Against the Chinese Room Argument

By
Robert James Wood

A Thesis Submitted to
Saint Mary's University, Halifax, Nova Scotia
in Partial Fulfillment of the Requirements for
the Degree of Master of Arts (Philosophy)

Dr. Peter March
Thesis Supervisor

Dr. Mark Mercer
Internal Examiner

Dr. Darren Abramson
External Examiner

Date: December 6, 2007
Revised: August 1, 2008

Your file Votre référence
ISBN: 978-0-494-44669-0
Our file Notre référence
ISBN: 978-0-494-44669-0

Against the Chinese Room Argument

By: Robert James Wood

Abstract

The Chinese Room Argument is a *reductio ad absurdem* argument, intended to show that a contradiction follows from the principles of the theory of computer functionalism. If a contradiction follows from these principles, then the Chinese Room Argument is a good argument. This thesis argues that the Chinese Room Argument is invalid; that is does not demonstrate a contradiction, and, therefore, is a bad argument against computer functionalism. The argument of this thesis is made in four chapters. (1) An analysis of the concepts employed in the original presentation of the Chinese Room Argument. (2) An analysis of the principles of computation addressed by the Chinese Room Argument. (3) An analysis of category mistakes made in philosophies of mind. (4) An argument that the category mistakes enumerated in (3) occur when the concepts in (1) are not properly applied to the principles in (2).

August 1, 2008

Table of Contents

Thesis Introduction

Chapter 1

Chapter 2

Chapter 3

Chapter 4

0.0 Thesis Statement

Computer functionalism in the philosophy of mind is, broadly speaking, the thesis that minds are like computer programs and that the properties of mind are those of software. Using the Chinese Room Argument, Searle argues that the thesis of computer functionalism is false in principle. The goal of this thesis is to show that John Searle's Chinese Room Argument is not a good argument against the thesis of computer functionalism. I will argue that Searle's conclusion about how computer programs necessarily lack mental properties such as 'understanding' does not follow from the premises that he adopts in his description of the Chinese Room Argument.

0.1 Method

In order to demonstrate this thesis, I will refer to material from five authors: John Searle, Alan Turing, Daniel Dennett, Richard White, and Gilbert Ryle. Work by the first four authors will be used to illustrate the logical form of the Chinese Room Argument; work by the fifth author will be used to show its logical errors. Searle presents the original version of the Chinese Room Argument and the metaphysical framework justifying its validity. Turing introduces the concepts of digital computation, which form a different metaphysical framework from that employed by Searle. White further explains the metaphysics of digital computation with regards to concepts invoked in the Chinese Room Argument, such as levels of abstraction/description, multiple realizability, and computability. Dennett discusses the Chinese Room Argument in Turing's metaphysical framework of digital computation. Finally, Ryle identifies a type of logical error that occurs in arguments of the same logical form as the Chinese Room Argument, in a metaphysical framework matching the concepts of digital computation.

My argument will be that although Searle's Chinese Room Argument is valid when making inferences about the metaphysics of computation described by Searle, it is invalid when making inferences about the metaphysics of computation described by Turing, White, and Dennett. This argument will be carried through in three steps, in four chapters. The first three chapters will introduce each of the three steps in the argument, while the fourth chapter will present the argument as a whole.

The first chapter of this thesis will introduce the Chinese Room Argument, and the metaphysical framework that Searle uses to certify its validity. I will note that Searle is using the Chinese Room Argument against the principles of computer functionalism, which he calls the "Strong AI Thesis", and review the rebuttals that Searle gives to various criticisms of the Chinese Room Argument in order to establish its validity (Searle, p.282). Briefly, the Chinese Room Argument is that if instantiating a program adds no additional mental properties to whatever instantiates that program, despite the appearance of doing so, then programs in general necessarily lack any mental properties. Given Searle's own explanation of the Chinese Room Argument, and the methods and terms that he employs in its construction and defence, I will argue in the conclusion to chapter one that the Chinese Room Argument appears to be valid. Furthermore, I will argue, in agreement with Searle, that if the Chinese Room is valid, then it presents an intractable problem for any theory of computer functionalism resembling the Strong AI thesis. This will be a good argument against the Strong AI thesis.

The second chapter of this thesis will introduce the concepts of digital computation, and particularly the concept of the Turing Machine, and its generalization in the Universal Turing Machine. The Universal Turing Machine is the model of digital computation addressed by the Chinese Room Argument. It is a schema of computability capable of computing any Turing-computable function and, thereby instantiating any special purpose Turing Machine. It is an abstract automaton that can be instantiated by a variety of different concrete machines. After noting some key concepts involved in the relation of such Machines to machines, via exegesis of White and Turing, I will discuss how the Universal Turing Machine is related to the test for the presence of putative mental properties that Turing calls the “Imitation Game” (Turing, pp.265-266). I will follow Turing’s discussion of the Imitation Game with Dennett’s analysis of that Game, the ‘Turing Test’. I will do so in order to demonstrate how Searle’s assumptions about the logic and limits of computability, multiple realizability, and abstraction in the Chinese Room Argument are not accurate and do not address the Strong AI thesis according to the conventional metaphysics of digital computation.

In the third chapter I will examine logical errors as they apply to reasoning about phenomena such as that of minds. I will introduce both Ryle’s notion of a category mistake and the metaphysics of the ‘Official Doctrine’ that involve that kind of mistake. I will use this chapter to explain how the Category Mistake is a type of logical error, and discuss two distinct category mistakes that Ryle describes (Ryle, p.17). The first mistake that Ryle describes involving categories leads us to think the mind to be a non-physical thing, to reify it. Such a category mistake is made when a category, such as “university”, is numbered amongst its own members such as specific faculty buildings (Ryle, p.17-18). The second mistake that Ryle describes involving categories also leads us to suppose the mind is a physical thing. Such a mistake is made when something in one category, such as “esprit de corps”, is numbered amongst the members of another category such as positions on a cricket team (Ryle, p.18).

Finally, in the fourth chapter of this thesis, I will put the contents of the first three chapters together. I will first substitute the metaphysics of digital computation for Searle's metaphysics of symbol manipulation. Once the Chinese Room Argument has been put into a context from which it can be a good argument against the Strong AI thesis, I will go on to show that it is invalid and therefore a bad argument against the Strong AI thesis where Searle purposes to show that thesis to be wrong in principle. Where Ryle points out two particular category mistakes to argue against the mis-categorization of mental predicates as spurious mental entities or the wrong category of things, I will point out the same two mistakes in the Chinese Room Argument. Therefore I will argue in the conclusion to this chapter that, considered within the metaphysical framework of digital computation described by Turing, White, and Dennett, the Chinese Room Argument commits both category mistakes in deducing its conclusion.

0.2 Literature Review

As mentioned in section 1.1, this thesis is concerned with the papers of five authors, John Searle, Gilbert Ryle, Alan Turing, Daniel Dennett, and Richard White. Each of these authors provides an important insight into the problem at hand. Despite these equally important contributions, however, the work of these authors can be divided into two groups: the works that directly address the subject of the Chinese Room Argument, and the works that address the argumentation and metaphysics involving that subject. The former includes: Alan Turing's "Computing Machinery and Intelligence", and John Searle's "Minds, Brains and Programs." The latter includes: "Can Machines Think?" by Daniel Dennett, Gilbert Ryle's "Descartes' Myth", and "Some Basic Concepts of Computability Theory" by Richard White.

Alan Turing's article "Computing Machinery and Intelligence" provides a formal model of computation and an alternative to the question "Can machines think?" The alternative proposed by Turing is the question whether a digital computer could successfully imitate the "...intellectual capacities of a man" (Turing, p.265). He hypothesizes the answer to be affirmative, that a digital machine might, in principle, successfully imitate the intellectual capacities of men. Hence Turing considers that while this imitation may appear to be like thinking, it may not be thinking. This is the contention of Searle's argument, that regardless of how well a machine may imitate some capacity, it may not be duplicating that capacity.

Turing also says that: "we need not be troubled by this objection" if "a machine can be constructed to play the imitation game satisfactorily" (Turing, p.266). The principles behind the imitation game are thus the principles by which Turing believes something like the Strong AI thesis can be tested. Searle disagrees, and argues to the contrary that we should be troubled by this objection, and that we should reject out of hand the hypothesis that a digital computer could ever successfully imitate the intellectual capacities of a man. This dispute is the issue at hand. Clearly, Turing's description of the imitation game, the machines involved and their capacities, and the significance of passing the imitation game are relevant as a contrast to the content of Searle's article.

Richard White helps to further explain the metaphysics of Turing's model of computation, the Turing Machine, to the instantiations, or tokens, of that model. In particular, he points out an essential misunderstanding on the part of Searle of what might instantiate a digital computer, and what the relation of programs are to the mechanisms that carry them out. In doing so, White's article supplements Turing's description of digital machines by showing directly how Searle's incorrect exegesis of the digital computer affects the Chinese Room Argument. These misunderstandings facilitate Searle's development of a logical framework in which his conclusion about the capacities of a digital computer is deductively valid, but in conflict with that of Turing.

The crux of this disagreement can be found in the first chapter of Gilbert Ryle's The Concept of Mind, "Descartes' Myth". This article provides an analysis of mental terms as predicates rather than terms denoting objects. Using this analysis, Ryle argues against mental predicates, most notably as properties of a mental object or Mind. Ryle calls this notion of mind-as-object the "Cartesian Myth" and the type of logical error made in reaching that conclusion the "Category Mistake". While the Cartesian Myth is certainly pertinent to a discussion about the ontological status of mind, it seems that the Category Mistake Ryle identifies, and its complement in the logic of categorizing mental predicates correctly, is more directly relevant to this thesis: If I can show that Searle make a category mistake in the Chinese Room Argument, I have demonstrated my thesis.

A more rigorous argument will be made in the body of this thesis to the effect that Searle makes category mistake with regard to programs that Ryle warned against in the case of minds. Where the properties cited by Searle, such as understanding, are categories of physical states rather than something non-physical, the deduction that computer programs cannot, in principle, have mental properties is invalid. Ryle's argument is used to show that where mental predicates such as understanding are a matter of being in the appropriate states, we need not be troubled by the conclusion of the Chinese Room Argument, that understanding is not necessary for the appropriate response to take place.

John Searle's article "Mind, Brains and Programs" presents the Chinese Room as a model of the formal digital computation proposed by Turing. The Chinese Room is simply a human being providing the mechanisms by which a digital computer, a

Universal Turing Machine as proposed by Alan Turing in his article, might engage in computation. The Argument is that where a Chinese Room is concerned (and thus all digital computers) such a machine can successfully pass the imitation game using Chinese symbols without understanding their meaning. Such a computer might pass the Turing Test, but in doing so it will not successfully imitate the intellectual capacities of a Chinese-literate person who understands the meaning of the symbols they employ.

Quite usefully, the version of "Mind, Brains and Programs" that I will use in this thesis contains Searle's report and consideration of replies made to his Chinese Room Argument by artificial intelligence researchers. In reporting, considering, and answering these replies Searle more fully qualifies the problems that he thinks the Chinese Room demonstrates with the Strong AI thesis. In particular, by addressing these replies, Searle explains his concern about the property of intentionality with regard to properties like understanding. Properties like understanding, he argues, are the result of brains having a biological property such that they can be about something. Describing the metaphysics by which machines such as brains can have *intentional* properties, Searle further sketches out his argument against digital machines like the Chinese Room having capacities with that property. This involves discussion of such concepts as a syntax/semantics dichotomy, levels of description, and a relation between such levels called 'instantiation'.

Finally, Daniel Dennett, in his article "Can Machines Think", helps to explain the relevant properties of the argument that Turing employs in proposing the imitation game. Contrary to Searle's understanding of the imitation game, and consistent with a Category Mistake-free reading of Turing's argument, Dennett argues that by winning the imitation game a computer does license us to believe that what the computer is doing when it plays that game is importantly like what a person is doing when they play the imitation game.

In doing so, Dennett provides a conclusion that is relevant by its contrast to Searle's invalid conclusion that passing the imitation game fails to indicate understanding. Instead, passing the imitation game not only indicates understanding, it does so in a way that makes Turing's dismissal of Searle's concerns seem less pre-emptory. In this article Dennett argues that if imitation may appear to be like thinking, then what is important is how much it appears to be like thinking rather than whether it is the result of thinking or not.

Altogether, I suggest that a conjunction of arguments found in these articles shows two arguments that draw contradictory conclusions from the same premises, and a way of determining which argument is valid. The conclusion of one argument is that artificial intelligence, in the strong sense of being something like natural intelligence, is possible. The conclusion of the other argument is that artificial intelligence in that strong sense is not possible. Two articles represent these two main arguments, Turing and White against Searle. Three articles (White's, Ryle's and Dennett's, respectively) represent the meta-arguments about the qualities of the two main arguments. The latter three articles act, as it were, as a filter by which the two main arguments can be compared and contrasted. Insofar as all of these articles are concerned, and in the absence of further information about them and their merits, it appears that the Chinese Room Problem that Searle presents does not entail his conclusion that qualities such as understanding are necessarily absent from digital computers.

1.0 Chapter Introduction

Recall that in the introduction I noted that the first chapter of this thesis would introduce the Chinese Room Argument, explain the concepts employed by John Searle in its construction, and explain the reasoning he uses to derive a valid conclusion from his premises. I hope to show that, as given, the Chinese Room Argument presents an intractable problem for any theory of computer functionalism resembling the Strong AI thesis described by Searle. By 'resembling' I mean that a theory of computer functionalism that is identical to the Strong AI thesis described by Searle, and may be substituted for it in a manner that preserves the accuracy of the description. By 'intractable problem' I mean a problem that falls outside the scope within which the principles of a theory can be used to compute a solution. The Chinese Room Argument presents an intractable problem for the Strong AI thesis if it shows that thesis to be false in its principles. An example of an intractable problem would be computing a solution to 'the *colour* of *grue*' in a finite arithmetic that admits neither 'colour' nor 'grue' as defined terms.

Briefly, the Chinese Room Argument is that if (P1) computer programs are purely syntactic in nature and may have semantic properties, and (P2) being purely syntactic in nature excludes something from having semantic properties, then (C) those programs necessarily lack semantic properties, such as that of intentionality. Given briefly and informally this argument seems sound. Where computer programs are purely syntactic, and a Chinese Room instantiates only those syntactic properties that allow it to resemble whichever program allows it to successfully play the Imitation Game, then there is no room in them for the semantic component that Searle claims minds to have.

As a positive argument for what might have intentional properties, Searle offers the human brain as an example of both the instantiation of a computer program and an object with semantic properties. According to Searle mental properties such as the property of intentionality, a semantic component of human brains and thus human minds, are biological properties. He also differentiates between what he calls the 'intrinsic intentionality' found in human brains and the 'extrinsic intentionality' or 'information' found in artefacts such as books and natural objects such as tree rings. The latter is derived from the former, so that the meaning that an observer attributes to the symbols manipulated by devices is a property of the observer rather than a property of the symbols themselves. Searle's levels of description are divided into physical objects with physical (including biological) properties, and non-physical objects with syntactic or structural properties.

Given that Searle conceives of digital computers as having non-physical properties derived from observers of physical objects that engage in "manipulating uninterpreted formal symbols", it follows that no property of a syntactic system could be a semantic property (Searle p.284). That is, where syntax is the relation of symbols to each other, and where semantics is the relation of symbols to whatever fixes their meaning, and where these two sorts of relation are distinct, it seems quite reasonable that Searle would find in his Chinese Room Argument an intractable problem for the Strong AI thesis. It is a case of mistaken identity where identity is at issue. However, simply saying that it seems to be reasonable raises the question of what counts as a good reason. Rather than merely remarking that this argument is reasonable, then, I will seek in this chapter to demonstrate that it is so.

In the following sections I will first present an informal exegesis of "Minds, Brains and Programs" based on the three sections into which it can be divided. I will draw some tentative conclusions and note some possible critiques of the analysis that I have given. Then I will set aside the conclusions of this chapter until Chapter Four when I will bring the conclusions of this chapter into the context developed in Chapter Two and into the light of the critique cast in Chapter Three.

1.1 The Chinese Room Argument

Searle introduces the Chinese Room Argument with a discussion of precisely what he means when he argues against the possibility of artificial intelligence. There are, he says, two different sorts of artificial intelligence, "strong AI" and "weak AI" (Searle, p.282). Weak AI is the idea that the digital computer is simply "a very powerful tool" which allows "us to formulate and test hypotheses in a more rigorous fashion" (Searle, p.282). Weak AI is the project of refining those tools. Strong AI is the thesis that "the appropriately programmed computer really *is* a mind" (Searle, p.282). Searle explains the difference as the difference between testing "psychological explanations" and being "themselves the explanations" (Searle, p.282).

Having established the target of his arguments, Searle turns his attention to specific exemplars of the Strong AI thesis. An exemplar of artificial intelligence for Searle is a computer program, of the sort used by Roger Schank to "simulate the human ability to understand stories" (Searle, p.283). Searle describes Schank's computer program as being able to answer questions about the story like humans would "even though the information that they give was never explicitly stated in the story" (Searle, p.283). The program ostensibly has the background information that a human might use to answer questions about the story, in the form of a script about the story. Searle claims that a proponent of the Strong AI thesis would say that, in answering questions about the story, Schank's program or some program like it simulates the human thought process and duplicates this process such that it "can literally be said to *understand* the story" (Searle, p.283). How it duplicates this process is an explanation of "the human ability to understand the story and answer questions about it" (Searle, p.283). Searle employs the Chinese Room Argument to dispute that a program, and hence this program, understands the story and that how it understands the story to explain the human ability to understand the story.

At this early juncture I would like to point out two things. First, the division that Searle draws between the Weak AI project and the Strong AI thesis is a false dichotomy. If a dichotomy is a choice between two mutually exclusive options, which together exhaust the field of options, then a false dichotomy may be two options that are not exclusive, or it may be two options that do not exhaust the field of options, or it may be both. That is why I am referring to one as a "project" and another as a "thesis". According to Searle's interpretation of how the Strong AI thesis might apply to a story-answering program like that of Schank, the Weak AI and the Strong AI concepts seem to be interchangeable at two points. Where the Weak AI project is to simulate human thought processes, and use the properties of the simulations to help explain human mental powers, the Strong AI thesis is that these simulations can duplicate human thought processes, and that the properties of these electronic minds may be used to explain human mental powers.

Supposing that 'help explain' is interchangeable with 'explain', if we assume no finality of explanation such that any explanation only helps to explain, then there seems to be little difference between the Strong AI thesis and the Weak AI project. The difference, such that there is one, is one of scope. I suggest the difference can be thought of like this: The scope of the Weak AI project is the project to model the human thought process using computers, while the Strong AI thesis is the thesis that at least one model has a 1:1 congruency with the phenomenon it models, that at least one model is accurate. Since the Weak AI project and the Strong AI thesis are not mutually exclusive options, the dichotomy that Searle draws between them is false. This is the first indication that Searle is misunderstanding the metaphysics of computation.

The second thing I would like to point out is that, if Searle is actually putting forth a false dichotomy between the Weak AI and the Strong AI, then the fact that Schank's program in particular might genuinely fail in principle to duplicate human thought processes does not falsify the Strong AI thesis in general. It may be the case that, even if Schank's program and programs like it fail to duplicate human thought processes, a different program using a very different process will succeed in duplicating human thought processes. The principles that might genuinely fail to duplicate human thought processes are only vaguely defined. Charitably supposing that there is only one bona fide Human Thought Process, showing how one model does not duplicate all of that process's properties does not show that no model can duplicate all of the properties of that Process. Indeed, showing how one process fails to duplicate all of the properties of another process suggests exciting avenues within the Weak AI project according to which the Strong AI thesis may yet be proven true. If only one type of process fails to duplicate any properties of the process it models, then there may be processes that can rightly be called programs, but fail to be the right kind of program. The principles that fail must be those of computation in general, rather than any particular computer program. Hence Searle lays out his assumptions for his "*Gedankenexperiment*", the Chinese Room Argument, with the intent to prove this impossibility in the principles of computer science.

As a thought experiment, Searle explains, the purpose of the Chinese Room Argument is to call our attention to what it might be like "if my mind actually worked on the principles that the theory [Strong AI] says all minds work on" (Searle, p.283). In proposing any experiment, whether genuine empirical test or more subtle argument about principles, some note should be given to the experimental or argumentative apparatus lest the results it returns are artefacts rather than data about its subject. So, in the following, recall the principles of the theory of Strong AI that Searle has laid out so far:

1) The digital computer may simulate human thought processes (which Searle does not dispute, as it is a principle the Strong AI thesis reiterates from Weak AI project in which it is embedded).
2) The simulation of human thought processes duplicates their mental states (such as *understanding*).
3) The process of the simulation explains (or helps to explain) mental states in humans (which Searle does not dispute in the case of Weak AI and disputes in the case of Strong AI).

Furthermore, I should note that Searle does not refer to 'the process of the simulation', but rather "what the machine and its program do" (Searle, p.283). I will shortly relate how Searle justifies this curious grammatical construction, rather like 'what a book and its words do', but for now it should be noted that he apparently considers these things to be distinct.

So what is the experimental apparatus that Searle employs? He employs a locked room, three batches of Chinese symbols, a booklet of rules, and John Searle the Chinese-Illiterate. John Searle the Chinese-Illiterate is locked in the room with the booklet of rules and is passed the batches of symbols (perhaps under the door). Recall that Schank's program involved a computer program that gave answers to questions about a story, the telling of which leaves plenty to the imagination, and requires one to understand the story. In Searle's model, the Chinese Room, John Searle is the machinery that implements this 'Chinese story' program.

The first batch of Chinese symbols is the "script" in Schank's model, the second batch is the "story", and the third batch corresponds to the "questions". Symbols from the script include the symbols that Searle describes as "the symbols I give them back in response to the third batch" or "the answers" (Searle, p.284). The booklet of rules is the "program" which enables John Searle the Chinese-Illiterate to transform the questions about the story he receives into the answers he passes out of the room (Searle, p.284). The rules allow Searle to compare the questions to the story, the story to the script, and extract answers from the script that, for the sake of argument, read as though composed by somebody who is Chinese-Literate.

Furthermore, Searle contrasts the situation described by the Chinese Room with the situation that appears to be the case otherwise, when an agent performs the same task with understanding. John Searle the Chinese-Illiterate is also John Searle the English-Literate. John Searle can perform the same 'answering-questions-about-what-stories-imply' game without using the booklet of rules, if the game is conducted in English. Unlike the answers mechanically formed in Chinese, Searle says that the answers in English are "indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker" (Searle, p.284). The answers about a story given by John Searle in English thus differ from those given by John Searle in Chinese in an important aspect: Searle understands the English and does not understand the Chinese. Whereas Searle finds the answers when the story is presented in English by understanding the relation of the questions to the story, he produces the answers to the story when presented in Chinese "by manipulating uninterpreted formal symbols" (Searle, p.284). The Chinese Room Argument thus depends upon a metaphysic of mental properties such as understanding being something other than the mechanical manipulation of formal symbols or computation as understood by Searle.

Because John Searle lacks the property of understanding in relation to the story, questions, and answers given in Chinese, the conclusion of the Chinese Room Argument is that the Chinese Room and all other physical instantiations of computer programs (the digital computers) similarly lack the property of understanding in relation to the formal symbols they manipulate. Searle is using the conclusion of the Chinese Room argument against the thesis that a program could simulate the human thought process and duplicate that process such that it "can literally be said to *understand* the story" (Searle, p.283). The 'answering-questions-about-what-stories-imply' program simulates at least one sort of human thought process: it answers questions about what stories imply. If Searle's conclusion regarding the Chinese Room argument is correct, however, that program does not duplicate that thought process. The Chinese Room in the eponymous argument does not duplicate that thought process because it cannot be said to understand the story it provides answers about. Hence the Chinese Room Argument contradicts at least one of the principles of the Strong AI thesis; that the simulation of human thought processes by digital computes could duplicate their mental states.

Given that the Chinese Room Argument has shown that a digital computer does not duplicate a human thought process, given certain assumptions about digital computers, Searle must generalize this conclusion to show that computer programs cannot duplicate any human thought processes. Clearly Searle understands that arguing against a specific computer program's inability to duplicate human thought processes will not show that all computer programs share this disability when he notes that "the computer and its program do not provide sufficient conditions of understanding since the computer and the program are functioning, and there is no understanding" (Searle, p.284). After all, Searle's argument is concerned with showing that the Strong AI thesis is false in principle, and not simply in the details of any one program such as Schank's. Referring to the Chinese Room set-up he asks: "But does it even provide a necessary condition or a significant contribution to understanding?" If the Chinese Room lacks even necessary conditions for understanding, then that lack would generalize the conclusion of the Chinese Room Argument to show that computer programs cannot duplicate human thought processes.

Clearly, a lack of properties in common between the 'answering-questions-about-what-stories-imply' computer program, and the human thought process of answering questions about what stories imply, would generalize the conclusion of the Chinese Room Argument by showing that the Chinese Room lacks even a necessary condition for properties of the same type as understanding. If this program did not have any properties in common with the thought process, then it follows that programs in general could not have some necessary condition in common with thought processes. As a result of not duplicating the thought process, by dint of having an exclusive set of necessary and sufficient conditions for being, it follows that no program could duplicate a thought process because their computational properties are exclusive with mental properties. The Chinese story program would not duplicate the thought process because programs do not have any properties in common with the human thought process of answering questions about stories. As a corollary it follows that no program would lend insight to, or explain, human thought processes, being so completely different.

In order to show that computer programs and thought processes have exclusive sets of properties, Searle makes a counter-argument to the following claim: "when I understand a story in English, what I am doing is exactly the same – or perhaps more of the same – as what I was doing in manipulating the Chinese symbols" (Searle, p.284). Searle says that this claim is based on the assumptions that there is a program with "the same inputs and outputs as native speakers, and in addition we assume that speaker to have some level of description where they are also instantiations of a program" (Searle, p.284). The counter-argument is that, given that the program seems irrelevant to the possession of understanding, it seems that construing native speakers themselves as being instantiations of programs giving the appropriate outputs to the input is likewise irrelevant to the possession of understanding. Searle says in conclusion to this counter-argument:

> "Rather, whatever formal principles you put into the computer, they will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything. No reason whatever has been offered to suppose that such principles are necessary or even contributory, since no reason has been given to suppose that when I understand English I am operating with any formal program at all" (Searle, p.285).

I suggest that this generalization of the argument against Schank's program may be expressed as the following: Computer programs operate according to formal principles. Formal principles appear to be irrelevant to human thought processes and the possession of properties such as 'understanding'. Therefore computer programs have a set of properties that are mutually exclusive of human thought processes. It follows then, that instantiating a program such as the Chinese symbol manipulator described in the Chinese Room Argument will fail to add any semantic or mental properties such as 'understanding' because it is not the sort of thing that has those properties.

Before ending this section, however, I should mention Searle's discussion of the property of understanding. Searle says, in reply to critics who suggest he may be ignoring the nuances of attributing understanding: "There are clear cases in which 'understanding' literally applies and clear cases in which it does not apply; and these two sorts of cases are all I need for this argument" (Searle, p.285). This rules out the metaphorical attribution of mental states, by making the Chinese Room Argument about whether the Chinese Room literally understands Chinese rather than metaphorically, if speaking that way about the Chinese Room is an insightful and productive way to discuss computing. Similarly this statement rules out partial cases; it is not about differences in degree, but about differences in kind. So for the purposes of presenting the Chinese Room Argument in this chapter I am going to treat 'understanding' as a single-part predicate, a monadic relation, a property that something either has or does not. The significance of treating properties such as 'understanding' as abstract nonsense will be discussed in Chapter Two, when I address Turing's metaphysics for digital computation. I put emphasis on this bivalency of properties like understanding now because, as Searle suggests, it is needed for the Chinese Room Argument.

1.2 Some Objections to the Chinese Room Argument

In *Minds, Brains, and Programs*, Searle answers a set of counter-proposals about how digital computers could avoid the problem of the Chinese Room and duplicate human thought processes. However, because computation or the formal manipulation of un-interpreted symbols is a matter of syntax, the relation of symbols to other symbols, and so appears to be irrelevant to the 'understanding' that is a matter of semantics, the relation of symbols to their meanings, Searle maintains a distinction between syntax and semantics. In answering each of these proposals Searle reiterates the semantics-syntax distinction elucidated by the Chinese Room Argument. In Searle's opinion they either miss this essential distinction, or they do not support the Strong AI thesis that Searle is addressing. The proposals themselves are named as follows: "The Systems Reply", "The Robot Reply", "The Brain Simulator Reply", "The Combination Reply", "The Other Minds Reply", and the "Many Mansions Reply" (Searle, pp.286-292).

The Systems Reply says that understanding results from the operation of the entire system rather than resulting from a particular component. According to Searle the Systems Reply is inadequate because the observation of a whole system appears to give no more reason to suppose such a system has the first-person experience of mind than the observation of any part of that system does (Searle, pp.286-288). This shows that Searle is aware that John Searle the Chinese-Illiterate is only a part of the Chinese Room, and does not consider it to be problematic. He argues that the other elements of the system such as the booklet of rules and the symbols can be treated the same if the person in a Chinese Room were to "internalize all of these elements of the system" (Searle, p.286). Such a person, supposing it is John Searle again, would then have two such systems inside of them, a Chinese-Literacy system and an English-Literacy system. However, as in the original Chinese Room Argument the person incorporating the Chinese-Literacy system would have no understanding of written Chinese while understanding written English. Searle says of this: "In short, the systems reply simply begs the question by insisting without argument that the system must understand Chinese" (Searle, p.287).

Searle argues that the Robot Reply is similarly deficient. Like the Systems Reply, the Robot Reply suggests that there is some greater system of which the Chinese Room is a part. Although the Robot Reply tries to give computational systems causal efficacy by embodying them as robots, Searle notes that it is inadequate because this embodiment in a robot does nothing to bridge the gap between being describable as a computational system, and actually having a mind. He says: "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" (Searle, p.288). Supposing that the Robot Reply does not concede that cognition is not solely a matter of formal symbol manipulation, Searle describes how such a robot might be conceived of as a Chinese Room. Such a robot, he says, would simply reiterate the Chinese Room's mechanical performance of symbol manipulation: "it [the robot] is simply moving about as a result of its electrical wiring and its program" (Searle, p.289). This would simply reiterate the Chinese Room's performance in answering questions about what a story implies, and "by instantiating the program I have no intentional states of the relevant type" (Searle, p.289).

The Brain Simulator Reply attempts to attack the principles of the Chinese Room Argument; the idea that a Chinese Room is fundamentally different from what it simulates. The Brain Simulator Reply proposes that the processor in the room simulates the physical operations of a Chinese-literate brain, rather than that sort of brain's content

(Searle, pp.289-290). Because the Chinese Room shares sufficient physical properties with Chinese-literate biological systems, as well as formal properties, the brain simulator would presumably have understanding of the symbols that it manipulates. Searle argues that the Brain Simulator abandons a central claim of computer functionalism, that things physically unlike brains may have minds (Searle, p.289). Where a Brain Simulator is limited to sharing the formal properties of a literate brain, Seale claims, it will not simulate "its causal properties, its ability to produce intentional states" such as understanding (Searle, p.290). The Chinese Room is constructed to show this divorce between formal properties and physical properties, and Searle clarifies this by proposing a Chinese Room where the person does not manipulate symbols directly, but via the valves of a set of water pipes (Searle, pp.289-290). Such an arrangement changes the physical properties of the Chinese Room, but maintains its formal properties; it produces the correct sort of answers about what stories imply because of its program.

The Combination Reply proposes that intentional properties such as 'understanding' would be the result of a robot system that simulated the entire human being, brain and all. Perhaps unsurprisingly, Searle claims that the Combination Reply fails for approximately the same reasons as the Systems, Robot, and Brain Simulator Replies do. Although such a robot system with a brain simulator might perform all the functions of a regular Chinese Room and many besides, it would only appear to have intentional states until we were made aware that it was just a complex puppet. It would still only be a machine and its functional description would still only concern whether it manipulates symbols in a manner sufficient to pass.

Even if it is only a difference in degree between such puppets and people, Searle argues, such puppets would be more physically alien to us than "apes and monkeys" (Searle, p.291). Such an alien nature would make such a robot too alien to suppose it had intentional properties regardless of how well its functional identity matches benchmark examples. Given Searle's objections to the prior three replies, though, a difference in degree would be unlikely because, as a formal system, the robot system would seem to lack intentional properties, since the robot system still relies on the properties of a formal system and not its physical system to perform. As a brain simulator the robot system would abandon the central claim of the Strong AI thesis by replicating intentional properties physically rather than formally.

The Other Minds Reply and the Many Mansions Reply both trade on the idea that the difference between a person with intentional states and a machine with a program is a difference of degree rather than kind. The Other Minds Reply suggests that no adequate explanation for literacy is forthcoming. Without this adequate explanation, without knowing the specific property a Chinese-Literate has that a Chinese Room lacks, no effective contrast may be drawn between a Chinese Room and a Chinese-Literate. Searle argues that such a problem in the construction of the Chinese Room Argument has been stipulated away: "The thrust of the [Chinese Room] argument is that it couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive state" (Searle, p.291). The Chinese Room Argument does not require an explanation of literacy to determine the specific property that a Chinese-Literate has that a Chinese Room does not. It turns on the idea that whatever Chinese literacy is, it is something that a Chinese-Literate has and a Chinese Room containing John Searle does not.

Likewise, the Many Mansions Reply suggests that the causal processes differentiating computational machines from conscious machines are restricted to particular classes of machines. The Many Mansions Reply suggests that while digital machines and their computational representations may not be conscious, analog machines may be conscious. Searle argues that the Many Mansions Reply misses the point that the Chinese Room stands as a counter-example to the claim that "mental processes are computational processes over formally defined elements" rather than any broad physical account of mind (Searle, p.291). Whatever machine may have a mind, Searle argues with the Chinese Room Argument, the formal representation of that mind cannot be the mind itself. According to Searle the fact that an analog machine may have a mind is immaterial to a general refutation of the Strong AI thesis because that refutation deals with the incommensurability of formal identity claims with physical identities.

In discussing the Many Mansions Reply, Searle gives a positive account of where intentional properties lie. He says: "I see no reason in principle why we couldn't give a machine the capacity to understand English or Chinese, since in an important sense our bodies with our brains are precisely such machines" (Searle, p.292). According to Searle, intentional properties such as understanding are the result of our physical structure as organisms, "and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena" (Searle, p.292). According to Searle, then, the problem with the Strong AI thesis is that it does not duplicate the relevant properties that cause intentionality, that cause understanding, but instead replaces those properties with representations or simulations.

Searle concludes that computer simulations, models of formal symbol manipulation, are not the things in themselves, but merely pictures of the things they simulate. He says: “All the arguments for the strong version of artificial intelligence that I have seen insist on drawing an outline around the shadows cast by cognition and then claiming that the shadows are the real thing” (Searle, p.292). The claim that minds are programs, according to Searle, is false because pictures of things are not the things being pictured, minds are things but computer programs are just pictures. Searle seems to be saying that if a simulation is a model of formal symbol manipulation, and formal symbolic manipulation is purely syntactic, and does not itself give rise to semantics, it follows that a simulation could not duplicate the semantic properties where those properties are physical. Hence the Chinese Room Argument is valid because the conclusion follows from the premises, and Searle provides a theory in which it is a fact that syntax alone does not give rise to semantics. None of the objections to the Chinese Room Argument reviewed by Searle in *Minds, Brains, and Programs* address Searle’s contention that formal properties are not physical properties, and therefore he rejects them.

1.3 Information and Intentionality

Having explained the principles that he uses to construct the Chinese Room Argument, and to derive its conclusion, Searle further explains the metaphysical principles underpinning the argument. We have already reviewed the conclusion of the Chinese Room Argument, that the addition of a computing program gives a man no additional understanding (Searle, p.293). Given that the Chinese Room Argument essentially adds a program to an intentional agent, and then subtracts the agent (and thus all of the intentionality) to show no intentionality is left, what other metaphysical assumptions can Searle be making about the natures of minds and programs?

Searle says that there are three particular "points" at which the identity statement "mind is to brain as program is to hardware" is false ("breaks down") (Searle, p.293). The first point is the discontinuity between the "program and [its] realization" that allows one realization of a program to possess a property of intentionality, say a Chinese-Literate person performing a Chinese Room program, and another realization of that program to lack intentionality entirely (Searle, p.293). Searle cites the example of "a roll of toilet paper and a pile of small stones" as an instantiation of a program that does not have intentionality (Searle, p.293). The second point is the incongruity of programs and what he calls "intentional states" (Searle, p.293). His point is that programs are "defined as a certain formal shape" while intentional states such as the "belief that it is raining" are "defined in terms of their content, not their form" (Searle, p.293). Put another way, programs are a structure while intentional states like beliefs are about things, and can be expressed in any number of structures (Searle, p.293). The third point is that minds are a product of brains while programs are not merely the products of a computer (Searle, p.293). As justification for the former claim Searle refers to an earlier claim that:

> "It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality (I am, I suppose, the instantiation of any number of programs), but as far as we know it is because I am a certain sort of organism with a certain biological (i.e., chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena" (Searle, p.292).

Upon examination Searle appears to be saying that while biological structures may cause biological entities to have intentional states, and thus intentional properties such as understanding, formal structures do not because the latter structures are too loosely connected to the material bearing causal properties. Two things with the same biological structure, then, would be much more physically similar to each other than two things with merely the same formal structure. Searle illustrates this looseness of fit by comparing computer simulations of minds and mental events to computer simulations of fires and rainstorms (Searle, p.294). Since, in the latter, there is no risks of burns or flooding such as there are in actual fires and rainstorms, Searle suggests it seems odd that simulations of things such as minds would be able to duplicate notable features of minds such as understanding.

This relates to something that Searle refers to as "level of descriptions" (Searle, pp.293-294). Levels of description, apparently, are the levels according to which representations more or less resemble what they represent. Higher levels of description are more abstract and more general, lower are more specific and concrete. Descriptions of biological structures, for example, are taken by Searle to be less abstract than descriptions of formal structures, and to be descriptions of the material of intentionality. This is because at the biological level of description material is what is described; grey matter, white matter, and chemistry, it is all material. Searle argues that considering the concept of 'levels of description' helps to dispel "ambiguity", "residual behaviorism or operationalism", and "dualism" associated with the Strong AI thesis (Searle, pp.294-295).

The ambiguity is about the "notion of 'information'" (Searle, p.294). Using the concept of levels of description, Searle argues that the information processed by a digital computer is not the same sort of information processed by a person when they compute things (Searle, p.294). The information processed by a digital computer, he says, can either be understood merely as something homonymous with information processing such that it is simply another term for the syntactic manipulation of symbols, or it can be understood to be information processing at a remove whereby many things without intentionality may be considered to process information (Searle, p.294). The argument is that while human beings understand the information that we process, information processed by digital computers is not understood until it is interpreted, so it either is not information processing or only partial information processing. Searle says: "And no similarity is established between the computer and the brain in terms of any similarity of information processing" (Searle, pp.294-295).

Similarly, Searle argues that the residual behaviourism or operationalism is there because "appropriately programmed computers can have input-output patterns similar to those of human beings," at the level of description of messages exchanged with a Chinese Room (Searle, p.295). When the difference in level of description is demonstrated via the Chinese Room Argument, then the intuitions behind the Strong AI thesis are dispelled. Searle links the residual behaviourism, with its 'information' abstracted from the material, to a similarly "residual form of dualism" (Searle, p.295). The independence of programs from their instantiations removes them from the physical world and the phenomena they are intended to explain. Searle notes that: "Unless you believe that the mind is separable from the brain both conceptually and empirically – dualism in a strong form – you cannot hope to reproduce the mental by writing and running programs since programs must be independent of brains or any other particular forms of instantiation" (Searle, p.295).

> "Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena" (Searle, p.295).

I suggest that these levels of description are necessary for the Chinese Room Argument to be fully expressed. Recall that Searle has noted that the instantiation of a digital computer by the person-component of a Chinese Room fails to add the intentional property of understanding to that person. Hence if the program adds no understanding to a person, then the Chinese Room has no intentional property like understanding. The level of description schema is necessary to differentiate between John Searle's property of understanding and any program or computation that he might be instantiating, in order to demonstrate that the discontinuity between hardware and software does not occur between brains and minds. By enabling John Searle to instantiate the Chinese Room computer program the property of multiple realizability also prevents the program from adding any intentional properties to him. If levels of abstraction are not involved, then the Chinese Room Argument is only expressed partially such that failing to have the intentional property is a result of incongruity between that particular program and the property it putatively adds, and its conclusion does not generalize to all programs. It could be the case that the program that produces intentional properties is not the program that produces intelligent conversational Chinese properties, for example, and such a version of the Chinese Room Argument could only conclude that John Searle's Chinese Room lacks understanding of Chinese. Instead, Searle argues that computer programs are fundamentally disconnected from the sort of material that has intentional properties, which makes the Chinese Room Argument an argument against the principles of the Strong AI thesis.

1.4 Chapter Conclusion

So where does this leave the Chinese Room Argument? Before answering that question I would like to restate the entire Chinese Room Argument, incorporating the concepts of the last two sections:

Understanding is either a property of computer programs or it is not. Suppose that there is a computer program that answers questions about the implicit details of stories as though it understands written Chinese and it can be instantiated by John Searle, several sets of symbols, a book of rules, and a locked room, or just by John Searle. In addition, John Searle is a person who understands written English, but does not understand written Chinese. John Searle alone, or any other arrangement, can instantiate the program, but John Searle gains no understanding of written Chinese, despite the appearance of being able to intelligently answer the questions set to him. Because Searle is not instantiating a program that answers questions about English when he appears to understand written English, and instantiating the Chinese Room program when he appears to understand written Chinese, understanding cannot be attributed to instantiating that program. Therefore understanding is not a property of computer programs.

If understanding is not a property of programs, does that exclude properties from being shared by programs? It seems that this argument can be generalized to any property of a program because the argument turns on the concept of instantiation, which is a term used to denote when something at a higher level of description is imposed over something at a lower level of description. The formal manipulation of uninterpreted symbols is, by definition, at a more abstract level than whatever it simulates because it leaves out the semantics of whatever it simulates. It is interpreted to simulate via some stipulated congruence between formal properties. Given that this property of the formal manipulation of uninterpreted symbols operates on a more abstract level of description from the level of description wherein mental properties such as intentional properties are produced, then there is a problem for a theory like the Strong AI thesis.

The problem is that the principles underlying such a thesis, that digital computers may have mental properties such as understanding, are principles that the Chinese Room Argument replaces with its own. Proponents of the theory may show that at some level of description John Searle is instantiating a program for understanding written English, but they cannot use the principles of the theory within Searle's metaphysics to show that instantiating a program for understanding written English will allow the instantiation to understand written English. The flexibility that allows John Searle, or merely a roomful of objects including John Searle, to be the program's computer disqualifies it from having content-properties like understanding. Pure syntactic representation is not enough to be anything by Searle's metaphysics of computation. That is why it must be instantiated, so that something represents it. If a program cannot have such important causal properties, then no computer program will be able to explain mental states and the nature of mind. This deficit makes the Chinese Room Argument an intractable problem for the Strong AI thesis.

However, recall that Searle does not object to the Weak AI project, which he agrees may help explain mental states. He does not expect the study of computer programs to yield an adequate explanation for the nature of mind, since they cannot be minds, but one might imagine that the scientific process of building and testing models would be one avenue to an adequate scientific explanation of the nature of mind. A picture may be informative about its subject, if not an exemplar of that subject. However, if Searle objects to the Strong AI thesis in principle, then he objects to the Weak AI project in principle because the Weak AI project is based on the principle that there is something importantly mind-like about programs, and, specifically, something structural; the structures of programs may provide insight into the structures of mind. Moreover one imagines that *Mind* in general is merely an abstracted description of the aggregate phenomena of minds, and that the study of the mind at such a level of description seems to be a similar (if not the same) level as that modeled using computer programs. Perhaps no computer program has understanding of its inputs and outputs while particular instantiations do; one could also say that Mind in general has no understanding of its queries and responses while particular people with minds do. In this chapter I have shown that Searle's objections to the Strong AI thesis are valid insofar as his metaphysics of computation and mind hold.

I agree that Searle's conclusion follows from its premises, and the gaps in the original sketch of a man enacting computation are filled by the background assumptions and arguments that Searle makes about the concepts involved, information, levels of description, and the formal nature of syntactic models. If we are to object and say that Searle simply assumes John Searle has intentionality and the Chinese Room does not, then Searle can note that intentionality appears to be biological. If we are to object and say that a machine could very well have a mind because of its embodiment we give up the thesis that minds are programs. If we are to object and say that John Searle may be running a program called 'understanding' in addition to the program of 'answering-questions-about-implicit-details-in-stories-in-Chinese', then it seems that we are confused about the difference between the biological level of description and its formal level of description as the manipulation of uninterpreted symbols. If the Chinese Room Argument is to be shown to be invalid, it will not be within the framework of Searle's metaphysics. Hence Searle's metaphysics will have to be shown to be inadequate for the task of modelling digital computation. I will return to the arguments of this chapter in Chapter Four in order to show that, despite appearances, the Chinese Room Argument is not valid and, thus, not an intractable problem for a theory like the Strong AI thesis.

2.0 Chapter Introduction

In this chapter of this thesis I will introduce the Universal Turing Machine and explain the formal terms and methods employed by Alan Turing in its construction. In doing so, I will argue that the test proposed by Turing in "Computing Machinery and Intelligence" is, as opposed to what John Searle argued in "Mind, Brains and Programs", a test for mental properties such as intentionality. I will explain how the metaphysics of digital computation employed by Turing differ in important respects from the metaphysics behind the Chinese Room Argument. In particular, Turing's metaphysics includes different assumptions about which inferences may be made at certain levels of descriptions, and about what conclusions may be deduced from the presence of purely syntactic properties.

In addition to analysis of Turing's article, I will introduce commentary on that article and its concepts by Daniel Dennett and Richard White. White's article "Some Basic Concepts of Computability Theory" will be used to clarify some of the concepts expressed in Turing's article, particularly the concept of the Turing Machine as a model or schema. The nature of the Universal Turing Machine as a schema or model will help to situate the digital computer described in the Chinese Room Argument as something less universal and less observer-relative than the model of computation supposed by Searle.

Dennett's article, "Can Machines Think?" will be used to demonstrate certain features of the Turing Test that are glossed over in the assumptions, made by Searle, mentioned in the previous chapter. In particular, Dennett's article will be used to discuss the implications what he calls the property of "a wider competence" possessed by any machine capable of passing the Turing Test (Dennett, p.7). I will show both that Turing addresses the nature of this property in his article, and that this property is necessary in order to satisfy Turing's claim that a machine capable of computing a solution to the Imitation game is possible. In short, I will argue that Turing's description of the Universal Turing Machine is relevant to the Chinese Room Argument because it is importantly different from the model of computation that Searle purports to address in his Chinese Room Argument.

Briefly, the Universal Turing Machine is the formal schema of computation that the Chinese Room is supposed to instantiate, and a schema capable of duplicating any special purpose Turing Machine or computing any Turing-computable function. Where some function is Turing-computable there is a special purpose Turing Machine capable of computing it; hence, if the functions performed by a Chinese Room are functions that are Turing-computable, then they may be enacted by a special-purpose Turing Machine and, hence, by a Universal Turing Machine. A Universal Turing Machine can do more than provide a basic architecture for a process by which questions about a story may be answered. Even more importantly, the Universal Turing Machine does so at a lower level of description than any particular program. The Universal Turing Machine provides insight into how notions of instantiation and levels of description work in the schema of computation in which proponents of the Strong AI thesis cash out the validity of their arguments.

Searle admits, for the sake of argument, that the Chinese Room may successfully compute a solution to the Imitation Game. However, he argues that it does so using purely syntactic methods that leave the symbols composing any such solution un-interpreted. Searle argues that this is the case, that the methods employed are purely syntactic, because he understands the Turing Machine schema to be broadly universal such that any collection of objects may count as a digital computer. White notes that this broad universality is not a feature of the Turing schema, and that its universality is limited even within the realm of what counts as computation. Even a Universal Turing Machine is limited in what it can compute: such a machine can compute only functions that are Turing-computable, which is merely a proper subset of all computable functions. It follows that only a certain collections of objects may be digital computers in the sense of embodying Universal Turing Machines.

Despite our powers of imagination, some collections of objects will never have the properties that a Universal Turing Machine must have, and thus computation such as Turing-computation (as opposed to the theory that describes Turing-computation) cannot be observer-relative in the broadly universal way that Searle argues that it must be. Thus the collection of artefacts that can instantiate a Chinese Room may quite reasonably include systems with obviously no mental states, and exclude any such system on the basis that they obviously have no mental states is presumptive. Hence software is not so thoroughly disconnected from hardware as Searle would suppose it to be, so excluding software from having causally efficacious properties misconstrues the relation of software to hardware that the Chinese Room Argument is predicated upon.

Aside from misconstruing the metaphysical generality of the Universal Turing Machine, Searle glosses over another important feature of the Universal Turing Machine: the specific process. By reducing the relation of hardware to software to a simple two-level description or abstraction, Searle oversimplifies the metaphysical framework that allows programs to be embedded in one another. The specific process of how something instantiating a Universal Turing Machine may also instantiate the process of answering questions about a story like an intelligent person is important because it is not a single universal process. While a Universal Turing Machine may compute a solution to any problem solvable via specialist Turing Machines, how it does so will be significantly different from the specialist Turing Machine thanks to different components, and may be more or less different depending on what other differences in architecture lie between the embedded programs. In particular this will be the case if the specialist Turing Machine in question has features that the Universal Machine does not, such as more than one read-write head. Dennett's article addresses this property of computational practicality, and particularly the complexity of programs for computing a solution to the Imitation Game. By addressing this feature Dennett provides a space for the property of intentionality to fit into the Turing Machine schema, as a category of the operations that a Turing Machine must perform in order to successfully compute a solution to complex problems such the Imitation Game.

In each of the following three sections of this chapter I will discuss the concepts of computation, instantiation, and abstraction that Searle used in the Chinese Room Argument as they relate to the theories of the Turing Machine and the Turing Test, and analyze these concepts in the context of the articles by Turing, White, and Dennett. The first section of this chapter will cover comments from Turing's article "Computing Machinery and Intelligence" White's "Some Basic Concepts of Computability Theory" on the Turing Machine. The second section of this chapter will cover Turing's argument concerning the significance of the Imitation Game, and expand on the metaphysics of the Turing Machine as it relates to properties such as intentionality. The third section will cover Dennett's "Can Machines Think?" and discuss the Imitation Game and the Turing Machine as they pertain to arguments like the Chinese Room Argument. This last section will expand upon the importance of how a program might solve the Imitation Game or Turing Test. The fourth section will review the implications for the Chinese Room

Argument that Turing's metaphysics raises. As with the previous chapter, I will end this chapter with some tentative conclusions based on the arguments put forward in the articles. These conclusions and critiques will then be left until the fourth chapter where they will be united with the conclusions and critiques of Chapters One and Three.

2.1 The Turing Machine

So what is a Universal Turing Machine and what are the metaphysics of it? White's article "Some Basic Concepts of Computability Theory", explains that: "A Turing Machine may be regarded as the control of a finite deterministic automaton coupled to an unbounded external memory in the form of a single tape, divided into squares like the input tapes of sequential machines" (White, p.207). A Turing Machine can thus be considered to have at least three parts: a control, a read-write head, and a tape. The tape is divided into cells upon which the read-write head may read a symbol or write a symbol. The read-write head detects the symbol in a cell and depending on its control may (non-exclusively) either write a new symbol or move along the tape to another cell. The tape represents both the input and output to the system, a storage medium, while the read-write head and the control represent whatever machinery might turn that input into output.

Compare such a machine with a finite deterministic automaton. A finite deterministic automaton is a device with two tapes, an input and output tape respectively, an input reader, an output printer, and a control that decides what is printed and what is read next given any particular input (White, p.206). Formally, such an automaton consists of a finite input set, a finite output set, a finite set of system states, a function that turns one input into the next input, and a function that turns each input into output. So, while any computation performed by a finite automaton will eventually be computed, it is possible for a Turing Machine to perform indefinitely, even though both machines have "finite number of internal states" (White, p.208).

Although any such machine is composed of a discrete and finite number of conceivable components it is possible that it might never come to a halt in its operation (White, p.208). This finitude is, in part, a result of its definition as a finite set of discrete components. Unlike an analog machine, its states are not infinitely divisible, and so can be represented on a machine table of finite size (White, p.209). This finitude is also in part a result of its conceptual power so that the theoretical performance of machines with different numbers of components, such as an infinite number of components, will change but the ability to compute functions will not. White says to that effect: "It can be shown that any function computable by a polycephalic machine is computable by an ordinary Turing machine" (White, p.209).

Such a "universal" machine can perform any Turing-computable function (White, p.211). Recall that although the Turing Machine is called a 'machine', it is not a machine in the sense of being a physical object. The Turing Machine is a type or class of machine, a formal structure any instantiation of which can be used to compute solutions to algorithmic problems, and the Universal Turing Machine is likewise the class of Turing Machines that can be used to compute solutions to any algorithmic problems. The members of these classes are the devices we commonly call 'computers'. Special purposes devices may be classed as Turing Machines, while all-purpose devices such as a desktop computer may be classified as Universal Turing Machines (White, p.211). The Turing Machine is thus a model or "abstract" machine that can be used to understand and predict the behaviour of mechanical devices, as well as providing a theoretical basis for the science of computation itself (White, p.212). Yet I should note that Turing Machines are not the limit for computational models. White tells us that with regard to "the practical importance of Turing's universal program":

> "...automata theory abounds with mathematical models of computers which are less powerful but more realistic than Turing machines, as well as powerful machines that (by Church's thesis) cannot compute any function not already computable by a Turing machine, but which may be much faster and more efficient than Turing machines" (White, p.212).

The implication of this comment, for my purpose in this chapter, is that such a machine can model any kind of device that could be used for computation, where we brush aside issues of performance as irrelevant at that level of abstraction. More ambitiously, there is the implication with which Searle agrees, that if the human central nervous system does computation in order to do what it does for a person, then it can be adequately represented as a Turing Machine or some similarly formal schema, and duplicated using physical parts that are not materially identical to nerve tissue. Since, Searle argues, the human central nervous system does not do computation in order to do what it does for a person, and can merely be represented as such, mere formal identity is not sufficient. So long as the formalism of the Turing Machine maps onto a system at some level of abstraction, then that system can be considered to be a sort of Turing Machine. So long as the formalism of the Turing Machine maps onto a device, then that device can be considered to be a token Turing Machine or digital computer. This property of identity is important because it is what allows a Chinese Room to be a model of digital computation. For the sake of argument, a Chinese Room can be instantiated by something describable as a

Turing Machine insofar as those Machines are digital computers. Turing says that: "The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer" (Turing, p.267).

Turing does not mean the operations of putting pen to paper, or adding columns of figures. Likewise he dismisses the commonality of the presence of electricity, electrical current, in both human brains and contemporary semi-conductor based computing technologies. He is referring to the concept of computation described by Searle, manipulating a set of symbols according to a set of rules, something in common to all instantiations of computation. Turing cites Charles Babbage's "Analytical Engine", a wholly mechanical (and hence inefficiently slow) computing device, as an example of what White calls "the multiple realizability of programs" whereby the Turing Machine is not physical (Turing, pp.268-269, White, P.215). Recall from Chapter One that it is this argument about the non-physical nature of computation that Searle used in "Mind, Brains, and Programs". It goes: If a Turing Machine is non-physical, and intentionality is a property of the physical world, then being a Turing Machine will not cause its instantiation to have that property.

However, Turing is not describing a para-mechanical machine that may or may not be attached to physical objects by a process called 'instantiation'. Turing is describing real-world devices with a certain sort of structure. He says: "The reader must accept it as a fact that digital computers can be constructed, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely" (Turing, p.268). The Turing Machine itself is a member of the class of what Turing calls the "discrete state machines", but the name

"Turing Machine" is not a token-term: it is a type-term (Turing, p.269). The tokens of the type 'Turing Machine' are the digital computers the existence of which Turing states we must accept. If "Turing Machine" is a type-term, then pending the existence of Platonic universals whose essence is cast into everyday objects by a conjuration known as 'instantiation', it seems quite reasonable that being a Turing Machine should be both physical, and non-physical, depending on whether we are referring to a particular device or the entire class of machines. The class of digital computers has physical referents just as 'the computer upon which this thesis was written' has a physical referent. The term 'instantiation', then is simply a reference to participation in a particular class; a device is a Turing Machine when it serves as a token of that type of machine.

So when Turing suggests that: "We may hope that machines will eventually compete with men in all purely intellectual fields" he is not suggesting that these digital computers will have mental properties because they have an addition to their physical mechanism called a 'program' or that they will merely be describable as performing the operations of those programs, but because what they do will be identical to thinking (Turing, p.282). I say that what these machines will do can be classified as thinking, but that does a disservice to Turing's original concept of the Turing Machine. Turing does not suppose that digital computers will be replicating human thought because, as he suggests when considering the possibility of replicating human thought by replicating a human being: "To do so would be a feat of biological technique deserving of the very highest praise, but we would not be inclined to regard it as a case of 'constructing a thinking machine'" (Turing, p.266). Instead, Turing seems to be suggesting that these devices will share a structure in both space and time with structures classified as 'thinking'. Note that because Turing has described a very different picture of computation to Searle's simple bifurcation of computation into machine-material and computer program or symbol manipulation, he disagrees with Searle's assumption that brains are thinking machines. This disagreement is because the structures of machines are, according to Turing's metaphysics of computation, a part of the physical material. Whereas Searle's model of computation is something imposed over objects, and observer-relative, Turing's model of computation is a structure shared by objects; a structure that observers can identify.

2.2 The Turing Test

This discussion of what counts as thought brings me to Turing's original formulation of what he called the "Imitation Game" and which is more colloquially called the 'Turing Test' (Turing, p.265). In his article "Computing Machinery and Intelligence" Turing explains that formulating an answer to the question "Can machines think?" will turn on the definitions of "the terms 'machine' and 'think'" (Turing, p.265). He says that on consideration these terms seem to be too ambiguously defined for any definitive answer to be forthcoming (Turing, p.265). Instead he proposes an alternative formalized as a game that will let us avoid question-begging definitional distractions. Whereas answering the original question depends on being able to derive some logical connection between the concepts associated with 'machine' and 'think', answering this new question depends on whether some candidate for thought measured up to some putative thinking thing. Essentially, Turing hypothesizes a formal schema for evaluating phenomena that might be colloquially called 'thought'.

Thus Turing introduces the "Imitation Game", a game wherein an interrogator must determine which of two players is a man and which is a woman according to some gender-neutral process such as judging the content of written notes (Turing, p.265). When applied to the question of judging whether a computer can think, the Imitation Game involves on an interrogator attempting to determine which respondent is a computer and which is a person according to some body-neutral process such as judging the content of written notes (Turing, pp.265-266).

Turing notes that having a neutral process by which a contestant might be judged has the advantage of not prejudicing the judge in favour of physical attributes. This is useful because the test is intended to determine if the inside of one contestant is more or less the same as the inside of another contestant, and if not the same then at least not distinguishable. Turing recognizes that this criterion of distinguishability may seem question-begging to people worried about the apparent flexibility of fit between the inner thoughts of some contestant and their outward behaviour. To that end, he asks: "May not machines carry out something which ought to be described as thinking but which is very different from what a man does?" (Turing, p.266). Perhaps more obliquely than necessary, Turing answers this question by noting that: "This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection" (Turing, p.266).

Turing's answer to this question may seem enigmatic. However he follows up this critique of the imitation game with a discussion of what sort of machines the game will be concerned with distinguishing from ordinary conscious people. These sorts of machines, Turing says, will be "a particular kind of machine, usually called an 'electronic computer' or 'digital computer'" (Turing, p.266). As noted above, requiring that the machine both play the imitation game satisfactorily and in the same way as a human being is beside the point when you want to figure out if a non-human machine can think. Turing points out that the idea is to explore the congruence between digital machines and human beings.

Recall that for any Turing-computable function an appropriately arranged Turing Machine exists that can compute it, and that a Universal Turing Machine exists that can also compute it. So two Turing Machines may operate very differently but can still compute the same functions. It follows that if a machine can be constructed to perform the functions required to satisfy the Imitation Game, then how it computes those functions is essentially cosmetic to the fact that it does, and in doing so satisfies the Imitation Game. We have those strong reasons because of the nature of the Universal Turing Machine and thus its tokens. He says: "Provided it could be carried out sufficiently quickly the digital computer could mimic the behaviour of any discrete state machine" (Turing, p.270). Where it could mimic the behaviour of a human being, and a human being could be classified as a discrete state machine, then there is an important equivalence. In satisfying the Imitation Game then, the machine gives us strong reasons for believing that it has done something significantly like thinking. This equivalence, however, is limited by precisely what must be imitated.

Turing says: "The short answer is that we are not asking whether all digital computers would do well in the game nor whether the computers at present available would do well, but whether there are imaginable computers which would do well" (Turing, p.267). Turing's original formulation of the Turing Test should therefore not be taken to make simplistic equivalences among behaviour, mind, and program. Some programs, such as those created during Turing's day, or just very simple computers in general, would simply not count as thinking just because a human judge could not distinguish between, say, a digital computer completing sums and a human computer completing the same sums. The implication of this argument is that if it shows that a computer can think, it does not suggest that the computer does so just because it has a program, but because it has a program that is equivalent to the program that a human being can be considered to run to accomplish the same activity – they share the same type of program, a type of program for thinking.

Of the objections to this proposal that Turing considers, what he calls the "Argument from Continuity in the Nervous System" is perhaps the most difficult one (Turing, p.276). After all, if a discrete state machine cannot adequately represent the human nervous system, then the argument that there is some significant equivalence between a Universal Turing Machine and a human being doing the same intellectual work must work from premises of mere correlation. This counter-argument should be familiar from Searle's discussion of the Chinese Room Argument; that while a Chinese Room's behaviour may not be distinguishable from that of a Chinese Literate person, that inability to distinguish the two does not serve as good reason to infer that the two Imitation Game players share significant properties. Indeed, Searle argues that this indistinguishability misleads us when, in principle, a Chinese Room cannot be a Chinese Literate person.

Interestingly, Turing's reply suggests that this possibility of mistaken identity is irrelevant, but not for the reasons one might assume from his discussion of difference and sameness in physical particulars. Recall that his point is that the artificial intelligence should not be the same as the natural intelligence if we are to take the hypothesis seriously. He raises the example of a machine called a "differential analyser" which a digital computer of the Universal Turing Machine type cannot mimic – presumably because the differential analyser computes functions that are not Turing-computable (Turing, p.277). He notes that although a Turing Machine might not be able to provide the same answers, they would be good enough to satisfy the judge of the Imitation Game: "Under these circumstances it would be very difficult for the interrogator to distinguish the differential analyzer from the digital computer" (Turing, p.277). After all, if the contestant gets the wrong answer, that can be attributed to human error. The correct answer can be attributed to access to a differential analyser.

However, indirectly, this raises a counter-part to the problem of being overly liberal in attributing minds to simple programs that are not doing anything mind-like. Dependence upon the skill of a human judge to tell the difference between a human computer and a digital computer like a Chinese Room may result in an unreasonably high rate of false positives. One could imagine a judge simply being incapable of distinguishing a Chinese Room from a natural intelligence where any difference, supposing there is one, is very subtle. Dennett answers this objection with the benefit of hindsight that Turing lacks, in his article "Can Machines Think?"

2.3 Dennett on the Turing Test

In answering the question whether a person could be taken in by a Chinese Room, and thus the question whether a test judged by a fallible human judge could ever be conclusively indicative of intelligence (if not other mental properties), Dennett says: "A failure to thinking imaginatively about the test actually proposed by Turing has led many to underestimate its severity and to confuse it with much less interesting proposals" (Dennett, p.5). Dennett explains that the Imitation Game expresses several insights. The first is that human judges are fallible and biased, and that they do fail in practice when given the opportunity. Dennett raises the practice of orchestra auditions being conducted with the aspirant behind a screen so as not to distract the judges with ephemera such as physical appearance, as an example of attempts to avoid bias (Dennett, p.5). Similarly, the Turing Test screens the judge from the participants to ward against bias; all the judge or interrogator has to go on is the responses returned in a suitably neutral fashion. This screening, Dennett argues, allows the test to avoid what obscures matters, which are "the capacity to understand, and think cleverly about, challenging problems" (Dennett, p.5).

Dennett cites an argument by Rene Descartes as possible inspiration for this identification of verbal behaviour with intelligence, to the effect that "the capacity to hold an intelligent conversation" is the most "demanding test of human[like] mentality" (Dennett, p.6). To Turing he attributes the following idea: "Nothing could possibly pass the Turing test by winning the imitation game without being able to perform indefinitely many other clearly intelligent actions" (Dennett, p.6). Dennett calls this the "quick-probe assumption" (Dennett, p.6). There are English sentences commonly used in casual conversation, Dennett argues, deciphering the sense of which requires massive amounts of contextual information (Dennett, p.7). He says: "The only way, it appears, for a computer to disambiguate [such] sentence[s] and keep up its end of the conversation that uses [those] sentence[s] would be for it to have a much more general ability to respond intelligently to information about social and political circumstances, and many other topics" (Dennett, p.7).

Furthermore, Dennett addresses the rebuttal that despite overcoming the computational challenge of managing such a large and complex database, one "may protest, something might pass the Turing test and still not be intelligent, not be a thinker" (Dennett, p.9). The possibility of such "false positive" results being a problem is dismissed by Dennett as being a similar problem for any such tests we might devise to map our physical world, and not a reasonable objection (Dennett, p.9). Dennett notes that the standard of reasonableness that underwrites such tests for properties is often obscured by accusations of "operationalism", that the Turing test simply defines intelligence such that passing the Turing test is defined as a sign of intelligence, and as such begs the question (Dennett, p.9). The amount of evidence required to pass the full Turing test, and not some restricted version, however, is counter to the criticism that the Turing Test is merely definitional. Passing the Turing test is not the definition of intelligence, or thinking: it is duplicating the behaviour of a natural intelligence. The test is evidentiary, not definitional. Dennett concludes his rejection of such objections by saying: "The moral we should draw is that as Turing test judges we should resist all limitations and waterings-down of the Turing test. They make the game too easy – vastly easier than the original test" (Dennett, p.11). To whit, given the importance of empirical results it is important not to assign credence to the results of the wrong sort of test. If the Chinese Room Argument waters down the Turing Test, then its preliminary conclusion is misleadingly generalized to its principled rejection of the mental properties of computer programs.

To give some idea of the computational obstacles to merely creating a well-indexed finite compilation of possible intelligent conversations Dennett offers the following conjecture: "So let's say, to be very conservative, that there are only ten to the fiftieth different smart conversations such a computer would have to store. Well, the task shouldn't take more than a few trillion years – given generous government support" (Dennett, p.12). Taking such computational tasks to be the sort that natural intelligences can solve on an everyday basis, Dennett argues that avoiding just this sort of computationally monumental task is one hallmark of intelligence that the Turing test could help to identify in programs that do not conveniently conform to "species-chauvinistic, or anthropocentric" prejudices about the nature of intelligence (Dennett, p.13). Extending Turing's original specification, such a test would not even require intelligence to be limited to trivialities such as the ability to understand sub-textual cues in stories and explain them in any human language.

This anthropocentrism works in other ways. Dennett notes that "The normal habit of human thought when conversing with one another is to assume general comprehension, to assume rationality, to assume, moreover, that the quick-probe assumption is, in general, sound" (Dennett, p19). In fact, he notes, it is not uncommon for people to mistake unintelligent computer programs, amusingly called "expert systems", for intelligent persons (Dennett, p.16). He gives the example of a program called "PARRY" that simulates "a paranoid patient who has delusions about the Mafia being out to get him" (Dennett, p.13). When psychiatrists were called upon to perform a Turing test, Dennett relates, they failed to do better than chance in distinguishing the program from real paranoid patients. Dennett says that any questions of the sort that did test the intellectual limits of such programs were prevented from being asked by a combination of scrupulous medical ethics whereby a paranoid patient is not challenged to be a digital computer imitating a paranoid patient, and the idea of only testing the program to see how well it acted the part of a paranoid patient (Dennett, p.14).

Turing addresses the property of wider competences that an intelligent digital computer would require when he discusses "Learning Machines" (Turing, p.278). When he addresses objections to his proposal, one such is attributed to Lady Ada Lovelace. Turing quotes her as saying: "The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform", which he parses as "the machine can only do what we tell it to do" (Turing, p.275, 278). That sets the limits of the machine well inside the limited competences of an expert system, something that does not run up against what Dennett describes as the "combinatorial explosion" of ten to the fiftieth intelligent conversations that would need to be stored in the database of such a dumb computer. Rather than making the combinatorial argument put forth by Dennett, however, and perhaps because he lacks Dennett's benefit of hindsight, Turing offers a couple of analogies.

The first analogy is that of critical mass: the idea that there is a degree of complexity at which Lovelace's dictum does not hold, and even "working out the consequences from data and general principles" (as Turing advises when one is prone to presuming that "as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it") will not be enough to provide a complete machine table before it changes. Turing says, by contrast, there are sub-critical minds such that: "An idea presented to such a mind will on an average give rise to less than one idea in reply." This describes expert systems well; an idea presented to an expert system will give rise to less than one idea in reply; it will give rise to no idea in reply, just a parroted selection from an appropriate index. Considering a machine that constantly adds to its own code, as well as constantly rewriting it may be such that: "An idea present to such a mind may give rise to a whole 'theory' consisting of secondary, tertiary and more remote ideas" (Turing, p.278).

The second analogy is that of an onion. In this analogy the functions of the mind are mapped onto computable functions until either the mind runs out of functions, or we run out of computable functions. In the latter case, Turing suggests, this remnant might be the "real mind" (Turing, p.278). More interesting, though, is the idea of encapsulation that Turing alludes to with the analogy of nested onion skin. As mentioned earlier, it is possible for programs to be constructed so that they provide a virtual device for running programs at a higher level of abstraction. Following this analogy, it may be the case that where higher-level competences are stripped away, one might not need to strip away everything before one finds architecture that provides no competencies in itself, but this is an essential foundation for higher-level functions that do map onto competencies. Dennett says: "That no nonhuman Turing test winners are yet visible on the horizon does not mean that there aren't machines that already exhibit some of the important features of thought" (Dennett, p.20).

Perhaps more importantly, the analogies of critical mass and onion skins raise a question of precisely what sort of metaphysical framework of syntax and semantics Turing presumes when, contra Searle, he argues that something very much like a semantic property might arise from something with suitably sophisticated syntactic properties. Both analogies suggest a concept of semantics or 'about-ness' that is not dichotomous with syntax or the formal manipulation of symbols. In the critical-mass analogy one might take an idea to have content, or to have a semantic value. The quantity (and perhaps formal arrangement, interaction, or rate of introduction) of such content would be a syntactic property of the system, making semantics and syntax at least entangled. Or perhaps an idea is simply an uninterpreted symbol until the system has a large enough collection of symbols to construct a sufficiently complex grammar and rules of inference that more symbols are produced than get introduced into the system. In the onion analogy one might take a competency or function to be understanding, which rests on a lower level of less abstract syntax, which rests on a lower level of less abstract physical structure of a machine, which rests on a set of familiar everyday devices, and so on.

Turing's argument with regard to the Imitation Game is then that a computational model such as the Turing Machine takes hand-waving about abstract symbol manipulation, and posits a framework within which hypotheses about the identity of computational structures can be tested. It provides a model of how structures at higher levels of description can be both instantiated by particular devices and identified as being so instantiated. If these structures have notable features, then a heuristic probe that tests for their presence can reliably identify those structures. Dennett's argument is that if formal properties are also natural properties, it will be the case that things with similar formal properties will share similar natural properties. Where intentionality is a natural property, and Turing's metaphysics of computational structures allows natural properties to be formally identified, then the Turing Test is a test for just such a property, contra the Chinese Room Argument.

2.4 The Implications for the Chinese Room Argument

Recall the earlier discussion of the discrepancy between any Universal Turing Machine's capacity to duplicate the computational capacity of more specialized Turing Machines, and the possible deficit in performance incurred by using a Universal Turing Machine to do a specialized Turing Machine's job. In part, this discrepancy may be due to the Universal Machine duplicating the performance of the specialized Machine with additional operations rather than additional components, and in part this may be due to the inefficient mechanism employing the principles of the Universal Turing Machine. While a micro-chip actuated Universal Turing Machine may have significant gains in performance over a mechanically actuated specialized Turing Machine that the former can simulate, this difference in performance can be attributed to differences at a lower level of abstraction rather than any essential difference at the level of abstraction at which these machines are equivalent. If a Chinese Room faced with such practical concerns cannot face a full Turing Test, then Searle is right and it cannot claim to be what it simulates.

Given the computationally daunting task that a Chinese Room faces with a full Turing Test, rather than an abridged version such as answering questions about implicit information in some predetermined stories, it is doubtful that John Searle could realistically hope to simulate a Chinese Literate person without supposing John Searle to have a superhuman capacity for mental computation even with the aid of pen and paper. Given that John Searle's brain is precisely the sort of powerful machine capable of handling this sort of computational task unconsciously, it seems more reasonable to consider his proposal to instantiate the entire Chinese Room without consciously following a set of rules. Whether John Searle does this via memorization of rote rules of

Chinese symbol manipulation, or by using his own brain as computing machinery, the important thing is that he is able to converse in written Chinese in the intelligent fashion described. This test would require John Searle to be able to converse in written Chinese about anything to which an intelligent Chinese Literate person might be imagined to offer an intelligent response.

What sort of Chinese Room would there be in such a circumstance? On one hand, Turing, and perhaps Dennett, would dismiss the effort as being of the same sort of thing as producing a computer by producing another human being: a fascinating project with little relevance as to whether a non-human machine might be able to think. On the other hand, Turing might agree that what John Searle is doing when he computes the proper manipulation of Chinese symbols is as much thinking as when John Searle is computing the proper manipulation of Arabic symbols to perform feats of arithmetic. Turing might argue that what John Searle is doing when he computes the proper manipulation of Chinese symbols is what he does with English symbols, particularly if John Searle manipulates Chinese symbols as intelligently as he does English symbols.

Recall from Chapter One that Searle would say that such manipulation of un-interpreted symbols would be purely syntactic, while the manipulation of interpreted symbols such as English symbols would have a semantic (physical) component. He assumes that one could not infer some semantic content from the mere manipulation of symbols because syntax is at a higher level of abstraction from semantics. Turing's Machine schema suggests that higher levels of description are not dichotomous with lower levels of description. As Turing points out, it is a fact that devices instantiating the Universal Turing Machine exist, and that they can duplicate the competencies of machines described at one level of abstraction with machines described at another level of abstraction. It seems, then, that Searle cannot assume that the formal nature of a program excludes it from having semantic properties that are either subvenient to, supervenient on, or simply entangled with its syntactic properties.

Another assumption that Searle makes about programs is that they are pictures of computation that an observer can usefully see in devices, rather than a kind of process that the device engages in when its mechanism operates. This leads him to make the inference that the computer program does not actually do anything, and that it merely represents. By contrast, Turing is more careful to note the difference between the type of program, tokens of programs, and the sort of thing that is being referred to when one suggests that a program is causing a device to do something. Turing's concept of the computer program is one in which a computer program can be coherently said to cause a device to do something because, when all of that sort of computer's states are at some state x, and that device is that sort of computer, then those states are sufficient and necessary for those actions to obtain. The device has a program, the token arrangement of its components and their operations, which is also the type of arrangement of those types of components and that type of operation. This particular programmatic physical arrangement has no dependence on an observer, assuming for the sake of materialist argument an observer-independent world, though its classification according to some theory of computer science obviously does.

Thus Searle's assumption that there is nothing going on in a digital computer that some observer with the property of intentionality does not provide is both question-begging and boot-strapping for his own argument that there is nothing going on in a digital computer. In the case of John Searle being able to hold intelligent conversations in written Chinese without having a clue about the content of what his calligraphy might mean would simply indicate an interesting mental phenomenon, like that of blindsight. While Turing dismisses the difference between how a digital computer and a human being might pass the Turing test, it seems the details of difference are not wholly irrelevant. The differences between a digital computer and a human being are irrelevant precisely because the concept of the Turing test is about evaluating difference in performance between different sorts of possible computer. What any particular digital computer does and how it accomplishes that task, the nature of its program, as Dennett explains, is quite important. The details of the type of computer program that a digital computer instantiates make the difference between a practically impossible computational feat requiring vast resources, and something that might be able to pull off similar economies of scale that a human central nervous system does when active and well educated with the right sort of resources.

2.5 Chapter Conclusion

The issue of performance is interesting when one considers the construction of the Turing Machine, and thus the Chinese Room, and Searle's conclusion that such a machine would not be intelligent. Given Dennett's argument about the sheer complexity of building an unintelligent Chinese Room that could pass the unrestricted Turing test, and his subsequent conclusion that accomplishing the same task with more realistic resources would indict something to be "intelligent", one would suppose that a successful Chinese Room would imply the Turing test to be a very good test for mental properties (Dennett, p.13). Whether a machine has access to a colossal amount of world knowledge, or simply an extremely clever design for intelligent conversation given a comparative paucity of resources, a Turing Test would be a useful empirical test if the metaphysics of Turing hold.

Such a quick-probe could indicate a vast amount of information stored at some level of abstraction which some mechanism for selecting conversational phrases at a higher level of abstraction could be said to be about; a program of a pattern of stored bits to indicate some things in the context of one level of abstraction and some other things in the context of a different level of abstraction would satisfy a definition of semantics whereby symbols are related to referents via context, for example. Or such a quick-probe could indicate a small amount of information being handled in a suspiciously intelligent and creative way, such as by a contestant interested in finding out what the questioning is about. This latter machine would be like the program discussed by Searle, except, instead of supplying information to questions about the story that could be answered by access to a stock of background information, it would be a program that would be able to formulate questions about the story, rather than just about the sentences used to supply answers. It would be a program with a specific sort of structure, an intelligent user of its resources, if not an intelligent correspondent.

Regardless of how the computer goes about satisfying such a quick-probe, such a quick-probe would be an effective (if not conclusive) test of mental properties such as intentionality in a subject in the Turing test where we abandon Searle's assumption that the answers given by the subject have no bearing on how those answers are generated. The point of the Turing test, considering Dennett and Turing's points about the relevance of performance, is that there is a significant link between a subject's performance in the unrestricted Turing test and how it accomplishes that performance. The significance of a machine like a Chinese Room passing a heavily restricted Turing test using a program thoroughly unlike a mind says only that the equivalence 'mind is to body as all software is to hardware' is false. Indeed, Dennett relates a comment about the significance of such restricted tests by Joseph Weizenbaum that even an electric typewriter, at most only a specialized Turing Machine, could pass a Turing test for infant autism (Dennett, pp.13-14). The equivalence 'all mind is to body as some software is to hardware' remains unchallenged by the Chinese Room when Turing's original conceptions of digital computation and the Turing test are substituted for some of Searle's assumptions about the same.

However, where the premises of an argument are changed, one might reasonably expect the conclusion of the argument to change and even to contradict the previous argument. Such contradiction would not render the original argument to be logically invalid. Where the premises of the original argument are in fact false, at best the original argument would be unsound. Since, in this thesis, I propose to demonstrate that the reasoning of the Chinese Room Argument is invalid, I must do other than show that its premises are untrue for the purpose to which Searle puts them. However, as I have shown in this chapter, Searle has not merely assumed false premises about the nature of the Turing Machine and the Turing Test, he has made several false assumptions about the sort of inferences that he can make regarding the subject matter. He cannot, for example, assume that type-terms denote para-mechanical causal explanations where the concept of levels of description/abstraction are at issue, or assume that simple disjunctive syllogisms can be indiscriminately applied to syntax and semantics where the metaphysics of formal computer science is concerned.

Before I can connect the arguments in this chapter to those in the first chapter, I need to show that the basic method of Searle's argument is invalid within its own framework. I will return to the arguments presented in this chapter and their conclusions in the fourth chapter of this thesis, when I bridge the arguments of the first two chapters using the arguments of the third chapter. In a sense, the criticisms alluded to in this chapter by iterating the differences between Searle's conception of the Chinese Room, and Turing, White, and Dennett's conception of the Universal Turing Machine and the Turing Test, will be explicitly argued. There is, I will argue, a difference between these two understandings of the concepts at hand, and that difference is a difference in the kind of valid inferences that can be made concerning identity claims such as the Strong AI thesis.

3.0 Chapter Introduction

In this chapter I turn my attention to Gilbert Ryle's analysis of two category mistakes, some logical errors that Ryle finds in the traditional discussion of minds, and to the task of explaining how that analysis is relevant to the Chinese Room Argument (Ryle, p.20). As in the previous two chapters, I will provide some exegesis of Ryle's chapter "Descartes' Myth" and the arguments found therein. If this chapter seems like a departure from the previous two, then recall that the subject at hand is whether a proposition like 'software is to hardware like mind is to body' is true. Whereas the previous two chapters dealt with the relation of computer programs to their hardware, this deals with the relation of minds to bodies. Ryle addresses the relation of the mind to the body by addressing the logic by which we ordinarily categorize, identify, and distinguish things such as objects, properties, and categories.

Recall that in the first chapter I argued that Searle derives the conclusion of the Chinese Room Argument, computer programs add no mental properties to people instantiating them, from premises that a person instantiating a computer program gains no understanding of its subject. The Chinese Room Argument goes something like:

(1) Mental properties are either the result of a computer program, or some physical property of its physical mechanism such as a human computer.

(2) A person instantiating a computer program does not gain a mental property from instantiating it.

(3) Therefore, mental properties are the result of some physical property of human computers.

Where my purpose in this thesis is to show that John Searle's Chinese Room Argument is an invalid argument, I will need some method of argument that can be used to show that the conclusion of the Chinese Room Argument does not follow from its premises. Therefore, in this chapter I will show that such a method of argumentation is available, and that it is applicable to arguments of the same form as the Chinese Room Argument. I will argue that Ryle's analysis of the relation of the mind to the body, and of how we properly categorize that relation, is that method of argumentation.

Ryle's identification of the category mistakes in discussions about the mind is particularly pertinent to the Chinese Room Argument because it analyses an identity statement in terms of levels of description or abstraction. Thus, given the option of identifying a mental/semantic property in a collection of syntactic operations such as a computer program, or finding a physical semantic component in humans, Ryle's analysis reaches the former conclusion in order to avoid logical error. Ryle identifies mental properties with the category of syntactic operations such as conscious behaviour because the behaviour itself is incongruent with the application of mental predicates. Because it is the category that is specifically mental, rather than the categorized, it follows that any hypothetical physical semantic component will be ontologically superfluous. This step up in level of abstraction avoids the generalization that if some syntactic operations are not mind-like, then no syntactic operations are mind-like, and hence that if no abstract syntactic operations are mind-like where the abstract is non-physical or non-material, then mental properties must be physical or material.

Therefore, although Searle accepts the necessity of a special physical semantic component such as the brain, and does so because he rejects the idea that computer programs are anything other than purely syntactic (and thus non-physical in their nature), Ryle does not thereby accept that the mind is non-physical. Such a position, that by accepting computer programs to have content we must accept the necessity of a special mental (or semantic) category, can be rejected as resting on a pair of category mistake. Ryle is not just opposed to the reification of what are properly categories. Ryle also rejects the idea that categories of different logical types are exclusive. A higher level of abstraction does not necessarily make something less real, or even less physical. Considering such categories to be exclusive, such that the abstract must also be non-physical, and to make inferences based on such exclusivity is to misapprehend a schema of computation including a notion of levels of description or abstraction.

In this chapter I will argue that the positive thrust of Ryle's analysis is not merely an argument against the existence of special semantic objects called 'minds', but an argument against the validity of arguments such as the Chinese Room Argument. Therefore, I will argue that the positive upshot of that analysis is a way to identify invalid argumentation in the reasoning employed by the Chinese Room Argument. By arguing that Ryle's point about categorization, and not just about reification, I will be arguing that Ryle offers a method for identifying deductively valid arguments in arguments like the Chinese Room Argument.

3.1 Descartes' Myth

Ryle discusses what he calls "The Official Doctrine" (Ryle, p.13). The Official Doctrine is that minds, whatever they might be, are not located in space like material objects, such as bodies, and do not have some physical existence that might permit observation (Ryle, p.13). They occupy, as it were, a "mental world" apart from the usual "physical world" (Ryle, p.13). Ryle contrasts these worlds by way of the epistemic familiarity we are afforded of them; the physical world only provides “great or small uncertainties about concurrent and adjacent episodes in the physical world,” while an agent “can have none about at least part of what is momentarily occupying his mind” (Ryle, p.14). This “bifurcation” of mind and body is, Ryle says, expressible via metaphor by the bifurcation between the “inner” and the “outer” where we ignore the oddity of using such a spatial metaphor for the difference between something spatial and something notably non-spatial (Ryle, p.14).

Despite a congruency in the Official Doctrine by which some things may be said to have a “physical existence” and others to have a “mental existence”, minds are not connected via mental non-space as bodies may be via physical space (Ryle, p.15). Ryle says: “Only through the medium of the public physical world can the mind of one person make a difference to the mind of another”, with the possible exception of “telepathy” (Ryle, p.15). This privacy of other minds is further contrasted with the immediacy of a subject’s own mind and its contents, though, Ryle notes, the full and immediate consciousness of a subject’s own mind is perhaps in doubt. Because of these confusions with regard to the specifics of this Official Doctrine, and their disagreement with our everyday familiarity with minds and their contents, Ryle argues it is “entirely false, and false not in details but in principle” (Ryle, p.17).

The principle that Ryle refers to is the principle the violations of which are the category mistakes he discerns. But before I go further, and relate what this principle is, and how Ryle argues that it is violated, I would like to lay some groundwork by further discussing the metaphysics that Ryle has attributed to the Official Doctrine. In particular, Ryle has described a dichotomy, that of either mind or body, where these things are entirely separated, and whose interactions are deeply mysterious and problematic.

Recall from the first chapter how John Searle describes the Strong AI thesis as being the acceptance of the following equivalence: mind is to body as software is to hardware. If Ryle is correct about the nature of the mind under the Official Doctrine, then it is easy to see how this equivalence breaks down on contact: If the mind is a mysterious non-physical thing, and software is the sort of thing that one encounters in the physical world, then there is at least one way in which one side of the relation is not equivalent to the other. But Searle does not argue against the Strong AI thesis on the grounds that one might occasionally trip over a computer program and not a mind; to the contrary as noted Searle considers minds to be physical phenomena and software to be non-physical. Recall that Searle goes so far as to accuse the Strong AI thesis of a sort of 'dualism', the sort of dichotomy between mind and body for which Ryle criticizes the Official Doctrine.

Such a dichotomy, if it is not to be a false dichotomy, must take as its options those options that may exclude each other at the same level of description or abstraction, and exhaust the range of options available. Where more options, or combinations of options, are unpalatable, the result may be the establishment of such a dichotomy. In assigning motivation for the Official Doctrine Ryle suggests that it arises out of a tension between science and religion (Ryle, p.20). He argues that the consequences of supposing some mechanical explanation for the mental would be either religiously or scientifically unpalatable. The solution would be that:

> "Since mental-conduct words are not to be construed as signifying the occurrence of mechanical processes, they must be construed as signifying the occurrence of non-mechanical processes; since mechanical laws explain movements in space as the effects of other movements in space, other laws must explain some of the non-spatial workings of minds as the effects of other non-spatial workings of minds" (Ryle, p.20).

This lack of physicality would formally preserve traditional religious doctrine on the soul from dissent, while allowing the application of science to be uncontroversial. However, the philosophical effect of assuming such ontological superfluity is to allow ordinary connectives to be employed to make grammatically correct sentences denoting absurdities. The Official Doctrine does not merely make for a fanciful para-mechanical turn of phrase with regard to psychological explanation, Ryle argues: It licenses such nonsensical phrases as: "he bought a left-hand glove, and a right-hand glove and a pair of gloves" (Ryle, p.23). This is the sort of nonsensical grammatical construction that Searle employs in his discussion of digital computers, as noted in the first chapter.

3.2 The Category Mistakes

Once the Official Doctrine is understood, or at least understood as Ryle understands it, then we can see that any mistake people make about categories is the violation of a logical principle, and an invalid inference, rather than just the reification of a category. To illustrate this absurdity (where 'absurdity' is taken as meaning 'logical error' rather than an instance of contradiction) Ryle employs three examples to clarify the logical errors that could be categorized as 'categories mistakes':

(1) That of a university being mistaken for a building in that university,

(2) That of a military division being mistaken for one of its constituent units,

(3) That of an *esprit de corps* being taken for a role on a cricket team (Ryle, pp.18-19).

In the case of (1) & (2) a type-term or category is mistaken for being one of the objects it types or categorizes. In the case of (3), however, one category is being mistaken for another category; Ryle describes *esprit de corp* or "[t]eam spirit" as: "the keenness with which each of the special tasks is performed, and performing a task keenly is not performing two tasks" (Ryle, p.18).

Mistakes involving categories not an isolated violation of the usual logical rules for making inferences over categories; they are the source of the Official Doctrine (Ryle, p.19). Given that the Official Doctrine is a result of category mistakes, and that the Official Doctrine permits the body to be distinguished from the mind, and that the Chinese Room Argument resembles the Official Doctrine, then the falsity of the equivalence Searle attributes to the Strong AI thesis (that the mind is to the body as software is to hardware) looks like it involves some sort of category mistake. If software is not separable from hardware like, according to the Official Doctrine, the mind is from the body; then the equivalence Searle attributes to the Strong AI thesis may not be absurd.

Ryle indirectly puts us in the mind of the congruency between the Official Doctrine, and Searle's understanding of the relationship between software and hardware, when he speculates on "The Origin of the Category Mistake" (Ryle, p.20). Recall that the Official Doctrine is concerned not only with distinguishing minds from bodies, but with leaving minds open to similar modes of mechanical explanation as are applied to bodies. Ryle says: "The differences between the physical and the mental were thus represented as differences inside the common framework of the categories of 'thing', 'stuff', 'attribute', 'state', 'process', 'change', 'cause', and 'effect' (Ryle, p.20). And moreover that the theory went that: "Minds are not bits of clockwork, they are just bits of not-clockwork" (Ryle, p.21). Likewise, Searle's programs are non-material structures, things that do not exist yet can map onto what exists in an efficacious manner.

Given this mis-categorization of mind as something complementary to the profane material of everyday life, it should not be surprising that one of its results, at least according to Ryle, is a problem of access. Given that physical bodies only have access to other physical bodies in addition to their own associated minds, and that no access to the non-physical causes of a person's speech acts, that: "It would have to be conceded, for example, that [a person] could never tell the difference between a man and a Robot" (Ryle, p.23). This sort of argument should be familiar from the principles of the Chinese Room Argument. That we cannot, in principle, distinguish between persons and artifice by observation alone is precisely the problem that Searle attributes to the Turing Test. If these principles are faulty, then it may be the case that any argument predicated upon them is also faulty.

So, while Ryle's purpose in describing these category mistakes is mainly critical, to dissolve what he sees as a problematic theoretical framework for thinking about minds, his explication of particular category mistakes allows us to determine some non-problematic inferences about categories. The first sort of category mistake would be mistaking a category for an extra member: An example of the complement to this category mistake would be to suppose there is a class A to whose members we are being introduced. Upon the conclusion of introductions we agree that we have met A, although we have not met A in the same way as we have met the members of A, we cannot shake A's hand like we can shake the hands of its members (Ryle, p.19). There is an A; it is all the members of A in the ways that they are A. That is to say: A exists at the level of abstraction where only those properties shared by members of A are relevant.

The second sort of category mistake would be mistaking one category as excluding another category that shares members, such as all positions on a cricket team and the players exhibiting team spirit. Ryle distinguishes it from the first sort of category mistake by noting that: "Certainly exhibiting team-spirit is not such that we can say that the bowler first bowls *and* then exhibits team-spirit or that a fielder is at a given moment either catching or displaying *esprit de corps*" (Ryle, p.18). Classes do not necessarily exclude each other even at the same level of description, so while displaying team spirit is an action and bowling a ball is also an action, they do not exclude each other in principle. In practice, a player will display team spirit only in how they perform an action.

3.3 Chapter Conclusion

In the introduction to this chapter I proposed to argue that Ryle's analysis of these category mistakes could be used to show that the Chinese Room Argument is an invalid argument. To explicitly iterate that argument here in the conclusion to this chapter: There are two congruencies between the Official Doctrine that Ryle criticizes as making category mistakes and the premises of the Chinese Room Argument. The first congruency is mis-categorizing classes as things. The Official Doctrine considers the class of mental behaviour to be a non-material thing called a 'mind'. The second congruency is mis-categorizing non-identical classes as exclusionary. The Official Doctrine considers the class of mental behaviour as something exclusive from the class of physical behaviour. Searle considers the class of machines describable as computer programs to be non-material things called 'software'. Searle considers such non-material things, representations, to be something exclusive from material things.

As well as these two congruencies between the premises of the Chinese Room Argument concerning the nature of computer programs, and the Official Doctrine, there are congruencies between the subject matter that is addressed in each case. Ryle's criticism of isolating physical behaviour from mental phenomena (with the result being that an automaton could be mistaken for a person) can be applied to Searle's worry that mere behaviour is not enough to inform us about the presence of mental phenomena. Where Searle is concerned with arguing that mind is not separable from body as computer program is from its instantiating hardware, Ryle is concerned with arguing that mind is not separable from body if the relation of mind to body is the proper categorization of some body's states.

Therefore, there is good reason to believe that Ryle's critique of an argument like Searle's Chinese Room Argument can be used to critique the validity in the Chinese Room Argument. That is to say: If it can be shown that Searle's Chinese Room Argument requires an inference that mis-categorizes something by either multiplying members as well as categories or reifying categories at the level of description they categorize, then the Chinese Room Argument may be proved to be invalid. In the next chapter, I will combine the arguments of the first three chapters to make the argument that the Chinese Room is invalid.

4.0 Chapter Introduction

The first three chapters of this thesis cover John Searle's Chinese Room Argument, Alan Turing's Imitation Game, and Gilbert Ryle's analysis of the category mistakes that lead philosophers to dualism. In the first chapter I argued that the Chinese Room Argument given by Searle appeared to constitute an intractable problem for theories like the Strong AI thesis ('computer functionalism'). In the second chapter I argued that Searle had misunderstood the metaphysics of computation, and distinguished the metaphysics employed by Searle from the metaphysics described by Turing, Richard White, and Daniel Dennett. In the third chapter I described the category mistakes that Ryle finds behind theories like the Official Doctrine, and argued for their relevance to arguments like the Chinese Room Argument.

In this chapter, I will argue that the Chinese Room Argument is not actually a valid argument to the effect that no digital computers can have mental properties such as intentionality. While the Chinese Room Argument may be a good argument against some computer programs, in principle, having mental properties such as intentionality, it is a bad argument for the purpose that Searle employs it, as an argument against the possibility that a computer program may have mental properties.

My argument will be developed in three sections. The first section of this chapter will apply the Chinese Room Argument to the Imitation Game cited in the second chapter. The second section will show how the Chinese Room Argument involves the two category mistakes cited in the third chapter. Finally, in the third section, I will argue that the Chinese Room Argument is only a good argument against some digital computers having mental properties.

4.1 The Chinese Room Argument and the Imitation Game

Recall the Chinese Room Argument from the first chapter:

> (P1) If instantiating a program is sufficient for understanding written Chinese, and John Searle can instantiate a program, then John Searle can understand written Chinese by instantiating such a program.
>
> (P2) Instantiating a program does not enable John Searle to understand written Chinese.
>
> (C) Therefore instantiating a program is not sufficient for the attribution of mental predicates such as 'understanding' to programs.

As well, recall the conceptual framework of digital computation discussed in the second chapter:

(1) The notion of a computer as a material object, a token of some machine-type that implements computation,

(2) The notion of a computer as a mathematical model, or a type of machine that implements computation,

(3) The notion of abstraction by which groups of tokens can be considered as a single type,

(4) The notion of instantiation by which tokens can exemplify types.

(5) The notion of levels of description (or abstraction) which limits the scope of identity claims concerned with types/tokens, and by which multiple types of tokens can instantiate a type.

(6) The Imitation Game, or Turing Test, whereby a judge communicating with two subjects via an otherwise subject-neutral medium attempts to distinguish between the two, and where one of the subjects is a digital computer.

The notion of a computer as a material object allows us to refer to devices like the Chinese Room as computers, and for them to be appropriate subjects for the Turing Test. The notion of a computer as a type of machine licenses us to say that the same computer can be instantiated by different sorts of hardware. In this sense the term 'computer' refers to both hardware and software. The notion of levels of description limits the scope of identity claims, however, so that what is subject to the Turing Test is the notion of the computer as a program, rather than the qualities of the device implementing that program. John Searle instantiating an English Room and passing the Turing Test thanks to John Searle's understanding of written English, for example, would not count as evidence for the claim that the English Room program understands English.

Conversely, the notion of levels of description means that, at the level of description of the Turing Test, the program implemented by the Chinese Room is the same whether it is implemented on any suitable arrangement of hardware. This notion of levels of description also means that the computer program instantiated in the Chinese Room Argument is, at some further abstracted level of description, essentially the same as all computer programs. Hence, if a contradiction is derived from the premises of the Chinese Room Argument, then that absurdity can be said to apply to the Strong AI thesis in general, and not just to that program in particular. As related in the first chapter, Searle argues that the absurdity of the Chinese Room Argument is because programs are essentially non-material.

If, at the level of description with which the Turing Test is concerned, the computer is indistinguishable from the human subject, such that John Searle's writing is indistinguishable from that of a native Chinese writers when implementing the Chinese story program, and at some other level of description the computer is clearly distinguishable from the human subject, as when John Searle's body is distinguishable from the Chinese Room apparatus, then there is a question of why disagreement between differing levels of description is absurd. After all, if the scope of identity statements is limited to the levels of description at which they are made, then it is a mistake to use token-terms interchangeably with the type-terms that they instantiate, and from which they are abstracted, as that would involved a confusion of categories.

Therefore, I argue that there is an ambiguity in the Chinese Room Argument whereby it appears that the first premise is concerned with what might happen, and under what conditions, while it appears that the second premise is concerned with what does happen. What might happen is that John Searle might be able to understand Chinese by instantiating the Chinese story program. However, John Searle does not understand Chinese by virtue of instantiating the Chinese story program. The category of what computer programs may do is on a different level of description, and hence of a different category, from what actually happens when a particular computer program is instantiated.

Searle misuses the notion of levels of description by claiming it to be a relation of material objects to non-material schemas, such that any schema may be instantiated or realized by any material objects; his claim that multiple realizability entails universal realizability. As I note in the second chapter, Richard White argues that this entailment is, at least, false in the case of computation. By forcing the notion of levels of description into a dualist cast, Searle forces the lack of understanding in the Chinese Room Argument to apply to all possible programs and not just to the Chinese story program. Since this misuse of the notion of levels of description is not part of the premises of the Chinese Room Argument, the generalization it licenses does not generate an absurdity from the principles of computer functionalism.

An analogous situation might be where one supposes that the implementation of wheels on a car might make the car roll, and, when the car fails to roll, to conclude that the theory of wheels is absurd. Wheels, like computers, can be realized by a variety of mechanical substrates. It strains credulity that the failure of some particular wheels to roll is due to a critical failure of existence inherent to the nature of wheels, rather than a specific defect of design in the wheels being employed. Likewise, it strains credulity to concurrently suppose that the efficacy of some other particular wheels, call them English Wheels, is not the result of being wheels, but the fact that they are implemented with substrates of sufficient tire pressure and roundness to make them roll under their load. Searle's Chinese Room Argument shares this strain on credulity.

4.2 The Chinese Room and Its Category Mistakes

Searle's accusation of dualism on the part of the Strong AI thesis, discussed in the first chapter, is unwarranted since it depends on a misunderstanding of the notion of levels of description. Searle's accusation of dualism is better applied to his own interpretation of the Chinese Room Argument, since it considers the abstract to be divorced from the material instead of being entangled with the material instantiating it. The Chinese Room Argument in conjunction with such a dualist notion of levels of description makes the two category mistakes raised in the discussion of Ryle in the third chapter.

By categorizing mental properties as something other than the demonstration of capacity for those mental properties, via his interpretation of the Chinese Room Argument, Searle makes the category mistakes described by Ryle in the third chapter. Searle makes the first category mistake, that of confusing a category with one of the things in that category, the Chinese story program, with the category of programs. Searle also makes the second category mistake, confusing one category for another, when he interprets the software/hardware distinction involved in the Turing Test as though it were the dualist mind/body distinction.

Dennett gives a helpful counter-argument to this sort of epistemological dualism: "A playful god, or evil demon, let us agree, could fool the world's scientific community about the presence of H_2O in the Pacific Ocean" but "the tests they rely on to establish that there is H_2O in the Pacific Ocean are quite beyond reasonable criticism" (Dennett, p.9). Likewise one could suppose that a Turing Test might be limited to demonstrating only that something that looks suspiciously like understanding, but is not understanding. However: consider the practicalities of the tests to which one can subject water in order to demonstrate its chemical properties and affirm its identity. Where testing samples are judged to be statistically significant, disqualifying the properties of programs from being mental properties because they only appear to be the same is unreasonable.

To be fair, it is easy for Searle to mis-categorize following a set of instructions as instantiating a program because he considers computer programs in a very abstract way, as the manipulation of symbols rather than as the implementation of computation at some level of description. At the level of description where the implementation of computation may be accomplished by manipulating symbols, Searle's categorization is valid. At the level of description where the implementation of computation is put to the Turing Test, Searle's categorization mistakes the type 'program' or 'software' both for the token program, the Chinese Literacy program, and the sub-type 'hardware'.

4.3 The Chinese Room Argument Revisited

Returning to the Chinese Room Argument, let us consider it without making the categories mistakes that I argue Searle's original makes:

If John Searle is able to instantiate a program to communicate as if he were a Chinese Literate person, then he is doing the same thing that a Chinese Literate person is doing when that person is communicating effectively in written Chinese. If instantiating such a program by following a list of instructions means that John Searle is aware of the process and how it works, and does not let John Searle understand written Chinese, then that program is for communicating effectively in written Chinese and not for understanding written Chinese. It follows, therefore, that the particular program in question may be insufficient for mental properties such as understanding written Chinese but not for mental properties such as the skill to communicate effectively in written Chinese.

Consider that if there are programs with mental properties, then it may be the case that a program for communicating effectively is not also program for understanding, and that a machine might just as effectively communicate a complete lack of understanding. It follows from this conceptions of the Chinese Room Argument that the Turing Test, as Dennett has argued, could be usefully employed to distinguish between a program that could communicate effectively in written Chinese and a person who could both communicate effectively and understand the written Chinese characters. But the Turing Test could not be conclusively employed, as a program for understanding a language may not be able to communicate that understanding if it lacks access to a program for communication whose sophistication suits the subject matter.

As Dennett argues, the Chinese story program featured in the Chinese Room Argument is precisely this kind of isolated expert system that the Turing Test should be able to expose. But suppose that we take communicating effectively to mean the same thing as understanding, under the assumption that one must be able to understand a language in order to communicate effectively in it. In this case it appears that the premises of the Chinese Room Argument still entail an absurdity since passing the Turing Test and not understanding the language that the test is passed in should be mutually exclusive.

However, this is not the case because the Turing Test is not a kind of schema that *a priori* defines intelligence or the possession of mental properties. As Dennett argued, the Turing Test is a quick probe. It may even be the case that passing the Turing Test is not enough to judge whether some subject understands written Chinese in a significantly similar way to Chinese Literate people, despite the quick probe assumption. A particularly clever expert system that excludes mental properties such as understanding might do as well, and it seems that they do as well if we take PARRY, from the second chapter, as an example of problems with limited imitation games that might generalize to the full Turing Test.

It may be the case that the full suite of mental properties eludes linguistic functions. As mentioned in the second chapter, Dennett warns against using this sort of linguistic chauvinism as a loop-hole in the Turing Test. Hence it may be that while no program enables a digital computer to both pass the Turing Test and have the property of understanding, this does not generalize to the conclusion that no program can enable a digital computer to have the property of understanding. As mentioned in the first chapter, it may be the case that passing the Turing Test without understanding constitutes an interesting case of aphasia rather than a whole-scale lack of mental properties.

Divorced from its original purpose as an argument against the principles of the Strong AI thesis, this revised Chinese Room Argument is a heuristic for deciding what to do with the results of the Turing Test on a case-by-case basis. One can only draw the conclusion from it that, for the testing so far, the subject has yet to pass.

4.4 Chapter Conclusion

In the introduction to this thesis, and to this chapter, I said that I would put together the work of the previous three chapters to make the argument that the Chinese Room Argument was an invalid argument, and hence a bad argument against the Strong AI thesis. In the first chapter, I argued that, taken at face value, the Chinese Room Argument looked like a valid argument and as such presented an intractable problem to proponents of what Searle labelled the Strong AI thesis, computer functionalism. In the second chapter, I argued that the notion of computation described by Searle as the mere syntactic manipulation of symbols actually obscured a more complex metaphysics in which the dualism that Searle attributed to the Strong AI thesis was not native to it. In the third chapter, I described those mistakes Ryle identified with concepts of dualism in relation to mind, and how they applied to the Chinese Room Argument.

In this chapter I put Searle's Chinese Room Argument into the context that I had earlier argued proponents of the Strong AI thesis such as Dennett and Turing understand their claims about the nature of mind and computer programs. In doing so I argued that Searle makes category mistakes in taking the Chinese Room Argument out of its proper context, and hence that it is invalid. If Searle's Chinese Room Argument is invalid, and it is intended to be an argument against the principles of the Strong AI thesis, a *reductio ad absurdem*, then it does not pose an intractable problem for proponents of the Strong AI thesis; no contradiction follows from its premises. If the Chinese Room Argument is to be a good argument against the principles of the Strong AI thesis, then it must demonstrate an intractable problem following from those principles; it does not, and so is a bad argument.

I also presented a revised version of the Chinese Room Argument that I have argued is valid, and noted how its conclusion disagreed with the conclusion of the invalid version proposed by Searle. Where the category errors do not occur, and the conclusion is not over-generalized from the premises, the Chinese Room Argument leaves us with the practical problem of distinguishing between task-oriented expert systems and genuine intelligences that have mental properties (if not the specific property of understanding).

Of course, if I have misunderstood Searle's Chinese Room Argument, it may very well be the case that the Chinese Room Argument remains an intractable problem, and that the reasonable efforts given here to demonstrate otherwise have been in vain. Nonetheless, I hold that the Chinese Room Argument is not a good argument for Searle's purposes where computer functionalism is concerned, and for the reasons given in this chapter.

Bibliography

Dennett, Daniel C. "Can Machines Think?" Brainchildren: Essays on Designing Minds. MIT Press, Cambridge MA. 1998. Pp.3-20.

Ryle, Gilbert. The Concept of Mind. Penguin Books, Middlesex. 1949.

Searle, John. "Minds, Brains and Programs". A Historical Introduction to the Philosophy of Mind: Readings with Commentary. Peter A. Morton, ed. Broadview Press, Peterborough ON. 2003. Pp.282-289.

Turing, Alan. "Computing Machinery and Intelligence" A Historical Introduction to the Philosophy of Mind: Readings with Commentary. Peter A. Morton, ed. Broadview Press, Peterborough ON. 2003. Pp.265-282.

White, Richard. "Some Basic Concepts of Computability Theory". The Place of Mind. Brian Cooney, ed. Wadsworth, Belmont. 2000. Pp.204-216.