Syntax, Semantics and the Room with No View: Why Searle is Wrong to Think Computers Can't Think

Steven Frei

A Thesis

in

The Department

of

Philosophy

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts at
Concordia University
Montreal, Quebec, Canada

January 1996

ISBN 0-612-18461-7

Canada

ABSTRACT

Syntax, Semantics and the Room with No View:
Why Searle is Wrong to Think Computers Can't Think

Steven Frei

In this paper I argue that John Searle's Chinese room thought experiment is not, as Searle claims, based on the simple logical truth that syntax is not sufficient for semantics, but rather on the view that subjective conscious states are necessary for cognition, and that computers could not possess such subjective conscious states. I show that the only way Searle can claim that the underivability of semantics from syntax is a logical truth is by equating semantics with the *consciousness* of semantics. But this can only be accomplished by adopting the "first-person" approach to the mind, an approach which, I argue, makes it impossible to know which systems have conscious states and which systems lack conscious states. If Searle's methodological approach rules out the possibility of knowing whether computers possess conscious states, then he has no warrant for his claim that computing systems can't think.

For Kirsten

Contents

Introduction

In his book *The Rediscovery of the Mind* (1992), John Searle claims it is a logical truth that syntax is insufficient for semantics. This logical truth, says Searle, ensures that no computer could think in the same way that humans do. Because computers are defined solely in terms of their formal operations, and because these formal operations cannot produce the semantic content necessary for genuine thought, no computer could be a genuine thinking system, regardless of how intelligent it might appear. In the pages that follow I would like to respond to this view in two ways: first, I want to show that Searle's claim about the underivability of semantics from syntax only holds if semantics is equated with the consciousness of semantics. In other words, I want to show that Searle's claim that computers can't think is based not on a simple logical argument but rather on the view that a) subjective conscious states are necessary for cognition, and b) computers could not possess subjective conscious states. Second, I want to show that Searle's views on the relationship between consciousness and cognition are based on a methodological approach to the study of mind -- the "first-person" approach -- which makes it impossible for him to know whether or not computers could have conscious states.

Chapter one of this thesis outlines the cognitivist view that cognition is a form of computation. Cognitivists believe that computers, like humans, think in virtue of the ability they have to manipulate physical symbols in a meaningful way. I begin the chapter with a detailed explanation of what a

computer is and then move on to discuss the origins and fundamental tenets of the cognitivist or computational theory of mind. In chapter two, I look at Searle's Chinese room thought experiment, a critique of the cognitivist view that thinking is the same thing as symbol manipulation. The *Gendankenexperiment* is meant to show that computers can't think because the syntactic operations they perform cannot produce the intentional states necessary for semantics. I argue, however, that we cannot rule out the possibility that computers are intentional systems. Finally, chapter three shows how Searle's view of consciousness as an ontologically subjective phenomenon lies at the root of his criticisms of cognitivism. Searle believes that only beings with subjective conscious states could have the intentional states necessary for thought. He thinks it is just obvious that computers do not have subjective conscious states and so concludes that they cannot be thinking systems. In this chapter I show that the methodology of Searle's first-person approach to the mind rules out the possibility of telling one way or another whether computers are conscious. As a consequence, I argue that Searle has no warrant for the claim that computers can't think. The chapter ends with an examination of the Turing test and the role it plays in the cognitivist account of cognition. I argue that Searle's criticism of the claim that computers can think is based in part on an inaccurate assessment of the significance of the Turing test for the cognitivist research program.

Chapter 1

Cognitivism and Computers

By ratiocination, I mean computation
-- Thomas Hobbes (1588-1679)

1 Introduction

A cognitivist[1] is someone who believes that thinking is computation. On this view, computers, like humans, think in virtue of the ability they have to manipulate physical symbols in a meaningful way. What is particularly interesting about this claim is that the computations performed by computers are said to represent the same process that humans undergo when they think (Pylyshyn 1986, xv). In other words, computers think in the same way that people do. In this chapter I would like to take a detailed look at the cognitivist claim that computers can think. Toward this end I propose to do two things: 1) describe what a computer is and 2) briefly sketch the origins of the theory that cognition is a form of computation. By proceeding in this manner, I hope to show how the claim that computers can think is based on the fundamental idea that we -- in so far as we are thinking beings -- are ourselves computers.

[1]Cognitivism is also known as computational functionalism (Bechtel 1988), classical cognitivism (Clark 1989), Strong A.I. (Searle 1980, 1992) , GOFAI or Good Old Fashioned Artificial Intelligence (Haugeland 1985) and no doubt many other labels I've yet to come across. Although there may be slight differences in the views expressed under these labels, they all equate thinking with computation.

2 Interpreted Automatic Formal Systems

In his book *Artificial Intelligence: The Very Idea* (1985), John Haugeland describes a computer as an interpreted automatic formal system.[2] What he means by this, roughly speaking, is that a computer is a machine that formally manipulates meaningless tokens in such a way that when the tokens are interpreted as symbols (representations that stand for things in the world) the manipulations performed on them can be seen as meaningful. The next several pages will be devoted to trying to make some sense of this description. This will require spelling out in detail 1) what a formal system is, 2) how such a system is automated, and 3) what it means for an automated system to be interpreted.

According to Haugeland, a formal system can best be described as a kind of game in which tokens are manipulated according to formal rules. Chess is an example of a formal game (Haugeland uses "formal game" and "formal system" interchangeably). The pieces are the tokens, and the rules of the game are the rules according to which the tokens can be manipulated. Thus, advancing a pawn from E2 to F4 is an instance of token manipulation according to formal rules. It is important to note that a token need not be a small physical object like a chess piece. In tic-tac-toe, also a formal game, the tokens (Xs and Os) can be represented, among other ways, by making the appropriate marks with a pen or pencil. It should also be noted that there can be more than one type of token in a formal game. Tic-tac-toe has two different token types (Xs and Os) and chess has six different types of tokens per side.

[2]For the most part, my description of a computer comes directly from Haugeland (1985, ch. 2).

Formal games have three essential features: 1) they must involve token manipulation; 2) they must be digital; and 3) they must be finitely playable. Manipulating tokens involves at least one of the following operations: 1) relocating them; 2) altering them (or replacing them with different ones), 3) adding new ones to the position of the game; and/or 4) taking them away. To use chess again as our example, moving a piece can be seen as an instance of relocating a token, promoting a piece as an instance of altering a token, and capturing an enemy piece as an instance of taking a token away. Writing Xs and Os in the game of tic-tac-toe is an illustration of what it means to add new tokens to a formal game.

To define properly and exhaustively a token manipulation game it is necessary to establish the types of tokens, the starting position of each game, and what token manipulations the rules allow in any given position. If you had to give someone a complete description of the game of chess you would have to tell her what each piece was (e.g., a pawn or a king, etc.) the exact position each piece must be in at the beginning of the game, and the rules according to which the pieces could be legally moved. The ways in which tokens can be manipulated (the rules of the game) depend on the position that the formal game is in at any given moment. Any given position (any configuration of tokens) in a formal game can be changed into another position only by a token manipulation that is legal given the first position. In formal games, position is everything. A pawn that can be advanced two squares given one token configuration (the starting position) is restricted to advances of one square in all other token configurations. The position not only determines the rules of a formal game, but ultimately the types of tokens

as well. Consider the starting position of a chess game. Before any move has been made all the pieces except the pawns and knights could be interchanged without making any difference to what moves white could make to begin the game.

The importance of position to token manipulation games -- formal systems -- cannot be overemphasised. When a computer manipulates tokens it does so only according to the form or shape of the tokens. As we saw from our chess example, the form that tokens are in (i.e., their type) depends on their configuration or position and nothing else. It is in this sense that formal games are said to be self-contained. The fact that these tokens might stand for something in the outside world is irrelevant to the manipulations performed. As Haugeland puts it, "meaning is not a formal property" (1985, 50).

As stated earlier, formal games must be digital. What this means is that the tokens of a formal game are always of a definite type, or in other words, the tokens are always in one discrete position or another. A formal system that operates digitally must be able to recognise and produce tokens that are in discrete states. If a system can do this then it is a digital system. According to Haugeland, a digital system is a set of positive read/write techniques. To write means to produce a token, and to read means to identify what type a token is and what position it is in. Making the move E2-E4 to begin a chess game is an example of writing a token and identifying the resulting position is tantamount to reading a token.

But what about the term "positive"? What makes a read/write technique positive? A positive technique for reading and writing tokens is one that is guaranteed to succeed absolutely and without qualification. To use Haugeland's example, a positive technique for producing a board between five feet eleven inches and six feet one inch would be to measure off six feet on the board, draw a line and saw along the line (1985, 53). Such a technique is positive because it is possible for it to succeed perfectly. On the other hand, if the goal of the board-cutting exercise were to get a piece exactly six feet long, then there would be no positive technique available since such exact measurements (say, to the millionth of an inch) are impossible to make. So whether a technique is positive depends on what could count as success. A positive technique for advancing a pawn from E2-E4 is to move the pawn from the E2 square to the E4 square. This may sound ridiculously obvious, but it illustrates a significant point about digital systems. If the rules stipulated that pieces had to be placed exactly in the middle of the square, then chess would not be a digital game, since there is no possible method for positively determining whether a figure is in the exact centre of a square. A chess game is always unambiguously in one discrete position or another only because digital systems allow for a margin of error. It is this margin of error that permits digital systems to achieve perfection. A knight at G1 at the starting position of a chess game will always be the same token, no matter where it is on the square and no matter what it looks like.

It is the digital nature of formal games that makes medium independence possible. A game that is medium independent is one that can in principle be instantiated in any number of different kinds of material. What medium is used to play the game is unimportant. What are important in formal games

are the rules according to which tokens are manipulated. Thus chess could be played using regulation pieces and a standard tournament board or it could be played with figure skaters on an ice surface (this was actually done in London in 1953 when a famous game between the American Paul Morphy and the Duke of Brunswick was played out as part of the pantomime "Sinbad the Sailor"). But it is only because there is a positive read/write technique in chess that medium independence is possible. Moving tokens and differentiating positions relies ultimately on the configuration of tokens, and these configurations can be exactly duplicated in any number of media given the margin of error in positive read/write techniques. Given the same configuration of tokens, a pawn is a pawn whether it is made of ivory and sitting on a marble chessboard or whether it is made of cheese and sitting on one of 64 appropriately laid out Ritz crackers.

The third essential feature of any formal game is that it be finitely playable. This means that no player should need infinite powers to make moves or recognise positions in a formal game. All that should be required of a player is that he be able to follow the rules of the game. Following rules requires that the player is 1) able to tell for any proposed move whether such a move would be legal and 2) capable of producing at least one legal move or showing that no legal move is possible. To perform (1) and (2) a player must have a limited repertoire of logically primitive operatives that can be combined to form complex operatives. Constructing complex operatives out of primitive ones necessarily involves more rule-following. The rules to follow in this case are algorithms which, for automatic formal systems, must be understood ultimately in mechanical terms.

An algorithm is an infallible step-by-step procedure for accomplishing a task in a finite number of steps. The steps to take in algorithmic procedures are fully determined and completely obvious. There are two types of algorithms. the straight schedule algorithm and the branched or conditional schedule algorithm. To execute the former, a player obeys one primitive instruction and then moves on to the next. Such a schedule is inflexible and dictates the same sequence of moves regardless of input or the outcome of any of the steps involved. In a branched schedule, on the other hand, the next step in a sequence depends on the results of the previous step. A player must be able to answer "yes" or "no" to questions about the previous step and then branch out to one of two possible sequences depending on the answer she gives. A branched schedule with disjunctive branching conditions and primitive instructions is a primitive algorithm which can be followed by any finite player.

Having discussed formal systems we can now go on to look at automation. Our discussion of algorithms can be seen as a link between formal systems and automation since, as I have already noted above, being able to produce legal moves (i.e., following primitive algorithms) should be understood ultimately in mechanical terms. To see how this is so we must first say a bit about what an automatic system is.

Simply put, an automatic system is one that works by itself. A formal system is automatic if it is a physical system with two essential features: 1) some of the parts of the system are tokens in some formal system, and 2) the system automatically manipulates these tokens according to the rules of the formal system. Any automatic formal system must consist of tokens, one automated

referee, and at least one automated player. If we think back now to the feature of finite playability we will remember that the requirement for a player to be able to play a formal game was the capacity to follow rules. These rules turn out to be primitive algorithms and it is precisely these kinds of primitive algorithms that are carried out by the automated player(s) and referee of an automatic system. The referee follows the recipes of the algorithms and the player performs the primitive tasks. The connection between formal systems and automatic systems is summed up in the automation principle which states that whenever the legal moves of a formal system are fully determined by algorithms, then that system can be automated.

Up to this point I have described interpreted automatic formal systems as machines that manipulate tokens according to formal rules. But until these manipulations are interpreted they cannot be described as meaningful. I'll now try to show what it means for something to be an *interpreted* automatic formal system.

To interpret the activity of an automatic formal system, its atomic tokens must be assigned meanings -- i.e., they must be seen as symbols that stand for something in the outside world. In addition, the way in which these symbols combine to form complex symbols (the rule of structure, or the syntax) must be specified. Once an automatic formal system has been interpreted the tokens can be seen both as meaningless markers in a self-contained formal system *and* as meaningful symbols. To see what is meant here, think of the 1s and 0s that form the syntactic heart of any digital computer. These 1s and 0s are tokens that are instantiated in ranges of electrical voltages so that one voltage level designates a 0 and another level a 1. As in any automatic formal system

the tokens are manipulated strictly in virtue of their form. In the case of 1s and 0s, the form is the voltage level. Now imagine a token manipulation consisting of adding a new token, say, 01. Such a token is produced only when both the voltage level representing 1 and the voltage level representing 0 are present in the automatic system. If either of the voltage levels is absent then 01 will not be produced. It's important to remember that the production of the token 01 is, if uninterpreted, nothing more than meaningless gear-grinding (so to speak). It is purely the result of physical forces (in this case electricity) operating according to certain automated algorithms.

But now imagine that 0, 1, and 01 are symbols that stand for atomic elements in propositional logic. The 0 stands for the proposition "John loves Mary", the 1 for "Mary loves John", and 01 for the conjunctive proposition, "John loves Mary and Mary loves John." Suddenly the gear-grinding becomes meaningful because the ways in which the tokens are automatically manipulated mirror the rules of propositional logic. That is, the tokens are manipulated in a way that preserves the *semantic* relations that hold between the propositions the tokens represent in their capacity as symbols. The computer can produce the token 01 (the conjunctive symbol) only if both 1 and 0 are present in its system just as the happy state represented by the proposition "John loves Mary and Mary loves John" can occur only if it is the case that both "John loves Mary" and "Mary loves John". In other words, the physical operations of a computer are what Andy Clark calls *semantically transparent*. According to Clark, a system is semantically transparent if "there is a neat mapping between states that are computationally transformed and semantically interpretable bits of sentences" (Clark 1989, 2). In an interpreted automatic formal system, symbol manipulations are structured so that they

echo the semantic structure of what they are meant to represent. Thus if the token manipulations are properly configured, semantic coherence is guaranteed and it is in this respect that the computations of computers can be described as meaningful. What is important about the notion of interpreted systems is the underlying assumption that the formal or syntactical nature of computations, when interpreted, is sufficient to produce meaning or semanticity. As we shall see in the next chapter, Searle challenges this position, claiming that it is a "simple logical truth" that syntax is not sufficient for semantics (Searle 1992).

To sum up, an interpreted automatic formal system is a physical machine that formally manipulates symbols in a meaningful way. Now that we know what a computer is, the next question to ask is, "Why do cognitivists believe that such a system thinks?" As a first step toward answering this question, I will look now at the origins of the idea that thinking is tantamount to token manipulation.

3 The Turing Machine

Although Thomas Hobbes first equated thinking with computation over 300 years ago, Alan Turing (1912-1954) was the first person to provide a formal account of the nature of computation.[3] This he did using a theoretical device known as the Turing machine. Daniel Dennett has characterised the Turing machine as the result produced from asking the question "What do I do ... when I perform a computation?" (Dennett 1991, 212) In other words, a

[3]Contemporaneously with Turing, the logicians Emil Post and Alonzo Church advanced (independently) analogous theories of formalised computation (see Newell and Simon 1976).

Turing machine, according to Dennett, is a formal account of the sequence of mental acts that are primitive operations involved in computation. Imagine, for instance, multiplying a four-digit number by another four-digit number without the aid of a calculator. The computation is performed by first recognising which rules apply (the rules of multiplication as opposed to, say, addition), then applying the rules (multiplying the digit on the far right of the bottom row by the digit on the far right of the top row), then writing down the result, etc., etc. Turing distilled the essence of these kinds of mental acts and created in the Turing machine what Dennett calls an "idealisation ... of a mathematician performing a rigorous computation" (1991, 212).

The basic architecture of the Turing machine consists of the following elements: 1) a potentially infinite tape divided into linear squares on which 1s and 0s are written and 2) a head that moves back and forth across the tape.[4] The head (which is always in one of a finite number of internal states) scans each square individually in order to perform the following operations: a) determine whether 1 or 0 is written on the square and b) write either 1 or 0 (replacing whatever digit was scanned). When it has performed these two operations, the head moves to a square immediately adjacent and repeats the same procedure. Which digit the head writes and which square it moves to are determined by a prespecified set of instructions which are conditional upon what digit is scanned and what internal state the head is in. For example, the instructions executed by the head could be as follows: if in state A when over a square containing 0, then replace the 0 by writing 1, move one

[4] My account of a Turing machine is taken from Bechtel (1988) and Haugeland (1985). It should be noted that Turing never built an actual physical Turing machine although it is possible to build one. The Turing machine is not really a machine at all but rather a theoretical entity whose principles of operation form the basis of all digital computers, from the largest mainframe to the smallest laptop (see Dennett 1991, 214).

square to the left and change to state B. Whenever a head has no further instructions for the state it is in and the digit it is scanning, it will come to a stop. Any problem a Turing machine is meant to solve is determined by the symbols on the tape, and the answer it produces are the symbols left on the tape when the head comes to a halt.

As Haugeland points out, a Turing machine can easily be described as an automatic formal system (1985, 135). Recalling our discussion of computers in section 1, we will remember that an automatic formal system is a formal system that works by itself. For example, the game of chess is a formal system while the *Chess Challenger 3000* program implemented on a Macintosh computer is an automatic formal system. A Turing machine is a formal system because it involves token manipulation (the scanning and writing of 1s and 0s), it is digital (each 1 and each 0 is discrete and always unambiguously in one and only one square on the tape) and it is finitely playable (the head manipulates tokens according to algorithms that are simple enough to be automated).

So not only does a Turing machine outline the "bare bones" or primitive operations essential to mental computation (Dennett 1991, 212), it also describes the workings of an automated formal system. In fact, giving a formal account of mental computation in terms of a rigorously specifiable procedure turns out to be the *same thing* as showing how computations can be mechanised (Pylyshyn 1986, 50). And showing how computations can be mechanised is just describing how a computer works. Remember, for a formal system to be automated it must be finitely playable -- i.e., the rules the player(s) and referee follow to perform token manipulations must be in the

form of primitive algorithms (step-by-step procedures for accomplishing a task in a finite number of steps). Turing's great accomplishment was that he came up with these primitive procedures which satisfied the finite playability requirement. By formalising mental computation, he simultaneously showed how these computations could be mechanised. In short, a Turing machine shows that mental computation and mechanical computation can be formally equivalent. On Turing's scheme of things, a machine and a person solving the same mathematical problem are thinking in *exactly the same way* (provided of course that they both get the right answer) because what defines thinking in this case is a set of formal, rigorously specifiable procedures consisting in the manipulation of tokens. Whether these token manipulations are carried out by a machine or by a human is unimportant (or, if you like, immaterial). What matters to thinking, according to Turing, is a formal process or function that, in principle, can be carried out by anything whatsoever. As Pylyshyn notes:

> Turing's work can be seen as the first study of cognitive activity fully abstracted in principle from both biological and phenomenological foundations ... It represents the emergence of a new level of analysis, independent of physics yet mechanistic in spirit. It makes possible a science of structure and function divorced from material substance ... Because it speaks the language of mental structures and internal processes, it can answer questions traditionally posed by psychologists. (1986, 68)

If, as Turing claims, thinking is an abstract formal process that has to do with structure rather than with substance, then *anything* which embodies the appropriate structure and carries out the appropriate processes would, by definition, be a thinking thing (this is the idea of medium independence discussed above in section 1). In other words, anything that manipulates

tokens in the appropriate way must be thinking, including, as Pylyshyn informs us, "a group of pigeons trained to peck as a Turing machine" (1986, 57).

We are now in a better position to understand why cognitivists claim that computers can think. As the discussion of the Turing machine makes clear, a computer is a mechanical instantiation of the formal processes which cognitivists believe constitute the very essence of thought. According to cognitivists, computers think because they do exactly the same thing we do, i.e., they *compute* by manipulating tokens. A Turing machine writing 1s and 0s on a tape and a person multiplying numbers together are executing the same formal processes, and it is these processes that constitute thinking.

But isn't there a crucial difference between a computer executing a formal process and a person executing the same process? Compare a Turing machine solving a multiplication problem by manipulating tokens with a human solving the same problem. In the case of the Turing machine (as is the case with all digital computers) the tokens being manipulated are meaningless, whereas in the case of the human's computations we want to say that they mean something. In other words, to a person working through the equation 7+ 5 = 12, the digits and signs are more than just markings identifiable by their form; they make sense together because the meaning of 7 + 5 is the same in some respect as the meaning of 12. So even though someone could arrive at the answer 12 strictly by manipulating the formal tokens 7 + 5 according to rules that apply to tokens in that particular form and position, the way she usually gets the answer is by accounting for meanings. A computer, however, cannot account for meanings since, as we have shown above in our

discussion of automatic formal systems, tokens are manipulated strictly in accordance with their form and nothing else. And form has nothing to do with meaning. Imagine a chess game played between two Medieval barons who agreed that the pawns would represent serfs to be won or lost. Despite the stakes, the pawns could be manipulated only according to formal rules. What the pawns represent -- the meaning of the pawns -- can have no effect on the way that they are manipulated because meaning is not a formal property. Because all computers are formal games (formal systems) the manipulations they perform are by definition meaningless.

4 Interpretation and Semantics

So how is it that a cognitivist can equate machine token manipulation with human token manipulation? Because, argues the cognitivist, once we *interpret* the token manipulations of a machine they become meaningful. As we have seen from our discussion of computers, an *interpreted* automatic formal system can be said to think because the token manipulations it performs become meaningful when the tokens are interpreted as symbols. Interpreted formal tokens, then, can be seen either as meaningless markers in a formal system or as symbols which are related to the outside world. As Haugeland puts it, these tokens "lead two lives", a syntactical life and a semantic life (1985, 100).

But why should cognitivists claim that token manipulations are meaningful just because the tokens are seen as representing aspects of the outside world? Why (to use the arithmetical example again) should a machine be described as engaged in meaningful activity when the tokens it manipulates are

interpreted as symbolising the Arabic numerals and signs used in arithmetic? After all, if meaning is a result of interpretation, couldn't we describe anything as meaningful? Why not interpret the firing of pistons in a motorcycle engine as Wittgenstein's *Tractatus*?

The answer, of course, is that every interpretation is constrained by the syntactical structures of a machine's token manipulations. A certain machine can be interpreted as adding 7 and 5 (i.e., the tokens the machine manipulates can be seen as standing for Arabic numerals and signs) only if the structure of token manipulations mirrors or echoes the semantics or meanings inherent in arithmetic. In other words, a machine can be interpreted as performing arithmetical functions only if it is syntactically designed in such a way that its token manipulations are (to use Andy Clark's phrase again) semantically transparent. In an important way, then, meaning is already encoded or programmed into an automatic formal machine. This is what Haugeland refers to as the Formalists' motto: "If you take care of the syntax, the semantics will take care of itself" (1985, 106). The meaning is in the form of the rules that automatic players and referees follow when manipulating tokens in a formal game or system. In one sense these rules are meaningless; indeed they have to be meaningless if they are to be followed by a machine. But in another sense they are meaningful because the structure they possess mirrors the interpretation that they are given. Arithmetic is a formal game. If you follow the rules of the game (i.e., the primitive algorithms that machines can follow) by manipulating formal tokens in the appropriate way, then the answers you get to arithmetical problems have to be true -- they have to make semantic sense. In theory you can do arithmetic just by following rules about how certain digits and signs combine to form other digits, without

understanding what you're doing. But when you're operating in this fashion you are acting meaningfully because the rules you are following can be interpreted semantically. A machine following the rules of the formal game of arithmetic acts meaningfully because the rules it follows (the syntactical nature of its manipulations) limn the essential structure of arithmetic. Think back to Dennett's conception of Turing distilling the *formal* nature of computation. The Turing machine captures the essence of computation in the form of primitive algorithmic rules, and it is these rules that constrain semantic interpretation.

We have now answered the question "Why do cognitivists believe that computers think?" Thinking, on the cognitivist scheme, is an abstract formal process that involves token manipulation. This process can be instantiated in machines by automating the rules of formal games or systems. Computers think, then, in virtue of the token manipulations they perform. And, as we have just shown, these manipulations are more than just physical gear-grinding, since once they are interpreted they become meaningful.

In examining the cognitivist claim that computers think I have up to this point provided an abstract account of a computer, as well as some details concerning the development of the theoretical underpinnings of the cognitivist view that computers think in virtue of manipulating tokens/symbols. In the next chapter I shall look at Searle's criticism of this view as outlined in his essay "Minds, Brains, and Programs" (1980).

Chapter 2

The Chinese Room

How on earth can one thing represent (or 'stand for', etc.) a different thing?

-- Hilary Putnam, *Reason, Truth and History*

1 Introduction

In "Minds, Brains, and Programs" (1980), Searle argues against the cognitivist claim that thinking is the same thing as symbol manipulation. His say is centred on the now-infamous Chinese room thought experiment which is meant to show that, contrary to what cognitivists believe, the appropriately programmed computer cannot understand natural languages in the same way that humans do. Searle claims that his *Gedankenexperiment* is based on the simple logical truth that syntax is not sufficient for semantics. Thinking, says Searle, requires "genuine semanticity" (real as opposed to derived meaning) which in turn requires intentional states. Because the formal operations of computers cannot produce intentional states, they cannot be equated with thinking. In what follows, I argue that Searle's claim about the underivability of semantics from syntax is not a logical truth and that we cannot rule out the possibility that computers are intentional systems.

2 The Gedankenexperiment

Here's how the Chinese room thought experiment goes: Searle has himself shut up in a room where he receives (through a slot in the door) batches of Chinese script along with instructions in English telling him to a) join together certain symbols from the various batches and b) send the product of his symbol-joining efforts out of the room. Because Searle doesn't understand Chinese, his symbol-joining operations are informed only by the shape of each Chinese symbol. Thus, the instructions in English might tell him to join the symbol that looks like a backwards "F" to the symbol that looks like a garden hoe. Unbeknownst to Searle, the batches he receives represent a story as well as questions about the story, and the collection of symbols he sends out of the room represent answers to these questions. After a while, Searle becomes so quick and proficient at joining symbols that observers outside the room are unable to distinguish his answers to the story from the answers a native Chinese speaker would give to the same questions.

The moral of the *Gedankenexperiment*, according to Searle, is that symbol manipulation is not sufficient for understanding. Or, to put it another way, taking care of the syntax does *not* take care of the semantics. Although, given the same input, Searle is able to produce output identical to what a Chinese speaker would give, he still does not understand a word of Chinese. This is because the answers he sends out of the room are based solely on the manipulation of *meaningless* symbols. Now, according to Searle, what he does in the Chinese room is exactly what any "Turing machine simulation of human mental phenomena" does (1980, 67). Contrary to the cognitivist's claim, a language-understanding program implemented on a digital

computer does not understand a word of Chinese because all it does is manipulate meaningless markers in exactly the same way that Searle shuffles meaningless Chinese symbols. Searle explains the similarity of the two cases in the following way:

> I have inputs and outputs that are indistinguishable from those of the native Chinese speakers, and I can have any formal program you like, but I still understand nothing. For the same reasons ... [a] computer understands nothing of any stories, whether in Chinese, English, or whatever, since in the Chinese case the computer is me, and in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing. (1980, 70)

To give the thought experiment extra punch, Searle asks us to imagine that while in the room he also receives stories and questions written in English to which he provides answers, also written in English. To observers outside the room, the answers Searle provides in Chinese and English are equally good -- they both show a complete understanding of the stories and questions. But while there is no difference on the outside, the difference on the inside makes all the difference in the world. In preparing the English output, Searle understands the meaning of the stories and questions he reads, as well as the answers he gives to these questions; in preparing the Chinese output, he understands nothing at all.

What is particularly clever about Searle's thought experiment is that it crystallises the common-sense intuition that machines can't think without giving any *reasons* why they can't think. What goes on in the Chinese room does not explain why or how Searle understands the English story but not the Chinese one, nor does it explain why an appropriately programmed computer

doesn't understand either story. In other words, it doesn't explicitly state why symbol manipulation in itself is not sufficient for understanding. It only shows that the kind of thinking or understanding that humans engage in *seems* to be different from what a computer does when it manipulates symbols. In short, the thought experiment leaves us with the claim that manipulating symbols is not thinking because thinking is something other than manipulating symbols. Put another way, it shows that syntax is not sufficient for semantics because semantics requires something -- we know not what -- that is more than syntax.

Daniel Dennett claims that Searle's *Gedankenexperiment* is not an argument at all but rather an intuition pump (1987, 324). In *The Rediscovery of the Mind*, Searle seems to admit as much when he claims that his thought experiment "*rests on* the simple logical truth that syntax is not the same as, nor is it by itself sufficient for, semantics" (1992, 200, my emphasis). In other words, the Chinese room is meant only as a colourful and vivid reminder of something that, according to Searle, is self-evident. The thought experiment provides no argument for the claim that we can't get semantics from syntax for the simple reason that it is *based on* the claim that we can't get semantics from syntax.

Let's turn now to the claim that syntax is insufficient for semantics. Searle would have us believe that the proposition "syntax is not the same as, nor is it by itself sufficient for, semantics" is of the same logical form as the class of propositions typified by such sentences as "No unmarried man is married." According to Quine, a statement that is a logical truth is one "which is true and remains true under all reinterpretations of its components

other than the logical particles" (1980, 23). For instance, we can substitute "dressed" for "married" in the above example and get the following logical truth: "No undressed man is dressed." Now, what if we were to reinterpret some of the non-logical components of the statement "syntax is not sufficient for semantics" in order to get the new statement "love is not sufficient for happiness"? Since it is not at all clear that love is not sufficient for happiness, the new statement could be false and as a consequence the proposition "syntax is not sufficient for semantics" cannot be a logical truth since it does not remain true under all reinterpretations of its non-logical components. As far as I know, Quine's definition of a logical truth is accepted as standard in philosophical circles and it is consistent with definitions of logical truth that appear in introductory texts on logic.[1] Such definitions are usually something along the following lines: "a sentence is logically true if and only if it is not possible for the sentence to be false." For example, the statement "a horse is a horse" would qualify as a logical truth because to deny it would be to utter a contradiction. Now clearly, the proposition "syntax is not the same as, nor is it by itself sufficient for, semantics" is not of the same form as "a horse is a horse", since it is not at all clear that its denial would result in contradiction.

It is possible, I suppose, that by "logical truth" Searle means something less formal and exact than the meaning expressed by Quine's definition. Perhaps he means something like "obvious" or "clear to anyone who has a moment to reflect". But this characterisation of logical truth won't do either given that there are a good number of people in cognitive science and the philosophy of mind (presumably with time to reflect) who see nothing obvious about the claim that syntax is not sufficient for semantics. Why then does Searle believe

[1] For examples see Teller (1989, 38) and Bergmann, Moor, Nelson (1980, 15).

it is a logical truth that semantics is not derivable from syntax? Part of the answer, as we shall see, has to do with the notion of intentionality.

3 Intentionality

After outlining his *Gedankenexperiment*, Searle asks

> ... what is it that I have in the case of the English sentences that I do not have in the case of the Chinese sentences? The obvious answer is that I know what the former mean, while I haven't the faintest idea what the latter mean. *But in what does this consist and why couldn't we give it to a machine, whatever it is?* (1980, 71; italics mine.)

What Searle has in the case of the English sentences but not in the case of the Chinese sentences is intentionality. Intentionality, says Searle, is best characterised as the "directedness" of our mental states (1979, 74). Mental states are intentional when they are directed at, or are about, objects or states of affairs in the world. Such things as beliefs and desires are intentional states whereas things like "raw feels" or pains are not. I can have a belief that there is a large cat on the mat in front of me or a fear that the large cat is in fact a hungry Siberian tiger, but there is no sense in which a sudden pain caused by the tiger's claws sinking into my arm is about anything or directed to anything in the world beyond the pain itself.

It is this "directedness" or "aboutness" of mental states that is the key to understanding Searle's claim that symbol manipulation does not in itself

constitute understanding. Formal operations on symbols are not *about* anything. Inside the Chinese room, Searle understands the English story but not the Chinese story because he has intentional states about the former but not the latter. The letters, words, and sentences that make up the English story *mean* something to Searle -- they are *about* something. The Chinese symbols, on the other hand, are just meaningless squiggles; in fact they aren't even symbols, because for Searle they don't stand for or represent anything. Since a computer implementing a formal program does no more and no less than Searle does in the Chinese case, it cannot have intentional states and so *ipso facto* does not understand in the way that Searle understands the English story.

But why does Searle have intentional states while the appropriately programmed computer has none? Because, says Searle,

> 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. (1980, 86)

The specific biochemistry of the origin of intentionality is, according to Searle, that of the brain. Humans have intentional states because they have biological brains and computers lack intentional states because they don't have biological brains. The only way for a computer to think, on this scheme of things, would be to endow it with the same causal powers possessed by the brain. Whether this could be accomplished using anything other than the biochemical matter that makes up *our* brains is an empirical question, says Searle. But if a computer could be endowed with intentionality by somehow artificially duplicating neuronal powers, then it wouldn't be thinking in

virtue of the formal manipulation of symbols; any intentional understanding displayed by such a computer would be the result of whatever duplicates the *causal* operations of a biological brain. For Searle, then, what is crucial to cognition is the stuff or matter that produces it. In direct contrast to the cognitivists, whose account of thinking has to do with abstract formal processes in principle realisable in any physical medium, Searle holds that thought without brains (or something with identical causal powers to the brain) is simply impossible.

To sum up, Searle claims that the formal manipulation of symbols is meaningless because meaning can only be found where there is intentionality and "no purely formal model will ever be sufficient by itself for intentionality because the formal properties are not by themselves constitutive of intentionality" (1980, 82). So it is intentionality, according to Searle, that accounts for the self-evident truth regarding the nature of the relationship between syntax and semantics that is brought to light in the Chinese room thought experiment. Syntax is not sufficient for semantics because semantics requires intentionality, and purely syntactical operations are not intentional because they lack the causal powers requisite for "aboutness". If semantics or meaning is intentional and syntax is not, then it is self-evident that syntax is not the same as, nor sufficient for, semantics.[2]

Why, though, should we claim with Searle that intentionality requires brains or something with equivalent causal powers? Why should "aboutness" be inextricably linked with causal powers, neuronal or otherwise? Why can't

[2] Once intentionality is brought into the equation we can formulate the following logical truth: "something non-intentional (syntax) is not identical to something intentional (semantics)."

machines have intentionality even thought they are not biological? There is no a priori reason, it seems to me, why intentionality should be a biological phenomenon. Indeed, Franz Brentano, the 19th-century philosopher who in many ways set the agenda for discussions of intentionality, gave a characterisation of intentional states that precluded any role whatsoever for the brain. For Brentano intentional states were the mark of the mental as opposed to the physical; no *physical* state (biological or otherwise) could be an intentional state because no such state could be directed toward objects in the same way mental states could be (Bechtel 1988, 41-42). I'm not suggesting, as Brentano did, that intentionality necessarily entails a form of substance dualism (Cartesian mind-body dualism would, in fact, rule out artificial intelligence altogether); I'm only trying to show that there is no obvious reason why one kind of substance rather than another should possess intentional states if both substances exhibit behaviour consistent with intentionality.[3]

If intentionality is not necessarily a function of causal powers, then the cognitivist theory of semantic transparency seems to offer a reasonable explanation of what "aboutness" could mean. As we saw in chapter 1, formal manipulations have "aboutness" in so far as they echo the semantic structure of the propositions they are meant to represent. And if they are *about* a projected semantic content, then they are intentional. As Andy Clark points out, "a system that performs systematically with respect to a certain semantic

---

[3] To those who respond to this by saying something like "I just don't see how a heap of silicon can have intentional thoughts," one only has to remind them that cutting open the top of somebody's head would probably evoke a similar reaction with regard to the grey matter they would find inside a skull. How in the world, after all, can a lump of wrinkled wet mushy stuff be *about* anything? As things stand now, neurobiologists do not (and nor does anyone else for that matter) have any idea how the brain causes intentionality (see Churchland 1986, 346).

description ... has the effect of making the semantic description a real object *for the system*" (1989, 19; italics mine). Searle inside the Chinese room "performs systematically with respect to a certain semantic description" and consequently can be said to be acting intentionally. In other words, once we drop the biological criterion for intentionality, it is not at all clear that, given the notion of semantic transparency, syntax is not sufficient for semantics.

4 Intrinsic or Original Intentionality

According to Searle, however, intentionality *is necessarily* linked to the brain because only the brain (or something with equivalent causal powers) can produce what he calls *intrinsic* intentionality. Intrinsic intentionality, says Searle, is the only genuine intentionality there is, and he opposes it to the derived intentionality which he claims is the kind of "aboutness" that computers have in virtue of being semantically transparent (1992, 78). To see what Searle means by this distinction between derived and intrinsic intentionality, consider the proposition "The cat is rather fat" both as a mental state (a thought) caused by neuronal activity in a human's brain, and as a written sentence on a piece of paper. Both instances of the proposition have semantic content (in both cases the proposition means something), but in the second instance, the written sentence only has meaning in so far as it is derived from some human's *intrinsic* intentional state. Without an attribution of meaning from a human agent, the marks on the paper that make up the sentence "the cat is rather fat" would be nothing but marks on paper. This is because they have no intrinsic or original meaning. The same type of marks in the same order could (in principle) have been produced by

two ants doing battle in the sand.[4] These marks clearly would be meaningless because the lines traced in the heat of battle were unintentional (in the common sense of the term as well as the philosophical sense). They were not "in themselves" representations of anything. The mental intentional state of the proposition "the cat is rather fat", on the other hand, is "in itself" meaningful since it does not depend on anything else for its meaningfulness or "aboutness". It is original because it is not derived.

For Searle, computers cannot have genuine intentionality or "aboutness" because computation is not intrinsically meaningful. The formal operations of a computer are like the marks on a piece of paper representing the proposition "the cat is rather fat": until they are interpreted by a human they are completely meaningless; they are not "in themselves", or intrinsically, about anything at all. According to Searle, computers only possess "aboutness" by virtue of the fact that we assign meaning to the symbol manipulations they perform. As he puts it: "Such intentionality as computers appear to have is solely in the minds of those who use them, those who send in the input and those who interpret the output" (1980, 83).

It would appear, then, that intrinsic intentionality rules out the notion of semantic transparency. Interpreting the formal token manipulations of an automatic formal machine does not endow the machine with "aboutness" or intentionality because the formal operations are not "in themselves" -- they are not intrinsically -- meaningful. The fact that a computer performs systematically in relation to a projected semantic content does not make the

[4] This example is in the vein of Hilary Putnam's Winston Churchill-depicting ant (Putnam 1981, 1).

formal operations of the computer *intrinsically about* the semantic content. Think back to the example in chapter 1 of the digital computer which produces the token 01 if and only if the atomic tokens 0 and 1 are simultaneously present in its system. The token 01 was interpreted as the proposition "John loves Mary and Mary loves John". But it could have been just as easily interpreted as "the cat is on the mat and the cat is rather fat" or any other simple conjunctive proposition. Now if such a syntactical operation can be interpreted in a variety of ways, how can it have intrinsic meaning? How can it be "in itself" about John and Mary rather than cats and mats? In short, how can computers possess intrinsic intentionality when their formal operations can mean different things depending on how they are interpreted? If, as Searle claims, intentionality or "aboutness" is necessarily intrinsic, then it would appear that computers cannot have intentional states, since the syntactical token manipulations they perform are not intrinsically about one thing rather than another. Now, if semantics is necessarily linked to intentionality, and if intentionality is necessarily intrinsic, there is no way, according to Searle, in which the syntactical operations of a computer can produce semantics because there is no way these operations can be seen as intrinsic. So if we agree with Searle that computers cannot have intrinsic intentionality then we must also agree with him that syntax is neither the same as, nor sufficient for, semantics.

Let's quickly retrace the path we have taken to get to this point: the Chinese room thought experiment is meant to show that computers can't think. According to Searle, the moral of his *Gedankenexperiment* rests on the "logical truth" that syntax is not the same as, nor sufficient for, semantics. But as we have seen, this claim about the underivability of semantics from syntax

presupposes that semantics is inextricably linked with *intrinsic* intentionality. It is intrinsic intentionality, then, that is at the heart of Searle's claim that syntax is insufficient for semantics. An interpreted automatic formal system (a computer) is not a thinking system because the syntactic operations it performs do not (even when interpreted) produce intrinsically intentional states. Indeed, the very fact that formal operations have to be interpreted to yield semantic meaning shows clearly, according to Searle, that computers have no intrinsic intentional states. If, as cognitivists claim, the token manipulations of a system have to be *assigned* a semantic interpretation, how could the manipulations be meaningful in themselves? Searle's notion of intrinsic intentionality has a great deal of intuitive appeal and at first glance it appears to deliver a knock-out blow to cognitivsm. But as we shall see, what initially seems obvious turns out upon closer inspection to be highly problematic.

According to Searle, a computer lacks intrinsic intentionality because the formal operations it performs are not meaningful "in themselves"; only when interpreted by a human with intrinsic intentionality do the syntactical operations become semantically significant. But how does Searle know this? The obvious answer is that formal operations can be interpreted in different ways. Syntactical operations have no intrinsic meaning because their meaning can change according to the way in which they are interpreted. But now what about cases where only one interpretation of a machine's formal token manipulations is possible? Can we still say with confidence that in such cases computers lack intrinsic intentionality?

5 Thermostats and Robots

Imagine an ordinary thermostat connected to a heat transducer and a boiler.[5] Such a thermostat could be described as a simple automatic formal system whose elementary token manipulations result in a boiler being turned on or off. For example, when the temperature reaches X degrees the thermostat executes token manipulation A which turns the boiler on, and when it reaches Y degrees it performs token manipulation B which turns the boiler off. Furthermore, the thermostat could be described as an *interpreted* automatic formal system if we assign to its token manipulations certain intentional states. We could say, for instance, that when it performs token manipulation B it has the intentional thought "the room is too hot". We could even fit the thermostat with a voice apparatus so it could utter the phrase "the room is too hot" whenever it carries out token manipulation B.

Now it seems somewhat ludicrous to assign intrinsic intentional states to such a simple device when the token manipulations it performs could be interpreted in any number of different ways. Given that the token manipulating system of the thermostat could be connected to different kinds of transducers, token manipulation B might be caused, among other things, by water in a tank exceeding a certain level, or by a train exceeding a certain speed. In other words, the same token manipulation that turns the boiler off and is interpreted as "the room is too hot" could just as easily be interpreted (given different transducers connected to different objects) as "the tank is too full" or "the train is going too fast". As Dennett points out, the thermostat system's "attachment to a heat-sensitive transducer and a boiler is too

[5] A transducer responds to physical energy patterns in the environment and transforms them into syntactic objects (see Dennett 1987, 141). The example of the thermostat is a modification of a thought experiment in Dennett 1987, 30.

impoverished a link to the world to grant any rich semantics to its belief-like states" (1987, 30). There is no sense in which we could interpret the formal operations of the thermostat as being *intrinsically about* the temperature of the room.

But now what if we were to add extra transducers to the thermostat so that it could gauge the temperature in other ways, thereby enriching its links to the world? For example, we might equip it with a complex multipurpose visual system that would be able to detect people shivering in the corner of the room or ice forming on the insides of the window panes. We might also give it an auditory system that would react to people complaining of the low temperature in the room. The addition of such sophisticated transducers would require vast complications of the inner structure of the automatic formal system (the thermostat) resulting in token manipulations with a much greater level of complexity. If, in addition to the extra transducers, we also gave the system the power to make inferences based on information from these transducers, it would be impossible to interpret the token manipulations performed by the system as about anything other than the temperature in the room. As Dennett puts it, somewhat more formally:

> the class of indistinguishably satisfactory models of the formal system embodied in its internal states gets smaller and smaller as we add such complexities: the more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which a unique semantic interpretation is practically (but never in principle) dictated. (1987, 30-31)

If abundant connections to the environment *dictate* a unique semantic interpretation of the system's syntactic operations, can Searle still claim that

such operations lack intrinsic or original intentionality? In other words, if the token manipulations can only *mean one thing*, does it make any sense to say that such a meaning is not intrinsic to the system? Searle could, I suppose, cling to the idea that any intentionality computers have is "solely in the minds of those who use them"; he could maintain that a unique semantic interpretation does not change the fact that it is we humans who assign semantics to a system and as a result no system (apart from we the assignors and a few other mammal species), no matter how complex, could have intrinsic intentionality. But how does this follow? Why should we assume, just because it is humans who assign a semantic interpretation to a system that the system itself lacks intrinsic intentionality? Could we not just be interpreting what is already intrinsic to the system? How, if there is only one coherent semantic interpretation of a system's operations, could we possibly know that the system has no intrinsic intentionality just because *we* assign meaning to its activities?

The widely-held view that a computer's intentionality "is solely in the minds of those who use them" can easily be made to look suspect, especially if we agree with Pylyshyn that

> the question of whether the semantic interpretation resides in the head of the theorist or in the [system] itself is the wrong question to ask. A better question is ... What latitude does the theorist have in assigning a semantic interpretation to the states of the system? (1986, 43-44)

Imagine an interpreted automatic formal system housed in a mobile and versatile robotic body capable of the full range of human activity -- e.g., writing poetry, acting heroically, solving complex problems, hitting a Randy

Johnson fastball, etc.[6] It goes without saying that any such system capable of making its way in the world would have an internal structure far more complicated than even the most impressive thermostat and as a result would also possess a unique semantic interpretation. Imagine further that the world is populated with a few billion of these systems and that the human race has long since disappeared from the face of the earth. These systems carry on in much the same way that their human predecessors did before them. They organise politically, send space shuttles into orbit, and attend professional sporting events. Now, could Searle claim that in the absence of interpreting human beings the activities of these robots would have no original or intrinsic meaning? Could he claim that the poetry these robots write, the jokes they tell and the political machinations in which they indulge are completely meaningless without a human mind to interpret them? Such a claim seems difficult to take seriously, to say the least. And yet, as we shall see, Searle thinks such automated formal systems could have no more intrinsic intentionality than a pocket calculator -- i.e., none at all.

Before going any further I should point out that the robots depicted above are, as things stand today, not even a remote empirical possibility; they are pure science fiction. Still, they cannot be ruled out on any a priori grounds and, as Searle says, "many of the most important thought experiments in philosophy and science are precisely science fiction" (1992, 70). In any case, whether such systems are empirically feasible is not important since Searle argues that even if they did exist they would not have intrinsic intentionality. In "Minds, Brains and Programs", Searle claims in answer to the "robot reply" that a computer inside a robot capable of "perceiving, walking, moving about,

[6] My description of such a system is taken from Haugeland 1985, 122.

hammering nails, eating, drinking -- *anything you like*" would not have intentionality (1980, 76; italics mine). Why not? Because, according to Searle, the Chinese room thought experiment applies to science fiction robots as much as it does to a language "understanding" computer. Searle asks us to imagine that he is inside the head of one of these robots manipulating meaningless symbols according to rules written in English. By effecting these formal manipulations Searle says he is

> receiving 'information' from the robot's 'perceptual' apparatus, and ... giving out 'instructions' to its motor apparatus without knowing either of these facts. I am the robot's homunculus, but unlike the traditional homunculus, I don't know what's going on. I don't understand anything except the rules for symbol manipulation ... I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. (1980, 77)

For Searle, computers that create poetry, hit fastballs, phone psychic hotlines and write essays on whether humans can think could all be Chinese rooms. As far as he is concerned, systems capable of emulating human behaviour are (like their semantically impoverished cousins, the thermostats and calculators) incapable of intentionality for the simple reason --the "simple logical truth" -- that *syntax is not sufficient for semantics*!

But wait a minute! How does Searle know that the syntactic operations of a system with a unique semantic interpretation are not sufficient to produce semantics? To put it another way, how is he able to tell that the "aboutness" of such a system (the science-fiction super-robot) is not original or intrinsic. It is one thing to dismiss the formal operations of a simple hand-held calculator; as Haugeland says "it would be tough to maintain that this crude

and limited device manages to refer to numbers all by itself -- quite apart from how *we* use it" (1985, 122). In such a case it is clear that syntax is insufficient for semantics (if with Searle we take intrinsic intentionality to be a necessary condition of semantics). But it is not at all clear that a complex system whose formal operations admit of only a single interpretation would lack intrinsic intentionality. How can we say with confidence that the "aboutness" of the science-fiction robots described above is not intrinsic or original? As we said earlier, it won't do to say that "such intentionality as computers appear to have is solely in the minds of those who use them" since there is no way to prove this. So why does Searle refuse to grant that robots have intrinsic intentionality? The reason, as we shall see in the next chapter, has nothing to do with the relationship between syntax and semantics; it has to do, rather, with the *consciousness* of semantics. For Searle, only conscious systems could possibly possess "aboutness", and no automatic formal system, on Searle's scheme of things, could ever be conscious.

In this chapter I have tried to show that Searle's Chinese room thought experiment is not sufficient in itself to show that computers cannot think. The *Gedankenexperiment* is meant only to bring to light what Searle considers to be a logical truth, namely that syntax is not sufficient for semantics. But as we have seen, there is no reason to believe that syntax is insufficient for semantics, even if, as Searle insists, semantics is inextricably linked with intrinsic intentionality, since the phenomenon of formal systems with unique semantic interpretations shows that computers can have intrinsic intentionality. I'd like now to take a detailed look at Searle's account of consciousness which, as we shall see, is the key to understanding his claim that digital computers can't think.

Chapter 3

Consciousness

I think that most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position.

-- Alan Turing, "Computing Machinery and Intelligence"

1 Introduction

In this chapter, we shall see how Searle's claim that syntax is not sufficient for semantics relies on his belief that subjective conscious states are necessary for intentional states. On this scheme of things, computers cannot have intentional states (and therefore cannot think) because they lack subjective conscious states. Searle's claim that thinking necessarily involves subjective conscious states results from his "first-person" approach to the study of the mind. On this approach, the critical feature of cognition -- consciousness -- is an ontological fact that cannot be limned by the objective, third-person methods used by the natural sciences. In what follows, I argue that this first-person approach to the mind, an approach which Searle uses to support his claim that computers cannot be conscious, in fact *precludes* him from making such a claim. If, as Searle would have it, conscious states are only accessible from a first-person point of view, then there is no way for him to tell which systems (apart from himself) have such states. And if he can't show that computers are *not* conscious, then he has no warrant for his claim that

computing machines lack intentionality, nor ultimately for his claim that syntax is insufficient for semantics. I end the chapter with a look at the Turing test and the role it plays in the cognitivist theory of cognition, and begin now with Searle's account of consciousness.

2 Consciousness

Searle describes consciousness as the "central mental phenomenon" (1992, xii), and claims that any study of the mind or cognition must include a study of consciousness since "we really have no notion of the mental apart from our notion of consciousness" (1992, 18). According to Searle, mental notions like intentionality, cognition, intelligence, etc. can only be properly and fully understood in relation to consciousness. There are two central features to Searle's account of consciousness: the first is that it is a biological phenomenon caused by neuronal processes in the brain, and the second is that this biological phenomenon -- unlike all other biological phenomena -- is ineliminably subjective. Before discussing these two features, let's take a look at what Searle thinks this biological, ineliminably subjective phenomenon actually is.[1]

Searle defines consciousness by giving examples of what it means to be in a conscious state. For instance, someone enters a conscious state when she awakens from a dreamless sleep and remains in a conscious state for as long as she is awake. If she falls into a dreamless sleep or dies, then she is no longer conscious. While dreaming, she experiences conscious states but ones

[1] Searle's account of consciousness is found in ch. 4 of *The Rediscovery of Mind*.

which are less intense and vivid than waking conscious states. Although this person may exhibit varying degrees of consciousness while awake -- she may fluctuate between being alert and drowsy -- she will always be either conscious or non-conscious. Consciousness, says Searle, is like an on/off switch. A word that comes close to describing consciousness, he says, is "awareness." Being in a conscious state is being in a state of awareness. There is a problem, however, with using the two words synonymously, since, according to Searle, "awareness" is more closely connected to cognition than is "consciousness". Also, it is possible to be unconsciously aware of something, as shown by the phenomenon of blindsight, a visual disorder in which patients are able to report on visual information that they do not consciously see. Obviously, consciousness, unlike awareness, can never be unconscious. Finally, conscious states must always have content. Whenever anyone is conscious he or she must be conscious of something. According to Searle, not all content, however, is the content of intentional states. Pains are an example of conscious states with a content that isn't intentional content, since the states don't refer to anything beyond themselves.

So much for Searle's description of consciousness. Let's move on now to what he sees as the two central features of consciousness -- its biological nature and its subjective nature. I'll begin by looking at Searle's claim that consciousness is a biological phenomenon.

3 Matter's What Matters

One of Searle's central objectives in *The Rediscovery of the Mind* is to situate consciousness within the scientific world view. He wants to show how it is

possible to acknowledge the reality of consciousness without resorting to dualistic theories of mind that split the world into material things measurable by science on the one hand, and elusive, non-quantifiable mind stuff on the other. Searle argues that a belief in the reality of consciousness is perfectly consistent with the facts of both the atomic theory of matter and the evolutionary theory of biology, the two theories he feels are fundamental to the scientific world view. In the following thumbnail sketch he shows how consciousness fits into the material world described by physicists and biologists:

> According to the atomic theory, the world is made up of particles. These particles are organised into systems. Some of these systems are living, and these types of living systems have evolved over long periods of time. Among these, some have evolved brains that are capable of causing and sustaining consciousness. Consciousness is, thus, a biological feature of certain organisms in exactly the same sense of "biological" in which photosynthesis, mitosis, digestion, and reproduction are biological features of organisms. (1992, 93)

According to Searle, consciousness, like photosynthesis, mitosis and other like phenomena, is a causally emergent system feature. To see what is meant by this, imagine a system S made up of elements X, Y, Z, say, a small piece of wood, made up of molecules. A system feature of the piece of wood is that it weighs two pounds. What makes weight a system feature is that it is something not necessarily shared by the molecules out of which the wood is made; the wood weighs two pounds but the individual constituent elements need not. Now some system features -- weight, velocity, shape -- can be figured out from the way the constituent elements are arranged and composed. Other system features, however, can only be discerned or explained in terms of the causal interactions that take place among the

constituent elements. Take liquidity, for instance. Water has the system feature of liquidity and this feature cannot be accounted for without explaining how the elements that comprise water (hydrogen and oxygen molecules) causally interact. To explain liquidity it is not enough to know what molecules make up water, we must also know how these molecules react with each other to create the system feature. So liquidity is a causally emergent system feature. According to Searle, consciousness is a causally emergent system feature in the same way that liquidity is. We can explain the existence of consciousness from the causal interactions involved in neuronal activity, but not from the mere physical structure and arrangement of the neurons themselves. Finally, although Searle claims that consciousness is caused by brains, he points out that nobody knows how conscious states emerge from neuronal activity. The neurophysiology of consciousness is at present still a mystery (1992, 91).

To this point I have explained Searle's claim that consciousness is a biological phenomenon caused by the brain. Now I would like to look at subjectivity, the other central aspect of his account of consciousness.

4 The Subjective Ontology of Consciousness

Consciousness, says Searle, is the only known natural phenomenon that possesses the feature of subjectivity. To become clear on what exactly Searle means by subjectivity we must first sort through some important distinctions he makes with regard to the subjective/objective dichotomy. According to Searle, what the subject/object split means on an epistemological level is completely different from what it means on an ontological level.

Epistemologically, the subject/object dichotomy refers to the degree to which claims can be said to be independent of personal vagaries, biases, emotion etc. The more objective a claim is, the less it is thought to rely on a particular point of view or a particular set of circumstances. Ontologically, the split refers to "different categories of empirical reality" (1992, 19). Some empirical facts about the world are ontologically objective facts whereas others are ontologically subjective.

According to Searle, consciousness is an empirical fact -- a real fact about the world -- that is ontologically subjective. In other words, the being of consciousness is necessarily subjective. To show what he means by the ontological subjectivity of consciousness, he asks us to consider the phrase "I now have a pain in my lower back" (1992, 94). This sentence, says Searle, is made true by the existence of an ontologically subjective fact, namely, the phenomenon of pain. The pain has a subjective mode of existence because it can only be felt by whoever utters the phrase "I now have a pain in my lower back". It is a fact about the world immediately accessible only from the first-person point of view; the pain must be a pain for somebody. Now according to Searle, consciousness is subjective in the same sense that pain is. In other words, it is a first-person phenomenon. The knowledge of what it is like for someone to be in a particular conscious state is only completely accessible to the person who is actually in that state. Searle takes great pains to point out that an empirical fact that is ontologically subjective is not therefore epistemologically subjective. The pain you experience is a real empirical fact about the world, as real as the intersubjective facts science discovers through an objective third-person approach to physical phenomena. It differs from ontologically objective facts only in so far as it is not equally accessible to

standard, objective, third-person tests (1992, 73). For Searle, what he considers the brute reality of a fact such as pain or consciousness is not impugned by its inaccessibility to the third-person methods of science.

5 What It's Like to Be Inside the Chinese Room

So far I have outlined Searle's description of consciousness as a biological phenomenon that is ontologically subjective. Now I'd like to try to show how his ideas about consciousness play a crucial role in the claim that syntax is insufficient for semantics (the key premise in his argument against the computational theory of mind). In the last chapter we saw that, for Searle, thinking or understanding requires intrinsic intentionality (original as opposed to derived "aboutness"). On this scheme of things, no computer could think because purely formal operations are not sufficient for genuine intentionality. The symbol manipulations of automatic formal systems do not have real "aboutness"; rather, they are "about" a semantic content only in so far as they are interpreted by humans with intrinsic intentionality. But we saw also that the claim "such intentionality as computers appear to have is solely in the minds of those who use them" made no sense with respect to sufficiently complex formal systems (like the super-robots) which possess unique semantic interpretations. If a system cannot be coherently assigned more than one semantic interpretation then there is no way we can be sure that the system lacks intrinsic intentionality. What would it mean, after all, to deny intrinsic intentionality to a robot which is behaviourally identical to a human? On what grounds could Searle claim that one of these super-robots lacks intrinsically intentional states? The answer he gives (and here we get to the crux of his argument against cognitivism and AI), is that such systems

can't think because "[o]nly a being that could have *conscious* intentional states could have intentional states at all" (1992, 132; my italics). Computers can't think, according to Searle, because they are not conscious beings! Now if, as Searle maintains, consciousness is essentially subjective, and if intentional states are necessarily conscious, then intentionality must be subjective. This is what Searle calls the aspectual nature of intentional states, the "what it feels like aspect of consciousness" that accompanies intentionality (1992, 132). It is this subjective, first-person "what it feels like aspect" which does all the work in the Chinese room thought experiment. Searle inside the Chinese room understands the symbols that make up the English story and questions but not those that make up the Chinese story, not because of the "simple logical truth" that syntax is insufficient for semantics but rather because he has a conscious, first-person experience of the English story but not of the Chinese story. As Dennett points out:

> Searle has apparently confused a claim about the underivability of *semantics* from syntax with a claim about the underivability of *the consciousness of semantics* from syntax. For Searle, the idea of genuine understanding, genuine "semanticity" as he often calls it, is inextricable from the idea of consciousness. He does not so much as consider the possibility of unconscious semanticity. (1987, 335; original italics.)

Given his view that only systems with conscious states could be systems with intentional states it comes as no surprise that Searle doesn't even consider the possibility of unconscious semanticity. For Searle, a system that lacks consciousness could not be an intentional system, since real intentionality or genuine semanticity (as opposed to the "derived" intentionality of semantic transparency) requires the first-person "what it feels like" aspect that only conscious systems possess. On this scheme of things, syntax can never be

sufficient for semantics because purely formal operations lack consciousness. The Chinese room argument, then, has nothing to do with whether a machine's formal operations are meaningful (the claim about syntax and semantics) and everything to do with whether a machine could have subjective, first-person conscious states about the formal operations it performs. For Searle the question of whether computers can think is really a question of whether they can be conscious, since, on his scheme of things, only systems with subjective conscious states could be systems with intentional states. To sum up, we can state Searle's position against cognitivism in the following way:

1) No system without subjective conscious states could be a thinking system since thinking requires intentionality and only systems with subjective conscious states can have intentional states.

2) No interpreted automatic formal system (i.e., no system defined by its formal operations) could have subjective conscious states.

Therefore

3) No interpreted automatic formal system could be a thinking system.

I see no reason why Searle would disagree with this formulation. It follows from his views on the subjective nature of consciousness and his critique of cognitive reason outlined in "Minds, Brains, and Programs" and *The Rediscovery of the Mind*.[2] It is this line of reasoning that gives warrant to Searle's claim that syntax is insufficient for semantics; the underivability of semantics from syntax only holds if intentional states must be subjective conscious states. What I would like to do now is challenge the soundness of the above argument by calling into question the second premise.

[2] See especially 1992, 44, 200.

6 Conscious Computers

Searle tells us that we can't seriously entertain the idea that computers are conscious (1992, 21). This claim can be explained in part by his view that consciousness is a biological phenomenon caused by an organic brain. Since computers do not have organic brains and since organic brains are the cause of consciousness, it stands to reason that computers could not be conscious. But this appeal to the biological nature of consciousness is not sufficient to guarantee the acceptability of premise 2 in the above argument, since Searle admits that "[f]or all we know at present, there might be no theoretical obstacle to developing consciousness in systems made up of other elements" (1992, 91). If no such theoretical obstacle exists -- if, as Searle says, it is an empirical question whether any matter apart from brain stuff can cause intentional states (1980, 82) -- then how can he be so sure that silicon computing machines could not be conscious?

Imagine that AI researchers together with neuroscientists construct a robot that is morphobehaviourally identical to a human. Imagine further that the robot's brain is a serial digital computer made of silicon (an interpreted automatic formal system) which duplicates, at the program level, the parallel architecture of a human brain.[3] What we would have in this case is a physical system that looks and behaves like a human system and that implements the same formal program implemented by a human brain. Would we say that

[3] As Dennett points out (1991, 215), programing a serial digital computer to duplicate the activities of a parallel distributed processing system such as the brain is possible in principle given Turing's thesis that a Universal Turing machine can perform any function that any computer, regardless of architecture, can perform.

this robot has a brain that produces conscious states? If, as Searle contends, it is an empirical question whether silicon could produce consciousness, couldn't we simply subject the robot to certain empirical tests and decide one way or the other whether it is conscious? For example, couldn't we just open up its computer "brain" and check for consciousness? At first blush this seems like it would be a fruitful empirical inquiry. But given that nobody knows how *human* brains produce conscious states (Searle 1992, 91), how would we be able to test for consciousness in silicon brains? If no amount of neurological probing can link neuronal activity to subjective conscious states, how would we know what kind of causal activity to look for in the silicon brain?

If an examination of the inner workings of the silicon brain won't yield a definitive answer, couldn't we attribute conscious states to the robot based on the fact that its behaviour is identical to a human's? If, for example, the robot tells us that it has conscious feelings of pain, and if, in addition, it behaves as if it were in pain, couldn't we conclude that the robot has conscious feelings of pain? This would seem a reasonable conclusion to draw given that this is the way that we humans usually interact with one another. A doctor who asks her patient "does it hurt when I press here?" usually assumes that the patient has a conscious sensation of pain when he responds affirmatively to her question. But according to Searle, *behaviour can tell us nothing about consciousness*. To prove his point he asks us to imagine a case where a patient has his biological brain replaced with a silicon one (1992, 66). While the patient's behaviour is unaffected by his new brain, he discovers, to his horror, that he no longer has any *conscious* vision. When doctors show him a red object he can't see anything at all but hears his own voice "in a way that

is completely out of control" (1992, 67) assuring doctors that he sees the red object. What this shows, says Searle, is "that as far as the ontology of consciousness is concerned, behavior is simply irrelevant. We could have identical behaviour in two different systems, one of which is conscious and the other totally unconscious" (1992, 71).

If we can't discern whether a human-like robot is conscious by examining its behaviour (including its verbal reports) or by investigating the inner workings of its silicon brain, then what could Searle mean when he says that it is an empirical question whether conscious states could be caused by silicon? He doesn't give us any clues but he does claim that "it is empirically absurd to suppose that we could duplicate the causal powers of neurons entirely in silicon" (1992, 66). But what makes such a supposition absurd? Empirically speaking, it is no less absurd than the supposition that *organic* brains cause subjective conscious states. After all, if neurophysiology can't tell us what makes humans conscious, and if behaviour is irrelevant to consciousness, then on what empirical grounds can Searle claim that organic brains cause the subjective conscious states necessary for intentionality? As Dennett points out:

> Perhaps only some organic brains produce intentionality! Perhaps left-handers brains, for instance, only mimic the control powers of brains that produce genuine intentionality. Asking the lefthanders if they have minds is no help, of course, since their brains may just be Chinese rooms. (1987, 334)

If scientific investigation can't reveal the difference between someone who is conscious and someone (or thing) who, like the blind patient, only appears to be conscious, then how can the question of what has consciousness and what

doesn't have consciousness be an empirical one? As Dennett says, "[s]urely it is a strange kind of empirical question that is systematically bereft of all intersubjective empirical evidence" (1987, 334).

For Searle, though, an empirical question need not be intersubjectively verifiable, and herein lies the key to his views on what kinds of things can be conscious. There are, he says, "lots of empirical facts that are not equally accessible to all observers" (1992, 72). One of these empirical facts is consciousness. We will remember that, according to Searle, consciousness is an empirical fact that is immediately accessible only from the first-person point of view. It is a real fact about the world -- a fact as real as any discoverable by science -- but one which, because of its subjective ontology, cannot be limned by standard, intersubjective, third-person tests. Subjective conscious states, the "what it feels like" aspect of phenomenal experience, can only be known by the person who experiences them. As Searle says, "every conscious state is always *someone's* conscious state" (1995, 95; original italics). Now, when Searle says that it is an empirical question as to what systems could produce subjective conscious states, he has in mind the kind of empirical question that cannot be approached intersubjectively, the kind that can only be answered from a first-person point of view. Searle's claim that biological systems cause subjective conscious states is based ultimately on an empirical fact that only he has access to, namely, the fact of his subjective conscious states. Searle is aware of *his own* subjective conscious states and concludes that they must be caused by his brain since all scientific evidence tells us that all our mental states are caused by neuronal activity.

But now what, on this scheme of things, gives Searle warrant for the claim that it is empirically absurd to suppose that silicon could produce subjective consciousness? If consciousness is a fact that can only be known from the first-person point of view, how can Searle know whether or not a robot with a silicon brain enjoys subjective conscious states? For that matter, how can he know that anything in the universe apart from himself has subjective conscious states? If, as he says, the ontology of consciousness is essentially a first-person ontology (1992, 20), then how can we find out about the subjective consciousness of other systems? Searle argues that we can know that other systems are conscious through the *causal basis* of their behaviour. If a system behaves like a conscious system, and if such behaviour is caused by similar physiology, then we know that the system is conscious. For example, Searle tells us he is "completely convinced" that his dog is conscious because it behaves in a conscious manner and has "the appropriate causation in the underlying physiology" (1992, 73). So though Searle thinks that behaviour alone can tell us nothing about the ontology of consciousness, he believes that behaviour caused by the right kind of stuff is sufficient to inform us about conscious states of other systems. He can claim that other humans, and some higher animals (including his pet dog) have conscious states because of a) the "empirical" fact of his own subjective consciousness and b) the similarities between his physiology and the physiology of other conscious beings. At the same time he can claim that robots with silicon brains cannot be conscious because, although they may behave as if they were conscious, they don't have the right underlying physiology.

But now how can a similarity in physiology between two systems be grounds for attributing subjective conscious states? If consciousness is ineliminably

subjective, if it is only known from a first-person point of view, then no amount of third-person, scientific evidence pointing to physiological similarities between two systems can give grounds for attributing a mental state (consciousness) which, because of its subjective ontology, cannot be accounted for by science. No amount of physiological data gathered through third-person, intersubjective methods will shed any light on facts which are ontologically subjective. Searle's dog may, from a third-person point of view, appear to have the appropriate physiological makeup for consciousness but perhaps it is only a furry zombie. If the dog yelps after having its paw stepped on Searle may want to conclude that the dog is in pain, but how would he know this as long as he holds that "no description of the third-person, objective, physiological facts would convey the subjective, first-person character of the pain, simply because the first-person features are different from the third-person features" (1992, 117)? The point being made here is that Searle's view of consciousness as an ontologically subjective phenomenon leaves him in a position where he can be certain only of his own consciousness. As a result, he can't claim that silicon computers necessarily lack conscious states. In this case, we need not accept premise 2 of his argument against cognitivism nor, ultimately, his claim that computers can't think.

To sum up, Searle's argument against cognitivism (his claim that syntax is insufficient for semantics) only holds if subjective conscious states are seen as a necessary condition for intentional states. If intentionality is a necessary condition for thinking and if subjective conscious states are a necessary condition for intentionality, then computers can't think because they don't have subjective conscious states. In other words, Searle's claim that

computers can't think relies on a first-person approach to the mind. On this approach, what is essential to cognition are subjective conscious states which are inaccessible to intersubjective empirical investigation. But the fact that such states can't be limned through the third-person methods of science make it impossible to make any claims about which systems do, and which systems do not, possess subjective conscious states. Since there is no way to tell whether or not computers have subjective conscious states Searle cannot preclude them from the class of thinking systems on the grounds that they *don't* have such states.

According to Searle, one of his primary goals in writing *The Rediscovery of the Mind* was to formulate a theory of consciousness that avoids the pitfalls of Cartesian dualism; he wanted to situate consciousness squarely in the world of atoms and evolutionary theory. Given this goal, it seems odd that he would include in his ontology phenomena such as subjective conscious states that cannot be accounted for by the standard methodological procedures used in the natural sciences. But Searle claims that this standard, third-person approach, while it works fine for every other scientific investigation, doesn't work for the mind because it fails to capture the "what it feels like" aspect of our "inner lives", an aspect that he believes is no less real just because it is not equally accessible to any observer. According to Searle, cognitivism (as well as AI and most contemporary philosophy of mind) is hopelessly misguided because its third-person methodology leaves out what is most important to cognition (1992, 95). But why should cognitivism, a well-defined empirical research program, countenance the existence of a phenomenon which cannot be limned by the empirical, objective methods of the physical sciences? Put another way, why should a research program committed to the

conviction that something real must be equally accessible to all competent observers accept "facts" whose ontology is subjective -- i.e., "facts" that can only be known from a first-person, private point of view Does it even make sense to talk about facts that aren't inter-subjectively verifiable?

A detailed examination of the merits and shortcomings of the first- and third-person approaches to the mind is beyond the scope of this paper, but I would like, in closing, to examine what Searle feels is one of the disastrous effects of the third-person stance -- the Turing test. Searle believes that the Turing test is a symptom of a theoretical stance which demands that whatever is objective must be equally observable to any observer. Such an approach, says Searle, shifts the questions about the mind away from subjective states and towards external behaviour. This leads in turn, he argues, to a situation in which we can no longer tell the difference between what really has a mind (i.e., humans) and what merely appears (from the outside) to have a mind (i.e., a computer).

Searle claims that the use of the Turing test by cognitivists reflects a kind of radical behaviouristic principle in their program, a principle which would force us "to conclude that radios are conscious because they exhibit intelligent verbal behaviour" (1992, 22). It is this characterisation of cognitivism which is responsible, I believe, for any success that Searle may have had in convincing people that computers cannot think. But, as we shall see, it is a characterisation that misrepresents the cognitivist position. Let's turn now to an examination of the Turing test.

7 The Turing Test

By 1950 Alan Turing and others had designed and built machines that they believed exhibited genuine intelligence. To address questions about whether or not these computers really were intelligent or whether they were just sophisticated electrical crank-shafts, he wrote "Computing Machinery and Intelligence" (1950). Turing begins the paper by stating that he will take up the question "Can machines think?" by defining the words 'machine' and 'think' in such a way that they would account for normal usage. He immediately abandons this definitional approach to the question, however, fearing that it will lead to endless and futile disputes over the meanings of words. Instead he proposes that the question "Can machines think?" be decided by the "imitation game". The "imitation game" is played by three persons, A, B, and C. A is male, B is female and C is either male or female. C is in a separate room from A and B, and communicates with them by teleprinter. The object of the game is for C to try to figure out which of the other two participants is male and which is female. This she does by asking A and B questions. For example, she might ask A what it is like to give birth to a child. A's role in the game is to try to trick C into making the wrong identification. So he might answer with something like "giving birth, although painful, was, without a doubt, the most joyous occasion in my life." The role of B is to try to help C make the correct identification by answering truthfully to all questions. B might also assist C by saying such things as "Don't believe the answer A gave you. It was trite and belied someone who obviously has never experienced childbirth." Naturally A can say the same kinds of thing in an effort to lead C astray.

Turing then asks the following questions: 1) "What will happen when a machine takes the part of A in this game?" and 2) "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" These two questions, says Turing, replace the question, "Can a machine think?" In other words, if a machine can fool an interrogator as often as a human can, then a machine can think.

A little reflection should be enough to realise that fooling an interrogator (providing she is sufficiently clever) would be no easy task, and could be accomplished only by an extremely sophisticated machine. This is because the question and answer format of the imitation game requires that player A be able to converse (intelligently, if it hopes to win the game) about any topic whatsoever, as is made clear by Turing's example of a possible exchange between C and A (1950, 53):

> Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?
>
> Witness: It wouldn't scan.
>
> Interrogator: How about a "a winter's day"? That would scan all right.
>
> Witness: Yes, but nobody wants to be compared to a winter's day.
>
> Interrogator: Would you say Mr. Pickwick reminded you of Christmas?
>
> Witness: In a way.
>
> Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.
>
> Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

In the course of this exchange, player A has displayed fluency in the English language, a pretty good understanding of poetry, the seasons, the way seasons are used metaphorically to describe people's character, the way people might feel about being ascribed such characteristics, etc. What makes the "imitation game" so compelling is that it is difficult to think of another test that could put such a strain on the cognitivist claim that computers can think.[4] The questions asked by the interrogator can be about anything, so a computer that successfully plays the game would have to able to "converse" intelligently about any topic whatsoever. Now how, one wonders, does Searle believe a radio could think given the demands of the "imitation game"?

The motivation behind Turing's "imitation game" (which in subsequent literature on the subject has become known as the Turing test) was to shift the focus away from questions like "what kind of things are capable of thinking?" to questions like "what would we all agree is an example of thinking regardless of its origin?" This approach eliminates any role prejudice might play in the attribution of intelligence to a mechanical device. Daniel Dennett (1985) likens the Turing test to the practice of symphony orchestras conducting auditions in which the jury is separated from the musician by an opaque screen. In this manner, candidates are chosen solely for their musical ability and not for attributes like gender or physical appearance which might have an effect (even if unconscious) on the jurors' decision-making process.

---

[4] As of 1985, no digital computer had even come close to passing the Turing test. Dennett, at the time he wrote "Can Machines Think?" (1985), guessed that no computer would pass the test for at least 25 years. Interestingly enough, he believes that only a computer with rich perceptual connections to the world (a super-robot of sorts) could have the necessary "life" experience to past the test.

This makes the Turing test an operationalist account of intelligence. Operationalism can be described as the view that if a difference can't be discovered, then there is no difference; or to put it another way, "If it quacks like a duck and walks like a duck, it is a duck" (Dennett 1991, 117). If the interrogator in the 'imitation game' can't tell the difference between the thinking performed by a human and that performed by a computer, then, as far as Turing is concerned, there is no difference to be found. This is not to say, of course, that Turing thinks computers are humans; the duck in question in the Turing test is intelligence, not personhood. As Turing states, "[the 'imitation game'] ... has the advantage of drawing a fairly sharp line between the physical and intellectual capabilities of man" (1950, 41). In other words, machines and people are the same only in so far as they are thinking things.

Having now said something about what the Turing test is, I would like to say something about what it is not. It is not -- contrary to what some commentators seem to believe (e.g. Moody 1994) -- the crux of the cognitivist theory of cognition. As we saw in the first chapter, cognitivism makes the claim that a thinking system is a system that formally manipulates symbols. But something that successfully passes the Turing test need not be a symbol-manipulating system. As Margaret Boden puts it "nuts falling from a wind-blown tree onto the keys of a teletype could conceivably 'fool' a human interrogator playing the imitation game" (1990, 5). Although this may sound nutty, it makes an important point: there is more to the cognitivist program than building machines that duplicate human verbal behaviour. As Pylyshyn points out:

> Even if it were true that a Turing machine can behave in a manner indistinguishable from that of a human ... producing such behavioural mimicry is a far different matter from providing an explanation of that behaviour ... explanation entails capturing the underlying generalisations as perspicuously as possible and relating them to certain universal principles. To do that ... it is necessary to appeal to, among other things, semantically interpreted representations. (1986, 54)

Searle, however, often talks about the cognitivist theory of cognition -- what he calls Strong AI -- as if it were centred on a behaviourist or operationalist ethic. He claims, for instance that "if AI workers totally repudiated behaviourism and operationalism much of the confusion between simulation and duplication would be eliminated" (1980, 86). In other words, he thinks that if cognitivists were not blinded by the dictates of these "isms" -- if they didn't think that behaving intelligently was the same as having genuine intelligence -- they would realise that duplicating the cognitive capacities of humans with symbol-processing systems is impossible, and that the best the computational theory of mind can hope for is simulation of intelligence. This type of criticism misrepresents the cognitivist stance by over-simplifying it. The reason cognitivists try to build intelligent computers is because they are interested in explaining how systems think. What Searle seems to forget is that cognitivists are convinced that there already exist billions of computers that think, namely, we humans. The goal of cognitivist research is not to build any old system that mimics human behaviour, but rather to provide an explanation of human cognitive processes in part by building machines that, like us, are interpreted automatic formal systems. Strictly speaking, cognitivists do not endorse the view of "same-behaviour-ergo-same-mental-phenomena" which Searle believes is the mistake enshrined in the Turing test and which he believes leads to the view (one

that cognitivists would not entertain for a moment) that radios exhibit intelligence (1992, 22).

But now why, if Turing believed that thinking was necessarily computational, did he devise a test that allows for the possibility of non-computational thought? Dennett suggests that Turing designed the "imitation game" solely in hopes of ending pointless quibbling over the "true nature" of thinking:

> Turing didn't design the test as a useful tool in scientific psychology, a method of confirming or disconfirming scientific theories or evaluating particular models of mental function; he designed it to be nothing more than a philosophical conversation-stopper ... a simple test for thinking that was surely strong enough to satisfy the sternest sceptic (or so he thought). He was saying, in effect, "Instead of arguing interminably about the ultimate nature and essence of thinking, why don't we all agree that whatever that nature is, anything that could pass this test would surely have it ..." (1985, 122)

In other words, just because Turing came up with the "imitation game" doesn't mean that he believed thinking was anything other than computation. He happened to think cognition had to do with manipulation of tokens; others thought at the time that it had to do with consciousness, others still with immaterial souls. What Turing hoped his operationalist definition of thinking would do would be to permit everyone to agree on what could count as intelligence regardless of how it was produced. This would let him get on with his project of designing intelligent machines.

As we can see then, the connection between cognitivism and operationalism or behaviourism is not as tight as Searle seems to suggest. The success of the

computational theory of mind depends as much on explaining cognition as it does on duplicating it. To put it another way, cognitivists do not believe that by duplicating intelligent behaviour they will have necessarily explained the nature of intelligence. Explaining intelligence has to do with semantic transparency, not behavioural mimicry. By exaggerating the stock cognitivists put in the Turing test, Searle creates a red herring of sorts. It is much easier to dismiss the claim that a radio thinks than it is to prove the shortcomings of the notion of semantic transparency, something which, as we have seen, Searle is only able to do by calling into question the role of the third-person, objective stance in the philosophy of mind.

In this thesis I have argued that Searle's critique of cognitivism is not based, as he says, on the simple logical truth that syntax is not sufficient for semantics, but rather on the belief that subjective conscious states are a necessary condition for thinking. Searle believes that we humans and some higher primates have such states and that silicon formal systems could not. But as we have seen, given the methodological approach Searle adopts to studying the mind (the first-person approach), he has no warrant for his conclusion that computers lack conscious states, and as a result cannot exclude them from the class of thinking systems.

Bibliography

Akins, K. A. (1990) "Science and Our Inner Lives: Birds of Prey, Bats, and the Common (Featherless) Biped," in M. Bekoff and D. Jamieson, eds., *Interpretation and Explanation in the Study of Animal Behavior*, vol. 1. Boulder, CO: Westview, pp. 414-427.

Bechtel, W. (1988) *Philosophy of Mind: An Overview for Cognitive Science.* Hillsdale: Lawrence Erlbaum Associates.

Bergmann, M., Moor, J., and Nelson, J. (1980) *The Logic Book.* New York: Random House.

Boden, M. A., ed. (1990) *The Philosophy of Artificial Intelligence.* New York: Oxford University Press.

Churchland, P. S. (1986) *Neurophilosophy.* Cambridge, MA: The MIT Press.

Clark, A. (1989) *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* Cambridge, MA: The MIT Press.

Dennett, D. C. (1985) "Can Machines Think?" in Michael Shafto, ed., *How We Know.* San Francisco: Harper & Row.

Dennett, D. C. (1987) *The Intentional Stance.* Cambridge, MA: The MIT Press.

Dennett, D. C. (1991) *Consciousness Explained.* Boston: Little, Brown and Company.

Dennett, D. C. (1993) "Living on the Edge," *Inquiry* 36: 150-153

Fodor, J. (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, MA: The MIT Press.

Haugeland, J. (1985) *Artificial Intelligence: The Very Idea.* Cambridge, MA: The MIT Press.

Moody, T. C. (1993) *Philosophy and Artificial Intelligence.* Englewood Cliffs: Prentice Hall.

Nagel, T. (1974) "What is it Like to Be a Bat?" *Philosophical Review* 4 LXXXIII: 435-450.

Nagel, T. (1986) *The View From Nowhere.* Oxford: Oxford University Press.

Newell, A., and Simon, H.A. (1976) "Computer Science as Empirical Enquiry: Symbols and Search," *Communications of the Association for Computing Machinery*, 19, 113-126.

Putnam, H. (1981) *Reason, Truth and History.* New York: Cambridge University Press.

Pylyshyn, Z. W. (1986) *Computation and Cognition: Toward a Foundation for Cognitive Science.* Cambridge, MA: The MIT Press.

Quine, W. V. O. (1980) *From a Logical Point of View.* Cambridge, MA: Harvard University Press.

Rorty, R. (1979) *Philosophy and the Mirror of Nature.* Princeton: Princeton University Press.

Searle, J. (1979) "What is an Intentional State?" *Mind* 88: 74-92.

Searle, J. (1980) "Minds, Brains, and Programs," *Behavorial and Brain Sciences* 3: 417-424. (Reprinted in Margaret Boden, ed., *The Philosophy of Artificial Intelligence.* New York: Oxford University Press, 1990, pp. 67-88.)

Searle, J. (1992) *The Rediscovery of the Mind.* Cambridge, MA: The MIT Press.

Siewert, C. (1993) "What Dennett Can't Imagine and Why," *Inquiry* 36: 93-112.

Teller, P. (1989) *A Modern Formal Logic Primer.* Englewood Cliffs: Prentice Hall.

Turing, A. (1950) "Computing Machinery and Intelligence," *Mind* 59: 433-460. (Reprinted in Margaret Boden, ed., *The Philosophy of Artificial Intelligence.* New York: Oxford University Press, 1990, pp. 40-66.)