever else it was dealing with. I say this because if anyone were prepared to claim understanding for the Chinese Room as it was first described--as the systems reply did--they would surely want to claim as much for this robot. There is perhaps more reason to think that this robot--into which Searle is incorporated--has a better chance of understanding, due to its greater complexity for example, than the orginal Chinese Room. But since Searle hasn't changed in terms of his speed, there is still the problem of the incredible slowness of *this* robot. As was pointed out in the chapter II, it would be reasonable to expect that for anything to pass the Turing Test, it should be able to respond in a reasonable period of time. This will not be possible in the robot in which Searle is a part. So both this robot, and the Chinese Room, will not be able to do what is crucial to Searle's argument against the fallibility of the Turing Test.

Of course, as I have maintained, the issue of the reliability of the Turing Test, is separate from that of the claims of strong AI. But this robot example described by Searle doesn't help his argument against those claims either. Searle as CPU is in no position to understand, and that fact doesn't imply anything about the rest of the system, and whether it can or cannot understand.

As I pointed out in Chapter III, I do not think the systems reply is correct in ascribing understanding to the Chinese Room system, although I do think it is correct in pointing out that Searle is merely the CPU. I have already stated my reasons for saying that the Chinese Room system

does not understand, and none of those reasons imply that the claims of strong AI are false. (My main reason was that that system was not functionally equivalent to a computer running a Turing Test or any other program.) I believe that the same may be true of the robot of which Searle is a part. Although it is not like the orginal Chinese Room in that it consists of more than pens, papers, pencils, and so on, it still has Searle as the CPU, and as the CPU he is still incredibly slow. It may be that although that robot doesn't understand, because of Searle as the CPU, the other robot described in the robot reply itself does understand since its CPU can work at speeds much greater than the Searle robot. That this is true does of course depend on the claims of strong AI being correct, and that this robot--the one in the robot reply--really does instantiate the appropriate program.

For the robot to understand, its understanding would not just depend on the speed of its CPU. At least part of what that robot needs to understand is to have some kind of internal representations of the things that it is dealing with (see, for example, Fodor, 1980; p. 431). Whenever I am thinking about something, or whenever anyone else does the same, there is some kind of mental image, picture, idea, association of ideas, what you will, that allows us to do the thinking necessary for understanding. All of these kinds of activities could be described as internal representations of some sort that are necessary for anyone to understand anything. It should be made clear that these internal representations do not necessarily have to be literally pictures of some sort that are stored in the brain. One form of internal representation that we probably have are representative

structures similar to those exhibited by Schankian scripts. Such representations would be constituted by a set of interrelated expectations, goals, inferences, and so on. They consist of "blueprint(s) for action without the details filled in" (Schank, 1984; p. 114) so that when we get new information about international politics, for example, we can deal with it, and understand it, at least to some extent, based on our ability to "fit" this new information in the framework(s) of these structures which we already have stored as some kind of internal representation. The question then is: Can we get a computer/robot to have similar types of representations without being anything more than a formal symbol manipulator? It would seem that some computers with the appropriate programs do already have these kinds of internal representations. (And from what I have already said in this paragraph, Schank's scripts may be considered as a species of internal representation suitable for computer run programs.)

An example is Terry Winograd's simulated robot SHRDLU (Haugeland, 1985; p. 185 ff). SHRDLU "lives" in an imaginary world of cubes, pyramids, cones and other similarly shaped objects. Without getting into too many of the details, SHRDLU moves the objects around as it is requested. New objects can be created out of these objects which, at first, it is not able to create until the new objects are described as being built out of the old ones. For example, a church with a steeple may be described as something made out of a cube and a pyramid on top of it. Once it is told in this way what a church with a steeple is, it will be able to build one. There isn't actually anything inside of SHRDLU in the shape of a cube, a cone, or

whatever, but that should not be considered a problem since we don't normally consider our own internal representations of things or people as, round, six feet tall, or built like a steeple. The point of these internal representations is that they enable us at least to change, for example, from one way of appreciating a situation to a new way of appreciating it. In the case of SHRDLU we can at least say that its imaginary blocks world enables it to keep track of what is the case with its blocks world at any given time, and how it can bring about changes in it.

The SHRDLU example may not show that anything is understood, and it may not show that its internal representations are relevantly similar to our own. However, it does show that some kind of internal representation is possible in a computer or robot, and it does not rule out the possibility that those representations are relevantly similar to the kinds that we need in order to understand. It also makes clear that a formal symbol manipulator can do more than simply manipulate symbols in a formal way. Such a system can at least have internal representations of some kind. I should also like to point out that although the robot reply does not make any mention of internal representations, I do not think that that reply rules this out as at least part of what a formal symbol manipulator needs to do in order to have an understanding about anything. This may also be another reason why Searle in the Chinese Room, functioning as its CPU, could not understand Chinese: As the CPU, *Searle* was not sophisticated enough to have internal representations. (Searle's activities as a normal human being, and not as a CPU, would of course enable him to have internal representations.)

Searle considers four more replies to his Chinese Room: The brain simulator reply, the combination reply, the other minds reply, and the many mansions reply. As part of the work of the rest of this chapter, I will only give serious consideration to the first two of these. In my opinion, the last two are really only worth very brief consideration.

The brain simulator reply is stated by Searle in the following way:

> Suppose we design a program that doesn't represent information that we have about the world, such as the information in Schank's scripts, but simulates the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories in Chinese and gives answers to them. The machine takes in Chinese stories and questions about them as input, it simulates the formal structure of actual Chinese brains in processing these stories, and it gives out Chinese answers as outputs. We can even imagine that the machine operates not with a serial program but with a whole set of programs operating in parallel, in the manner that actual human brains presumably operate when they process natural language. Now surely in such a case we would have to say that 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? At the level of the synapses what would or could be different about the program of the computer and the program of the Chinese brain? (Searle, 1980; p. 420)

In Chapter I I quoted Searle as saying that if you exactly duplicated all the causes of understanding, intentionality, etc., you could exactly duplicate the effects (that is, actually have understanding, intentionality, etc.). According to Searle, our ability to understand, have intentionality, and so on, was dependent on the causal powers of the human brain with its specific biochemistry. So, according to Searle, if you could exactly duplicate in a computer run program not only the formal structure of the

human brain, but its biochemistry as well, you would be able to have understanding, intentionality, etc. The brain simulator reply at least suggests that you only need to have the formal structure of the human brain to do this (and that may not include the formal structure of its biochemistry), that is, if you exactly duplicate in a computer run program the formal structure of the human brain, you will have exactly duplicated the effects of that structure such as understanding, intentionality, and so on.

I am not sure that anyone in strong AI would have to go as far as the brain simulator reply in order to respond to Searle's arguments against strong AI. It would seem that those in strong AI may only need to go this far if, for example, the systems reply had failed to show that Searle was merely the CPU, or that, for example, the robot reply had failed to show that a much more complex system was needed than the Chinese Room. (In fact, the replies are so ordered in Searle's article that this sequence of failures is implied.) But the purported failure of both the systems reply and the robot reply as responses to Searle's Chinese Room, and his claims concerning its lack of understanding, etc., are at least partly what is in dispute. Searle believes that those in strong AI do not have those two options any longer, since he believes that he has refuted these replies. (He may be correct in saying, for example, that the Chinese Room does not understand Chinese, but not because of the reasons that he gives.)

In his response to the brain simulator reply, Searle states that:

> ...it is an odd reply for any partisan of artificial intelligence (functionalism, etc.) to make. I thought the whole idea of strong artificial intelligence is that we don't need to know how the brain works to know how the mind works. (Searle, 1980; p. 421)

As I have mentioned already, those in strong AI need not have to go as far as the brain simulator reply if either the systems reply or the robot reply is sufficient to respond to Searle's claims exemplified in the Chinese Room. But even if they did, it is clear that in the brain simulator reply what is important is the *formal structure* of the brain, not what it is made of. Certainly the functionalist, or the proponent of strong AI, can claim that we have to know something about the way the brain works in terms of its formal organization. For example, the fact that the human brain deals with various kinds of information all at the same time may be a good reason for using parallel processors rather than serial processors. However, they may not believe that we actually have to go as far as duplicating exactly all of the formal structure and organization of the human brain, for example, its biochemistry, if there are only certain aspects of it that deal with higher level capabilities such as understanding. But in doing that, they need not be claiming that we have to replicate the material in which that formal organization occurs in us. That sort of claim is what Searle is maintaining. He believes that in us it is the specific biochemistry of our brains with its causal powers that allows us to have cognitive states such as understanding. At the same time, Searle wants to allow for the possibility of other beings with other kinds of chemical make up to also have these causal powers and cognitive states. However, on Searle's view of the matter, if an alien merely

replicated the formal structure of our brains without a chemistry that was exactly the same, or a different chemistry with exactly the same causal powers, it couldn't have things like cognitive states.

At least one of the problems with this kind of criterion for understanding is the following: If there are different brain chemistries with the same causal powers, then there is a difficulty in distinguishing those chemistries with the right causal powers from things that don't have them, unless those that have the right kind of chemistry with the right causal powers all have the same formal structure, while the things that do not have the right chemistry and causal powers have a different formal structure (and perhaps they also need to have a different external behavior from these special cognitive materials). But if they did all have the same formal structure, there would be no principled way of distinguishing them from the brain simulator if it had, for example, the same formal structure as our own brain's biochemistry (as well as the same external behaviors). On the other hand, if all of these brains do not have the same formal structure, there does not seem to be a way of distinguishing those kinds of things from other kinds of things that don't think, understand, etc. since there could be something that was apparently cognitive but it was not made of the right stuff to actually be cognitive, but there would be no way knowing this. Searle could simply claim that such and such a thing, although it had the same formal structure as our own brains simply couldn't do what we do because it didn't have the right kind of chemistry, but that just sounds like "neural chauvinism" (Cuda, 1985). There seems to be no principled way of saying what is and what is

not the "right kind" of chemistry if you allow for more than one kind (which Searle does).

After Searle questions the use of the brain simulator reply by those involved in strong AI, and in functionalism, he goes on to maintain that even if this reply does not abide by strong AI principles, the "brain" described in the brain simulator reply will not have any understanding:

> To see that this is so, imagine that instead of a monolingual man in a room shuffling symbols we have the man operate an elaborate set of water pipes with valves connecting them. When the man receives the Chinese symbols he looks up in the program, written in English, which valves he has to turn on and off. Each water connection corresponds to a synapse in the Chinese brain, and the whole system is rigged up so that after doing all the right firings -- that is, after turning on all the right faucets -- the Chinese answers pop out at the output end of the series of pipes. (Searle, 1980; p. 421)

Searle, of course, does not believe that such an "elaborate" system of water pipes could possibly have an understanding of Chinese even though it apparently exactly duplicates the formal structure of the brain. His reason for saying this, as I have indicated above, must be that it isn't made of the right kind of stuff. But, as I have also indicated above, although that kind of reason might appear obvious and right, it doesn't help his claim that other beings with other chemistries could have understanding. Although it is obvious that a statue of a person, for example, does not have understanding or any other cognitive states, even if it exactly duplicates the person's weight, height, color, and so on, because it does not have, for example, the ability to pass the Turing Test, it really is

not that obvious what we would think and should say if there really was a system of water pipes that could pass the Turing Test, which is exactly what this system is supposed to be able to do (Dennett, 1980; p. 434). (If Searle claims that the Chinese Room can pass it, there is no reason why a system of water pipes couldn't pass it.) As I have already said before at different points in this thesis, if something can consistently pass the Turing Test, it should at least give us pause for thought before we "obviously" conclude that such a system really doesn't nor couldn't have any cognitive states.

For Searle to maintain his position against strong AI, and functionalism, it would be much more direct for him to claim (but more difficult to defend) that the only material that could possibly have cognitive states, is what we know we are made of, and some animals are made of. But that would rule out any beings, regardless of their apparent abilities to understand us, that were not made of the same stuff. At least in the case of functionalism, and strong AI, you are not left with the difficult task of trying to find out which of many possible materials could really have cognitive states: If they are all functionally the same, they are all cognitive beings.

Another problem with this alleged counter-example to the brain simulator reply is that there is no obvious need for the man to have anything to do with this system. Perhaps he would be needed to start the whole thing going, but after that, if this really is a system that duplicates the formal structure of the brain, it would seem that it could take care of

any further job the man was supposed to do. For example, it could turn its own valves on and off after the initial starting up. So unlike Searle's earlier counter-examples, this one may not even need to have him or anyone else in it. Even if it did need him in some integral way, there is no reason to suppose that he would be in this purported counter-example anything other than the CPU. (And its job does not involve understanding anything.)

Unlike the other replies, the next reply, the combination reply, does not try to present a counter-example that is self-sufficient. Rather, the combination reply depends on the previous three replies being individually at least partially correct:

> Imagine a robot with a brain-shaped computer lodged in its cranial cavity; imagine the computer programmed with all the synapses of a human brain; imagine that the whole behavior of the robot is indistinguishable from human behavior; and now think of the whole thing as a unified system and not just as a computer with inputs and outputs. Surely in such a case we would have to ascribe intentionality to the system. (Searle, 1980; p. 421)

In this case, all internal and external behavior is formally/functionally equivalent to that going on in a human being, but even in this case, Searle wants to say that it is not enough for it to have cognitive states such as understanding:

> ...we would find it rational and indeed irresistible to accept the hypothesis that the robot has intentionality, as long as we knew nothing more about it....but as soon as we knew that the behavior was the result of a formal program, and that the actual causal

> properties of the physical substance were irrelevant, we would abandon the assumption of intentionality. (Searle, 1980; p. 421)

Clearly this response to the combination reply rests on the already discussed claim that it takes a certain sort of material to have things like understanding. But as I have already pointed out in this chapter, there seems to be no principled way in which you could distinguish arrangements of materials that really are cognitive from those that are not (especially because you could have a robot such as the above which to all intents and purposes does exaclty as we do even though, according to Searle, it is not intentional). So, according to Searle, even if this robot could, run, walk, talk, take and pass philosophy exams, tell jokes, cry, laugh, climb mountains, describe its past experiences and evaluate them, make dinner, and so on, it could not be a cognitive being because it was only doing these things in virtue of a formal program. But it seems more reasonable to suppose that if it could do these things solely in virtue of a formal program, that it really was a cognitive being with all the respect due to one. Even if the robot was only instantiating a formal program, its behavior internal and external, should make one consider other ways of deciding whether or not it really is a cognitve being.

Part of Searle's response to the robot described in the combination reply is to formulate a purported counter-example to the claim that the robot really does have intentionality, understanding, etc.:

> Suppose we knew that the robot's behavior was entirely accounted for by the fact that a man inside it was receiving uninterpreted formal symbols from the robot's sensory receptors and sending out

> uninterpreted formal symbols to its motor mechanisms, and the man was doing this symbol manipulation in accordance with a bunch of rules. Furthermore, suppose the man knows none of these facts about the robot; all he knows is which operations to perform on which meaningless symbols. (Searle, 1980; p. 421)

This purported counter-example is very similar to Searle's response to the robot reply which in turn was similar to the original Chinese Room example. I believe that this response to the combination reply is susceptible to many of the objections already raised against both the orginal Chinese Room, and the robot reply. In this case, as in the other two purported counter-examples, Searle does not have to be considered as anything other than the CPU. And as has already been pointed out, no one expects that the CPU understand anything, and in that case, its not understanding doesn't imply that the whole system does not as well. Another point that was made when discussing the robot reply was that the robot in which Searle plays a part may not understand, while the robot in the robot reply itself, may really understand. The reason for saying that was because of the problem that Searle, as the CPU, had in all of these systems: being so incredibly slow. The robot in Searle's purported counter-example to the combination reply may not, as a whole, understand for the same reason, while the robot of the combination reply itself, may understand.

From the point of view of those defending the combination reply, the only thing that their robot does not have which all human beings have, is the specfic material out of which human beings are made. According to them, in all other respects this robot is functionally/formally the same as

any other cognitive being such as ourselves. From Searle's point of view, it is because it is not made out of the right stuff, or more specifically because it doesn't have the right biochemical structure, that it cannot understand or have any other cognitive states. It might be the case that whatever stuff this robot is made out of actually does have the right causal properties for intentionality, but as Searle points out, that is not what those using the combination reply are claiming (Searle, 1980; p. 421). They are claiming that it has intentionality and understands in virtue of having the appropriate program(s) with the appropriate formal/functional organization, and not because it is made out of a particular stuff.

I have already discussed in this chapter some of the problems involved with Searle's criterion for having intentionality, understanding, and so on, and I do not want to repeat them here. However, I believe that they are relevant criticisms of Searle's response to the combination reply, because although he cannot infer that the robot doesn't understand because he doesn't, he tries to show that since the stuff it is made out of is considered irrelevant, it cannot have intentionality and understanding based on its instantiating the appropriate program. But his criterion, as I have pointed out, for intentionality and understanding is itself suspect. On the other hand, I believe that there is another type of criterion for saying, for example, that something is intentional, which is in keeping with the claims of strong AI and functionalism, and is in my opinion much more acceptable than Searle's criterion. I will get to this criterion in Chapter V.

The last two replies that Searle considers are the other minds reply, and the many mansions reply. I do not believe that either of these replies can be given serious consideration by those working in the field of strong AI. The other minds reply simply avoids the question as to whether or not the internal structure of something is important in deciding whether or not it has understanding:

> How do you know that other people understand Chinese or anything else? Only by their behavior. Now the computer can pass the behavioral tests as well as they can (in principle), so if you are going to attribute cognition to other people, you must in principle also attribute it to computers. (Searle, 1980; p. 421)

The other minds reply considers only external behavior, but, as has already been granted, this is not a certain guarantee of a system really having understanding. Searle believes that it is the causal powers of the brain, and not just external behavior, which are important. Strong AI believes that it is formal structure, and not just external behavior, that is important in determining understanding. That something could pass certain behavioral tests would be most certainly a very plausible reason for saying it really had cognitive states, but as I have pointed out in Chapter II, I could not say that it *must* have them purely on the basis of its behavioral success. Unless one is a behaviorist, one is attributing something more to a cognitive system than its behavioral success, and that something more should be something that is detectable by us. If you are in favor of strong AI or functionalism, that something more should be having the appropriate structure and organization. If you are in favor of Searle's views, you will look for the appropriate material with the appropriate

causal powers. But as has already been pointed out in this chapter, Searle's criterion for intentionality, understanding and other cognitive states, is susceptible to criticisms that functionalism's criterion is not susceptible to.

In the case of the many mansions reply, Searle's belief that only certain kinds of materials with certain kinds of causal powers capable of producing intentionality, understanding, and so on, is not questioned:

> Your whole argument presupposes that AI is only about analogue and digital computers. But that just happens to be the present state of technology. Whatever these causal processes are that you say are essential for intentionality (assuming you are right), eventually we will be able to build devices that have these causal processes and that will be aritificial intelligence. So your arguments are in no way directed at the ability of artificial intelligence to produce and explain cognition. (Searle, 1980; p. 422)

I agree with Searle when he says that this reply is no help to strong AI since it is conceding that AI needs to look at means other than computer run programs in order to create artificial intelligence.

CHAPTER V

THE INTENTIONAL STANCE AND CONCLUDING REMARKS

The final part of Searle's article is spent reiterating in a summary form what he has already said regarding the claims of strong AI. I think that the most concise way in which he says this is the following:

> ...formal symbol manipulations by themselves don't have any intentionality: they are meaningless; they aren't even *symbol* manipulations, since the symbols don't symbolize anything. In the linguistic jargon they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and who interpret the output.
>
> The aim of the Chinese room example was to try to show this by showing that as soon as we put something into the system which really does have intentionality, a man, and we program the man with the formal program, you can see that the formal program carries no additional intentionality. It adds nothing, for example, to a man's ability to understand Chinese. (Searle, 1980; p. 422)

There are two claims that Searle is making in this conclusion. The first is that formal symbol manipulation by itself will not have any intentionality, and is meaningless. The second claim is that the Chinese Room example shows this to be true, since the man--Searle--in this room, although he instantiates a formal program does not understand Chinese or anything else solely in virtue of instantiating that program. It has been the second of these two claims that I have been primarily been dealing with. Considering the strength of at least some of the replies dealt with in the previous three chapters--especially the the point made by the

systems reply that Searle is on the CPU--Searle may not in fact have a counter-example to the claims of strong AI. As has been repeatedly pointed out, no one should expect Searle to understand in the Chinese Room example, and his not understanding does not imply that the entire system does not understand. At one point in the article (Searle, 1980; p. 418) Searle does claim that the "computer" is nothing other than him. But in that case, that "computer" would most certainly not understand, or have the potential for understanding. The reason for that is that since Searle is only the CPU, and if the Chinese Room is only Searle, then that is not equivalent to a computer since a computer has more than a CPU relevant to its operations. In that case, although the systems reply would be wrong in saying the whole system understood though Searle did not (it would be wrong if the entire system is Searle and it is accepted that he does not understand Chinese), that would not really matter since Searle has not described a system which is formally/functionally the same as a computer, and that is what he needs to have a possible counter-example to the claims of strong AI. However, even though Searle does say at one point that the "computer is me", given the problems this entails for Searle's counter-example, and the fact that he has the intention of describing a counter-example which is functionally equivalent to a Turing Machine, it would be in Searle's own interests to ignore his stating that this computer is nothing other than Searle.

Even if Searle doesn't have a counter-example to the claims of strong AI and functionalism, he would probably still want to maintain that a formal symbol manipulator could not have intentionality, understand and

have other cognitive states, even if he hasn't shown it to be so by the use of Chinese Room example. It is at this point that I want to introduce another type of criterion for saying whether something has or has not got intentionality, understanding and other cognitive states.

In the first chapter of his book Brainstorms, Daniel C. Dennett considers a criterion for intentionality very different from Searle's, and one that is in keeping with the claims of strong AI and functionalism. Dennett explains an intentional system as one whose "behavior can be-at least sometimes-explained and predicted by relying on ascriptions to the system of beliefs and desires (and hopes, fears, intentions, hunches,...)." (Dennett, 1978; p. 3) When we describe/explain a system in this way we are taking what he calls the "intentional stance". He goes on to qualify this definition of an intentional system by saying that as he has defined them, a particular system is to be considered as an intentional system "only in relation to the strategies of someone who is trying to explain and predict its behavior." (Dennett, 1978; p. 3 ff)

According to Dennett, the intentional stance is just one of several stances we can take towards something. Instead of taking the intentional stance towards something one could, for example, take what Dennett calls the design stance towards it. If you took the design stance towards something, you would explain its behavior based on how its different parts function. For example, "As the typewriter carriage approaches the margin, a bell will ring (provided the machine is in working order)...." (Dennett, 1978; p. 4) In explaining the behavior of the typewriter, we look to see

how it has been designed. One could also take the physical stance towards something. In that case, one's predictions are "based on the actual physical state of the particular object, and are worked out by applying whatever knowledge we have of the laws of nature." (Dennett, 1978; p. 4) In the case of today's computers, it is very unlikely that one is going to be able to successfully predict behavior from either the design stance or the physical stance, most simply because there are so many critical factors involved in both its design and in its physical make-up that predicting its behavior based on those stances is next to impossible. Because of that fact, it is necessary that we take the intentional stance towards these kinds of computers.

Dennett uses a chess playing computer as an example of a system that is so complex that at least for practical purposes we have to take the intentional stance towards it: "when one can no longer hope to beat the machine by utilizing one's knowledge of physics or programming to anticipate its responses, one may still be able to aviod defeat by treating the machine rather like an intelligent human opponent." (Dennett, 1978; p. 5) So one may end saying things like: the computer *believes* that it will checkmate its opponent in three moves, instead of saying that the computer's program is designed so that it will checkmate its opponent in the next three moves.

Although one could take the unusual step of taking the intentional stance towards, for example, a light switch (in which case you would say things like: it now believes that it is switched on) there is no reason to

take such a stance to things as simple as a light switch. One can easily enough predict what these kinds of things will do from either the physical or the design stance. On the other hand, when it comes to predicting the behavior of a human being, an animal, or that of some of today's more complex computers, the intentional stance is wholly appropriate. I want to claim that if the complexity of something, its functional organization, and its input and output merits the intentional stance, then there is all the reason you need for saying that it is an intentional system. Although you could try taking this stance towards a light switch, or something equally simple, it would be wholly inappropriate (see Lycan, 1980; p. 435). (For example, if you were to try and treat a light switch as intentional, you would soon see that it only had two "beliefs": the belief that it was on, and the belief that it was off. But nothing that we normally treat as intentional is limited to only two beliefs.) Dennett claims that we can at will either choose to take this stance towards something, or not take it towards something such as a computer. But it is quite obvious in the case of a human being, for example, that if we do not take the intentional stance, we will not be able to successfully predict their behavior. That may also become more and more obvious in the case of computers as they, and the programs they run, become more and more complex. There isn't any better reason for taking the intentional stance towards other human beings, than there is in the case of a computer, if both perform in a functionally equivalent way--or at least in those ways that are going to be relevant.

Searle, on the other hand, claims that describing a computer's behavior in an intentional way is misleading, since, according to Searle, it "really" doesn't have beliefs, desires, and so on, although human beings do. According to Searle, intentionality is something other than a way of viewing and dealing with a human being. It is a real quality of that person, and it is something that a computer does not have in virtue of running a purely formal program. But Searle does not appear to have any basis for saying this other than his simply claiming that intentionality in all likelihood depends on the specific biochemistry of the being involved just as lactation, photosynthesis and other biological phenomena depend on theirs. But unless Searle wants to claim that intentionality is some kind of physical stuff, there is apparently no good reason to suppose that it depends on a certain sort of biochemistry. There are relevant differences between, for example, lactation and intentionality so that in the former case one is talking about a physical phenomenon that we can touch, smell, see and so on, but in the case of intentionality, none of these things appear to apply. Searle finds it strange that although no one supposes that a computer simulation of a thunderstorm will produce rain, many have claimed that a computer simulation of understanding will really "produce" understanding (Searle, 1980; p. 423). But this may not appear so strange if one believes that understanding is relevantly different from rainstorms or other natural phenomena (see Carleton, 1984), so that in the latter case one will only expect certain kinds of stuff capable of producing rain, while in the former case, one does not associate understanding, intentionality and so on, with the particular stuff that has these properties. There are many things that are square, but no one

expects them all to be made out of the same stuff. Why can't intentionality, for example, be like squareness, rather than rain?

It seems more appropriate to take intentionality as merely a convenient way of describing something's behavior, as well as its level of relevant functional complexity than as some kind of physical phenomenon like rain. Taking the intentional stance towards a human being is much more convenient and successful than taking, for example, the design stance towards a human being, which is likely to be highly impractical if not impossible. If taking the intentional stance is not for any better reason than because of its predictive, explanatory value, then there is no good reason not to take it towards computers when it is of more value than the physical or design stance.

In the case of Searle in the Chinese Room example, there was no good reason for describing his behavior and functions as the CPU as intentional since they could easily be accounted for by taking, for example, the design stance towards him. But if there was anything else--along with Searle--that was relevant in the Chinese Room and was able to pass the Turing Test, then since that entire system, and not just Searle, was dealing with things like, beliefs, hopes, questions, answers, and so on, there is good reason to say that it was intentional and therefore understood Chinese.

There may be certain systems to which we mistakenly attribute intentionality, but that would be because their organization was not suitable and complex enough to provide for the behaviors of what we

should describe as an intentional system. For example, I could easily enough create a computer program that printed the question: Do you think I am happy today? That kind of behavior could be exhibited by an intentional system, but that is not enough to say that this computer is intentional, or rather, should be taken as intentional. You must also be able to say that this computer wants your opinion, that it believes you have one, but none of these things would be exhibited by any of its further behavior since it is not complex enough an organization to have these kinds of intentionality. The same would be true in the case where the Chinese Room is nothing else than Searle acting as a CPU. Since that kind of functional organization can be explained in other ways, there is no reason to take it as intentional, or to have understanding. But as I have mentioned above, unless you take intentionality as some kind of mysterious physical property of a thing, there is no good reason not to take this intentional stance towards computers when they are relevantly the same as ourselves. When the intentional stance is taken towards a human being there is no good reason for supposing that that implies ascribing something more to the human being than we would towards a computer that is also appropriately described in an intentional way.

BIBLIOGRAPHY

BIBLIOGRAPHY

Anderson, Alan Ross, ed. (1964). Minds and Machines, Englewood Cliffs, NJ: Prentice-Hall.

Boden, Margaret (1977). Artificial Intelligence and Natural Man, New York: Basic Books.

Carleton, Lawrence R. (1984). "Programs, Language Understanding, and Searle," Synthese, 59, 219-230.

Churchland, Paul M. (1984). Matter and Consciousness, Cambridge, MA: Bradford Books/The MIT Press.

Cuda, Tom (1985). "Against Neural Chauvinism," Philosophical Studies, 48, 111-127.

Dennett, Daniel C. (1980). "The Milk of Human Intentionality," The Behavioral and Brain Sciences, 3, 428-429.

(1978). Brainstorms, Cambridge, MA: Bradford Books/MIT Press.

Dreyfus, Hubert L. (1979). What Computers Can't Do, New York: Harper & Row.

Dreyfus, Hubert, and Stuart Dreyfus (1985). Mind Over Machine, New York: Macmillan/The Free Press.

Fodor, Jerry A. (1981). Representations, Cambridge, MA: Bradford Books/The MIT Press.

(1980). "Searle On What Only Brains Can Do," The Behavioral and Brain Sciences, 3, 431-432.

Haugeland, John (1985). Artificial Intelligence, Cambridge MA: Bradford Books/The MIT Press.

ed. (1981). Mind Design, Cambridge, MA: Bradford Books/The MIT Press.

(1980). "Programs, Causal Powers, and Intentionality," The Behavioral and Brain Sciences, 3, 432-433.

(1978). "The Nature and Plausibility of Cognitivism," The Behavioral and Brain Sciences, 1, 215-226.

Hofstader, Douglas R. (1980). "Reductionism and Religion," The Behavioral and Brain Sciences, 3, 432-433.

Lycan, William G. (1981). "Form, Function, and Feel," Journal of Philosophy, LXXVIII, 24-50.

(1980), "The Functionalist Reply," The Behavioral and Brain Sciences, 3, 434-435.

McCorduck, Pamela (1979). Machines Who Think, San Francisco: W.H. Freeman.

McDermott, Drew (1976). "Artificial Intelligence Meets Natural Stupidity," SIGART, No. 57 (reprinted in Haugeland 1981).

Rey, Georges (1986). "What's Really Going On In Searle's 'Chinese Room'," Philosophical Studies, 50, 169-185.

Schank, Roger C. (1986). Explanation Patterns, Hillsdale, NJ: Lawrence Erlbaum Associates.

(1984). The Cognitive Computer, Pennsauken, NJ: Addison-Wesley.

(1980). "Understanding Searle," The Brain and Behavioral Sciences, 3, 446-447.

Searle, John R. (1984). Minds, Brains and Science, Cambridge, MA: Harvard University Press.

(1983). Intentionality, an Essay in the Philosophy of Mind, Cambridge, England: Cambridge University Press.

(1980). "Minds, Brains and Programs," Behavioral and Brain Sciences, 3, 417-424 (reprinted in Haugeland, 1981).

Turing, Alan (1950). "Computing Machinery and Intelligence," Mind, 59, 434-460 (reprinted in Anderson, 1964).

VITA

Kevin Martin Roddy was born in Larne, County Antrim, Northern Ireland on May 8, 1963. He attended elementary school in that town and was graduated from St. MacNissi's College, Garron Tower in June 1981. In September 1982 he entered The University of Ulster, Jordanstown, and in June 1985 he received a Bachelor of Arts Degree in Philosophy. In the fall of 1985 he accepted a graduate assistantship at The University of Tennessee, Knoxville and began study toward a Master's Degree with a major in Philosophy. This Degree will be awarded in August 1988.

Mr Roddy will enter The Queen's University, Belfast in October 1988 to begin research for his Doctoral Dissertation.