# *Turing's Test, Searle's Chinese Room Argument, and Thinking Machines*

*Peter Jackson*

The Open Polytechnic Working Papers are a series of peer-reviewed academic and professional papers published in order to stimulate discussion and comment. Many Papers are works in progress and feedback is therefore welcomed.

This work may be cited as: Jackson, P. *Turing's Test, Searle's Chinese Room Argument, and Thinking Machines*, The Open Polytechnic of New Zealand, Working Paper, May 2005.

Further copies of this paper may be obtained from

The Secretary, Research Publications Committee
The Open Polytechnic of New Zealand
Private Bag 31 914
Lower Hutt
Email: WorkingPapers@openpolytechnic.ac.nz

This paper is also available on The Open Polytechnic of New Zealand website:
http://www.openpolytechnic.ac.nz/

Printed and published by The Open Polytechnic of New Zealand, Lower Hutt.



ISSN — 1174 – 4103

ISBN — 0 – 909009 – 78 – 3

Working Paper No: 2-05

A list of Working Papers previously published by The Open Polytechnic is included with this document.

# Dedication

The working paper is dedicated to the memory of Dr Alan Mathison Turing (1912–1954). Turing was one of the most brilliant minds of his era. As a mathematician and logician, he made a profound contribution to computer science and to the debate addressed in this working paper. To him we owe the key concepts of the Turing thesis, the universal Turing machine and the Turing test, all of which are essential to the debate on thinking machines.

Turing's life was tragically cut short by his own hand at the age of 41. The backwardness and bigotry of British law at that time led to his being branded as a criminal because of his homosexual inclinations. The Court's punishment of hormonal treatment ('to quell his lust') caused Turing to grow breasts and led him into depression and despair. He ended his short life by eating an apple that he had laced with cyanide.

His untimely death robbed us of a great mind. Who can say how far and how quickly computer science would have progressed had he lived out a more natural span of years?

**Alan Turing at Cambridge circa 1936**

# Abstract

This paper deals with the debate on artificial intelligence (AI) thinking machines. In particular, it asks the question, 'Do AI machines think as we humans do?' The main thrust of this paper is philosophical and does not directly deal with technological platforms for AI. After a brief history of AI, there follows a discussion on the work of Alan Turing, in particular that on his logical computing machine (LCM), his thesis (also Church's), his paper in *Mind*, covering the 'Imitation Game', and the Turing test, which arose out of it. Turing is seen as the founder of the strong AI hypothesis (machines can think). The work of John Searle is then covered as it relates to this debate. Under particular discussion are Searle's Chinese Room experiment (CRE) and the Chinese Room argument (CRA) that arose from it, in which he attempts to refute the strong AI viewpoint and provide support for his alternative weak AI hypothesis (machines cannot think). The consideration of Searle's work leads to a discussion of issues critical to Searle's view, that of syntax versus semantics, and of intentionality. After a comment on artificial neural networks (ANNs) as a potential technological platform for thinking machines, there follows a discussion on the relationship between AI, thinking and consciousness, in an attempt to clarify what is meant by these terms in relation to the debate addressed here. Finally, a summary is made and tentative conclusions are reached, in which the following views are offered:

- The strong AI position is invalid, at least for von Neumann-type machines. However, the weak AI position is valid in so far as such machines can, and currently do, emulate human thinking.

- While ANNs provide a potential technological platform for thinking machines, the technology is too nascent as yet.

- If truly thinking machines ever do become a reality, their existence will raise a number of challenges, such as our ethical responsibility toward them (as sentient entities) and the threat to us as a species that they might represent.


  

# Contents



  

# Introduction

The idea of a thinking machine has a long history, especially within the field of science fiction literature. Isaac Asimov's *I Robot* series of novels, written in the 1940s, was based on the notion of a robot's having powers of thought and very stringent ethical standards. In many ways, these robots were superior to the humans that they served. In 1967 Arthur C. Clarke wrote a novel, *2001: A Space Odyssey* (which was then produced as a movie of the same name). Although this story was not specifically about thinking machines, one of the key characters was the IBM, HAL 9000 machine aboard the spacecraft *Discovery,* which was heading out into Jupiter space. HAL had a personality and could think, and concluded that the human commander and co-commander were jeopardising the mission and attempted to kill them. Very recently, in the 1990s, David Brin has written a number of novels within his *Uplift* series. In this series, there are several orders of sentiency, including humans and sentient machines.

The reason I started this working paper with a mention of science fiction (SF) literature is that SF has a reliable track record for predicting future technological developments. For example, we have seen computers, space vehicles and gene engineering emerge from science fiction into science fact. Will we see the same in regard to thinking machines? Of interest, the transition from fiction to fact in the case of computers and space travel occurred quite rapidly (within a decade or two). Gene engineering has been a little slower to make this transition. As yet, although thinking machines have been around as fiction for at least six decades, we have not yet seen them as fact. This anomaly is noticeable in view of the power of SF prediction in other areas of technology. One possible reason may be that the problems associated with making the transition seem to be many orders greater than those involved in earlier transitions because the issues are not simply technological. Thinking implies a mind that does this, hence the challenge to achieve this breakthrough. The central theme of this paper addresses this challenge in discussing the question, 'Can a machine think as we humans do?'.

As a prelude, I need to touch briefly on the historical developments within the field of artificial intelligence (AI).

# A brief history of artificial intelligence

Research on artificial intelligence began in the 1940s, soon after the development of the modern digital computer. Early investigators quickly recognised the potential of computing devices as a means of automating thought processes. The term *artificial intelligence* (AI) was first coined in 1956, at the Dartmouth conference (see shortly). Since then, AI has expanded to become a major aspect of computer science and technology.

Although the computer provided the technology necessary for AI, it was not until the early 1950s that the link between human intelligence and machines was really observed. Norbert Wiener's research (Wiener, 1948/1972) into feedback loops led him to theorise that all intelligent behaviour was the result of feedback mechanisms and that electronic machines might possibly simulate such mechanisms. This thinking strongly influenced much of the early development of AI.

Late in 1955, Newell (with Simon) developed *The Logic Theorist*, considered by many to be the first AI program (Newell, 1961). The program, representing each problem as a tree model, attempted to solve it by selecting the branch that would most likely result in the correct solution. The impact that *The Logic Theorist* made on both the public and the field of AI made it a crucial stepping-stone in the development of the AI field.

In 1956, John McCarthy (for example, McCarthy, 1956), regarded as the key founder of AI, organised a conference to draw together the talent and expertise of others interested in machine intelligence for a month of brainstorming. He invited them to Vermont for *The Dartmouth Summer Research Project on Artificial Intelligence*. From that point on, owing to McCarthy's influence, the field would be known as *artificial intelligence*. The Dartmouth conference brought together the founders of AI and served to lay the groundwork for the future of AI research.

In 1957, the first version of a new program, the *General Problem Solver* (GPS), was tested. Newell (1961), who also developed *The Logic Theorist* (see above), developed the GPS program. The GPS was an extension of Wiener's feedback principle and was capable of solving a range of commonsense problems. In 1958, McCarthy announced his LISP (LIST processing) language, which is still used today. LISP was soon adopted as the language of choice among most AI developers.

In 1963, the Massachusetts Institute of Technology (MIT) received a 2.2 million dollar grant from the United States (U.S.) Government to be used

in researching machine-aided cognition (artificial intelligence). The grant, made by the Department of Defense's Advanced Research Projects Agency (ARPA), was used to ensure that the U.S. would stay ahead of the Soviet Union in technological advancements. The project served to increase the pace of development in AI research, by drawing computer scientists from around the world, and continues funding such research. The MIT research was headed by Marvin Minsky (for example, Minsky & Papert, 1969), who remains an influential figure in AI circles. Other programs that appeared during the late 1960s were STUDENT, which could solve algebra story problems, and SIR, which could understand simple English sentences. The result of these programs was a refinement in language comprehension and logic.

The 1970s saw the advent of the *expert system*. This is an advanced computer program, comprising a database and inference engine, that mimics the knowledge and reasoning capabilities of an expert in a particular discipline. The software attempts to replicate the expertise of one or several human specialists to create a tool that can be used by the layperson to solve difficult or ambiguous problems. Some examples of the application of expert systems are forecasting in the stock market, aiding doctors in disease diagnosis, and leading mining companies to promising mineral locations.

During the 1980s, AI has begun moving at a faster pace, especially in the corporate sector. In 1986, U.S. sales of AI-related hardware and software surged to $425 million. Expert systems were in particular demand because of their cost savings and efficiency. Today, expert systems and programs simulating human methods have attained the performance levels of human experts and professionals in performing certain specific tasks.

In terms of thinking AI machines, some (for example, Dennett, 1988, 1990, 1995) claim that machines such as IBM's *Deep Blue* can think. This is argued on the basis of the now famous series of chess games between the Grand Master Gary Kasparov and *Deep Blue* in 1997. While *Deep Blue* could examine a vast number of potential moves in milliseconds and had programmed in a range of famous gambits, it certainly could not think in the way that Kasparov could. Although Kasparov lost the first game to *Deep Blue*, he realised that it was playing rather ugly and rule-bound chess. As soon as Kasparov made unorthodox opening gambits, he started winning. He took *Deep Blue* outside its rule set. Despite its great speed and power, it was no match for a human thinker as it could not deal with ambiguity and unorthodox situations. It was simply following the algorithms programmed into it. The real intelligence lay in the mind of its programmer. I will return to this point later.

This brief history shows that there have been enormous advances in AI hardware and software. However, it is clear that there is a long way to go before thinking machines become viable. I now wish to look at AI machines and the extent to which they might be able to emulate human thought.

# AI machines and human thinking

The current challenge driving AI research is to understand how the capabilities of computers must be organised in order to reproduce the many kinds of mental activity that comprise 'thinking'.

Recent progress in the development of AI computers has led a number of philosophers (for example, Dennett, 1988, 1990, 1995) to conclude that a suitably programmed computer with a sufficient memory capacity would have an actual mind capable of intelligent thought. Two questions are intensely debated in this field:

(1) What are the theoretical limits to what can be achieved in the way of artificial intelligence? Despite phenomenal progress in recent years, no computer yet devised even begins to approximate in its capacity the powers of human cognition.

(2) Secondly, assuming that the optimistic hopes of artificial intelligence researchers are realized, would such devices literally have minds or would they be mere imitations of minds?

We can see that the first question is not so much about actual thinking as about computing capacity and power. This is quite a different issue from that of thinking. There is little doubt that, even today, computers far exceed human capacity, speed and power in terms of computation. In the narrow sense of cognition, where we restrict it to mean problem recognition and solution, computers have already far surpassed us. However, cognition implies far more than just problem solving. It includes an entire range of processing of internal and sensory data. And thinking entails processing beyond this.

The second question goes even beyond the issue of the thinking of AI machines thinking and invokes the question of mind. Thus, while we might concede that, with sufficient capacity and power, an AI machine might emulate human thought, the claim that it has a mind is of another order. I will return to these questions throughout this paper.

Today, we tend to describe computers in anthropomorphic terms: in terms of having memories, making inferences, understanding one language or another, making decisions and so on. However, are such descriptions literally true or simply imprecision in the use of language? There appears to be two opposing schools of thought in this debate. One holds that computers will never be more than tools employed by human intelligence to aid its own thinking (for example, Searle, 1980a, 1980b). The other school holds that human intelligence

itself consists of the very computational processes that could be exemplified by advanced AI machines, so that it would be unreasonable to deny the attribution of intelligence to such machines (for example, Dennett, 1988, 1990, 1995). This debate tends to be couched in terms of the strong AI hypothesis (arguing for thinking machines) and the weak AI hypothesis (arguing for AI machines simply as unthinking, non-conscious tools).

To address these issues, we need to look at the work of two important contributors to this debate who anchor the two extreme views. The first is Alan Turing, who, one could argue, initiated this debate five decades ago. He held the view that, with the maturation of computer technology, machines would one day be able to think. This is the strong AI hypothesis. The second contributor is John Searle, who has strongly argued against the notion that machines will ever be able to think. This is the weak AI hypothesis.

# Alan Turing

Alan Mathison Turing[1] was born in London in 1912, the second of his parents' two sons. His father was a member of the British civil service in India, an environment that his mother considered unsuitable for her boys. So John and Alan Turing spent their childhood in foster households in England, separated from their parents except for occasional visits home.

Alan's separation from his parents during this period may have inspired his lifelong interest in the operations of the human mind, how it can create another world when the world it is given proves barren or unsatisfactory. At 13, he was enrolled at the Sherbourne School in Dorset, where he showed a flair for mathematics, even if his papers were criticised for being 'dirty,' that is, messy. Turing recognised his homosexuality while at Sherbourne and fell in love, albeit undeclared, with another boy at the school, who suddenly died of bovine tuberculosis. This loss shattered Turing's religious faith and led him into atheism and the conviction that all phenomena must have materialistic explanations. There was neither a soul in the machine nor any mind behind a brain. His question was: 'how, then, did thought and consciousness arise?'.

For the war effort, on the basis of his published work, Turing was recruited to serve in the British Government's Code and Cypher School. The task for Turing and his colleagues was to break the *Enigma* codes used by the Nazis in communications between headquarters and troops. Because of secrecy restrictions, Turing's role in this enterprise was not acknowledged until long after his death.

After the war, Turing returned to Cambridge, hoping to pick up the quiet academic life he had intended. However, the newly created mathematics division of the British National Physical Laboratory (NPL) offered him the opportunity to work on the ACE (Automatic Computing Engine), and Turing accepted. Finding most of his suggestions dismissed, ignored or overruled, Turing eventually left the NPL for another stay at Cambridge. He then accepted an offer from the University of Manchester, where another computer was being constructed along the lines that he had suggested back in 1937.

While addressing a problem in the field of mathematical logic, he imagined a machine that could mimic human reasoning. What Turing did was to dream up an imaginary machine, a fairly simple typewriter-like device capable of scanning (reading) instructions encoded on a tape of theoretically infinite

[1] An excellent biography of Alan Turing can be found on the Alan Turing Homepage at http://www.turing.org.uk/turing/

length. The scanner moved from one square of the tape to the next, responding to the sequential commands and modifying its mechanical response if so ordered. Turing demonstrated that the output of such a process could replicate logical human thought. This imaginary device quickly acquired a name: the *Turing machine*. In addition, since the instructions on the tape governed the behaviour of the machine, by changing those instructions one could induce the machine to perform the functions of all such machines. By varying the programming of this machine, the same physical hardware could perform a range of functions, such as arithmetic, chess-playing and so on. It thus acquired a variation on the original name, becoming known as the *universal Turing machine*. (Turing, himself, actually referred to what has become known as the Turing machine as a *logical computing machine*: *LCM* [Turing, 1950].)

It should be noted that the notion of the universal Turing machine is related to the Church-Turing thesis (known also as Turing's thesis and Church's thesis). This connection arose out of the seemingly independent, but closely contemporaneous, work of Turing and Alonzo Church. Quite independently of, and a few months prior to, Church, Turing drafted a paper on the replacement of the informal effective method procedure by a formally exact predicate. Although this paper contains ideas that have proved of fundamental importance to mathematics and to computer science ever since it appeared, publishing it in the *Proceedings of the London Mathematical Society* did not prove easy. The reason was that Church had already published his 'An Unsolvable Problem in Elementary Number Theory' in the *American Journal of Mathematics* in 1936 (Church, 1936). The Church article also proves that there is no decision procedure for arithmetic. Turing's approach was very different from that of Church, but Max Newman (Fielden Professor of Mathematics at Manchester University) had to argue the case for publication of Turing's paper before the London Mathematical Society would publish it (Turing, 1936). It is interesting to note, in this connection, that Turing was one of Church's doctoral students at Princeton University at this time. One might speculate here on who influenced whom. Was Turing's draft influenced by his supervisor, Church, or did Church borrow from Turing's ideas and get in first with his publication?

These speculations aside, Church worked in the field of mathematical logic, recursion theory and theoretical computing science. In particular, in 1936, Church developed his theorem (Church, 1936), which states that there is no decision procedure for the full predicate calculus.[2] In this, he extended the work done by Gödel (for example, Gödel, 1934).

[2] A predicate expresses a relationship. In mathematics, the relationship is algebraic, such as ab is a times b. In grammar the relationship is between the subject of a sentence and what that subject is doing or what the subject is like.

Gödel's work on mathematical logic led him to investigate proof within systems of mathematics. Gödel devised a way of mathematically representing the sentence, 'This sentence is not provable'. If the representative equation is true, then the equation is beyond proof. If the equation does not hold (that is, the sentence is false), then the equation has a proof. In logic, statements must be either true or false. They cannot be simultaneously both true and false. Thus we have that the equation is true, meaning that the system of mathematics is incomplete in that it contains equations that cannot be proved; or we have that the equation fails to hold true, meaning that the system of mathematics is inconsistent and contains proofs of false equations. Thus, Gödel showed that, if mathematics is to be consistent, it must contain true equations that cannot be proved; hence, it is incomplete.

In the literature (for example, Baum, 2004), many authors refer to Gödel's theorem in the singular. There are, in fact, two theorems. The first theorem states that, in any consistent formalisation of mathematics that is sufficiently powerful enough to define the concepts of natural numbers, one can construct a statement that can be neither proved nor disproved. This is the best known of the two theorems, hence the tendency to refer to Gödel's theorem rather than theorems. This first theorem is also the most misunderstood. The second theorem states that any consistent[3] system cannot be used to prove its own consistency. Prior to this second theorem, there had been the belief that complicated systems (for example, of mathematics) could be proved in terms of simpler (sub) systems. However, Gödel's second theorem shows that even basic arithmetic cannot be used to prove its own consistency and so cannot be used to prove the consistency of anything more powerful, such as mathematical systems. The relevance of these two theorems to the work of Turing lies in his belief that his LCM (Turing machine) could generate any valid proof. Gödel's first theorem says that one cannot do this.

Gödel's theorems more generally deal with meaning in systems, in that the theorems show that no system, such as mathematics, can explain itself. This bears on the topic of syntax versus semantics, which will be addressed later in this paper and which is fundamental to the debate here. As thinking entails a semantical dynamic, hence meaning, and as Gödel's theorems show that computational systems cannot explain themselves, it is difficult to see how a computational AI system can think. However, I will return to this issue when I look a Searle's views shortly. Of note in this context is the claim by Penrose in his book, *The Emperor's New Mind* (1989), where he uses Gödel's theorems to argue that the human mind can operate outside the axiomatic rules of

[3] In this context, a system is consistent if none of its proven theorems can also be disproved within that system.

mathematics and so is not circumscribed by the incompleteness theorems; that is, the human mind is not restricted to what can be proven by 'mechanical' systems, such as digital computers, showing (for Penrose) that the human mind is not mechanical and that it uses non-computational processes. This is a serious claim, which, if valid, undermines claims that AI machines can think in the way that we humans do. I will come back to this issue when I look at Searle's views, and intentionality and Brentano's thesis.

In essence, the Church-Turing thesis deals with effective or mechanical methods in logic and mathematics. In this context, *effective* refers to a method (M) for achieving a desired result. In this case, M is a finite number of exact instructions, which, when carried out without error, will always produce the desired result in a finite number of steps. Turing's (hence Church's) thesis states that, whenever there is an effective method for obtaining the values of a mathematical function, the function can be computed by an LCM (Turing machine). In more modern language, we could say that, if we are dealing with computable numbers and an algorithm exists, then the problem can be computed (solved) by using a suitably programmed digital computer (Turing machine). It is worth noting here that, when in 1936 Turing used terms such as, *computer*, *computable* and *computation*, he used them in relation to human clerks who worked in accordance with effective methods. Computers, as we know them today, did not of course exist at that time. Turing was able to assert later (Turing, 1950) that his proposed LCM would be able to do all that a human 'computer' could do. However, this claim is valid only in reference to the existence of a definable algorithm that will lead to the 'program' halting, hence completing its task (a solution). This relates to Penrose's claim, mentioned above, and restricts Turing's use of the term *computer* to computable processes.

When Turing's seminal paper, 'Computing Machinery and Intelligence' was published in *Mind* in 1950 (Turing, 1950), no one recognised that Turing's machine provided a blueprint for what would eventually become the electronic digital computer. In the *Mind* paper, Turing proposed the idea that a machine could learn from, and thus modify, its own instructions. Especially, he proposed a thought experiment that he called an *imitation* game. In its original form the imitation game consisted of a man, a woman and a judge. The judge was separated from the man and woman by a screen and could neither see nor hear them. The only communication between them was by means of a teletypewriter device. The man was instructed to answer the judge's question in such a way as to convince the judge he was a woman. The woman was instructed to lead the judge to assume that she was a man. The idea was that, after sufficient questioning, the judge would come to realise the deception going on. Turing soon modified this game, replacing the man and woman with a human (gender unimportant) and a computing machine. In this revised version, the judge's

task was to decide whether the responding entity was a human or a machine. The judge was allowed to ask any question and, after an appropriate number of questions, was to make a decision as to whether the respondent was human or a machine. If the judge was either confident that the respondent was human or was unsure, then we can assume, where the respondent was a machine, that the machine passed the test. This has gone down in AI history as the *Turing test.* It is still the yardstick used in the debate on thinking machines and so is crucial in the debate on whether such machines can think.

Turing, himself, raised objections to the notion that machines can think, in his own paper (Turing, 1950), where he grouped his objections as follows:

(1) the theological objection

(2) the 'heads in the sand' objection

(3) the mathematical objection

(4) the argument from consciousness

(5) arguments from various disabilities

(6) Lady Lovelace's objection

(7) argument from continuity in the nervous system

(8) the argument from informality of behaviour

(9) the argument from extra-sensory perception.

The above are the objections as Turing worded them. To each of his objections, he responded as follows (the numerical sequence follows his):

(1) The argument here is that thinking is a function of man's (sic) immortal soul. God has not given a soul to non-human animals or to machines. Turing admits to not being very taken by theological arguments but rebuts this objection by stating that it places a serious restriction upon the Almighty. Turing argues that, if the Almighty chose to do so (being omnipotent), he could confer the power of thought on any animal or machine.

(2) Here, the argument is that the consequences of thinking machines are too dreadful to countenance. Turing says that this objection arises out of our sense of superiority over the rest of creation and is connected with the above theological argument. He offers consolation rather than refutation!

(3) This objection was far more serious, in Turing's view, in that it centres on the limitation of mathematical systems. In particular, he refers to Gödel's theorems, and his and Church's (see above). These theorems show that there

are limitations to the powers of discrete state computing machines. Turing supposes that, for the present, the questions in his imitation game are of the kind where a *yes* or *no* is appropriate. He accepted that a machine will fail with questions of the type: 'what do you think of Picasso?'.[4] Turing argues that this objection assumes that the machine is subject to computational limitations, to which the human intellect is not subject, and further argues that this sense of superiority over machines may be illusory. He points to our fallibility in a whole range of activities. Were he alive today, Turing would see that modern computers far surpass human computational ability, in terms of both power and speed.

(4) In this objection, Turing appears to be equating consciousness with what in philosophy is referred to as *qualia* (feeling states) as opposed to cognitive processes. In other words, a machine cannot think because it cannot feel. Turing says that this objection appears to be a denial of his imitation test. He says that, in the most extreme form of this view, the only way that one could be sure that a machine was thinking was to be that machine, hence to feel what it was feeling as it thought. He dismisses the objection somewhat trivially by arguing that it is solipsistic to raise this objection from the standpoint of consciousness. However, while it is true that we can infer that another human is thinking on the basis of their behaviours, we are not entitled to make the same inference with regard to a machine's behaviour[5].

(5) This objection centres on the idea that, while machines can do many things, there are also many things that they cannot do. Turing cites examples such as displaying kindness and friendliness, and falling in love, among many others. Turing appears to attribute these disabilities in a machine to a lack of storage capacity. In more modern terminology, this is another way of saying that the potential of thinking machines is a function of the degree of maturation of computing science and technology (I touch on this point later). Turing argues that this objection is related to the objection from consciousness and implies that, if a machine could write a sonnet on love, then it could feel love, hence think.

---

[4] As an aside, we do now have AI machines that could provide a reasonable answer to a question of this type, as long as its database has information about Picasso that its inference engine can draw on, for example, the CYRUS program (Kalonder, 1983), developed several decades ago, where CYRUS stands for Computerized Yale Retrieval System. However, as Dennett (1990) points out, CYRUS was modeled on the memory of Cyrus Vance, the then Secretary of State within the Carter Administration. One could address this program as though one were addressing the real Cyrus Vance, with questions such as 'last time you went to Saudi Arabia, where did you stay?', to which the machine would respond, 'In a palace in Saudi Arabia on the 23rd of September in 1978'. CYRUS could correctly answer thousands of such questions. However, as Dennett reports, when he asked the question, 'Have you ever met a female head of state?', CYRUS failed to answer either yes or no. It seems the software could not make the connection between the facts that, for example, Margaret Thatcher was both a head of state and a female.

[5] This, in my view, is a crucial objection, in that Turing comes close to admitting that his imitation test is not a test of thinking but of the simulation of thinking.

(6) The Countess of Lovelace (who published in the mid-1800s under the pseudonym, Ada[6]) was a brilliant mathematician, who did work with Charles Babbage on his analytical engine and who forecast that analytical engines would be able to perform many human tasks, such as writing music. Her objection was that analytical engines have no pretensions to originate anything. They can do only what we order them to do. One implication here is that a machine cannot learn from its experience and thus modify its behaviour. (Another is that we are not consciously aware of all the steps we take in performing certain tasks, but Turing does not pick up on this aspect.) As Turing points out, Lady Lovelace was writing this well before the electronic computers of his time and, we could add, well before the machines of our time. Therefore, she was not encouraged to believe that machines could 'think for themselves'. The Lovelace objection entails the issue of novelty, with its implication that machines cannot generate novelty. Turing addresses this implication by reducing it to an issue of a machine doing something surprising or unexpected, thus somewhat trivialising the objection. My computer often takes me by surprise (for example, I type a C and it turns it into a ©). But this is no indication that it can generate novelty or has some inbuilt creativity. It is simply a matter of my having forgotten that my XP word-processing software will do this unless I instruct it otherwise. There is no evidence of thinking here!

(7) This objection contrasts discrete state machines (for example, digital computers) with continuous state biological systems, such as those in the neurons of our brains. Turing dismisses this objection, arguing that, in the imitation test, it is not an issue because the judge cannot know whether the responder is a machine at all, much less what type of machine. In Turing's time, there existed differential analysers that worked on analogue rather than digital principles (in fact, Babbage's engine was one such device); hence, they were continuous state devices.

(8) This objection centres on the fact that, in essence, we are not rule-bound systems; whereas a machine is such. We can perform well in ambiguous situations. Turing cites a faulty traffic light system where both red and green appear simultaneously, arguing that we make a decision based on conditions prevailing where issues of safety are paramount. Turing rightly argues that a machine could be built to make such decisions. In fact, today, AI systems make far more complex decisions under conditions of ambiguity.

[6] Lady Lovelace's full name was Augusta Ada Byron. She was the daughter of Lord Byron, the English Poet Laureate.

(9) Of interest, Turing regarded this objection (based on human extra-sensory perception — ESP) to be a strong one. His concern was that a human could provide correct answers (in the imitation test) under certain conditions, where the machine could not. For example, Turing assumed that a human having some degree of ESP might be able to correctly answer a question about the card being held in the judge's hand, whereas the machine could not do this, because it lacked ESP. Frankly, it is hard to see why this objection rattled Turing. Assuming that ESP does exist, it would seem that few of us posses it to any useful degree, and yet we still think.

At this point, before moving on to discuss Searle, we need to consider several things about the Turing test. Firstly, Turing used the terms *intelligence*, *consciousness* and *thinking* somewhat interchangeably, regarding a success in the test as evidence that the machine was conscious/intelligent and was thinking. I feel that it is confusing to use these three quite different-meaning terms synonymously as Turing did (and as some still do). From the human viewpoint, consciousness is the hierarchically senior term in that intelligence and thinking entail being conscious. In the same way, thinking entails being intelligent.

The Turing test is regarded as a severe test because the judge can ask any form of question, seek a viewpoint, and 'test' the responding entity in ways he/she regards as appropriate.

However, in my view, no matter how severe a test it is, it is not a test of the machine's ability to think or a test that it is actually thinking. It merely tests the machine's ability to *simulate* thinking (Turing himself used the term *mimic*). Some will argue that this is a subtle distinction. I will argue, not so. There remains a major difference between a machine's simulating (mimicking) thinking and actually thinking, and to test this difference will require something other than the Turing test as currently devised.

In relation to the Turing Test, Harnad (1990) raises the issue of total performance in his example of a 'pen-pal' that might well be a machine that has 'fooled' its correspondent for years into perceiving the pen-pal as human. The correspondent sees nothing of the robotic performance of the pen-pal, which limits the usefulness of the Turing Test (TT). Harnad suggests a total performance measure, which he terms the Total Turing Test (TTT) and which can test robotic performance capacities. However, Harnad questions even the TTT, in that, whilst it measures for indistinguishable performance, it does not measure indistinguishability down at the neuro-molecular level. He wonders if we need an even stronger test (TTTT) but concludes that the TTT version suffices.

The commemorative essays collected by Millican and Clark (1996) cover a wide range of views on Turing's Imitation Game and his test. Likewise, the insights provided by Preston and Bishop (2002) help us to come to an understanding of Turing's brilliance and the seminal contribution he made to this debate.

Unfortunately, reality caught up with Turing well before his visions would, if they ever could, be realised. In Manchester, he told police investigating a robbery at his house that he was having 'an affair' with a man who was probably known to the burglar. Always frank about his sexual orientation, Turing this time got himself into trouble. Homosexual relations were a felony in Britain at that time, and Turing was tried and convicted of 'gross indecency' in 1952. He was spared prison but subjected to injections of female hormones intended to dampen his lust. 'I'm growing breasts!' Turing told a friend. On June 7, 1954, he committed suicide by eating an apple laced with cyanide. He was 41.

# John Searle and the Chinese Room

John Searle, an American, is currently Mills Professor of the Philosophy of Mind and Language, at the University of California Berkeley campus. Searle has opposed the argument that AI machines can think. In 1980, he published (Searle, 1980a) an article that was and remains a cause of ferment in the debate on thinking machines. In his article, he proposed a Gedanken experiment, which has become known as the *Chinese Room*. In this thought experiment, Searle attempts to rebut Turing's assertion that a machine that passed his Imitation Game (the Turing test) was thinking. In his argument Searle asks one to imagine a non-Chinese-speaking person (Searle himself) sitting in a room with a long list of rules for translating strings of Chinese characters into new strings of Chinese characters. When a string of characters is slipped under the door, the person consults the rules and slips back an appropriate response under the door. If the incoming strings actually represented questions (like a Turing test), then a particularly cleverly contrived and exhaustive set of rules could conceivably allow the person in the room to produce outgoing strings that furnished answers to the questions. From the point of view of a person outside, the room would seem to contain an intelligent Chinese-speaking person who is responding to the questions. But the person in the room has no understanding of the content of these questions (having no understanding of the Chinese language) and is merely acting out a set of rules, translating one set of random symbols into another. It could, just as well, be an AI machine in the room, not a person.

In this thought experiment, Searle wished to disprove the notion of the strong AI hypothesis, which argues that an appropriately programmed computer really has a mind; that is, that a computer, given the right program, can be said to understand and have other cognitive states. Thus, strong AI argues that the programs are themselves the explanations (Searle, 1980a). Searle's Chinese Room experiment shows that, although the 'Room' appears to have an understanding of the Chinese language and is evidencing thought processes, no such thing is actually happening. The 'Room' is not thinking, nor does it possess intelligence. It is simply following a set of instructions in just the way that a stored program machine is.

Searle is willing to consider that a computer might pass the Turing test but considers that it will not think or possess intentionality, two attributes he considers central (we will return to the issue of intentionality). His formal argument is that computers operate on syntax, that thought is semantical, that syntax does not produce semantics, and therefore that computers do not think.[7]

Various objections were put forward to rebut Searle's stand during the drafting of his article (Searle, 1980a), when Searle had the opportunity to discuss his thought experiment with a number of workers in AI. In his article he responds to each objection, categorising them as follows:

- **Systems:** In this objection, it is suggested that the non-Chinese speaker is only a part of the system, where the whole system does understand the strings of Chinese characters; that is, it is false to claim that the entire system does not think and understand simply because one component (the human in this case) does not. This objection is linked to a related objection known as the *Virtual Mind* objection. In this linked objection, it is argued that the 'mind' that understands is not identical to the human in the system; that is, this objection is distinguishing between the mind and the realising system.

- **Robot:** This objection arises from the view that the person in the Chinese Room is prevented from understanding by a lack of sensori-motor connection with the reality that the Chinese characters represent. The idea here is to put the Chinese Room into a robot that has sensori-motor capabilities, which impart perception and locomotion, hence engagement with reality.

- **Brain simulator:** The argument in this objection is that the 'program' implemented by the person in the Room simulates the actual sequence of neuron firings at the synapses of the Chinese speaker who receives the outputs from the Room. The argument here is that, at the synaptic level of neuronal firing, there is no difference between the 'program' in the Chinese Room and the program in the Chinese reader's brain. This being so, the 'Room' has just as much understanding as the Chinese reader.

- **Combination:** In this objection, it is supposed that the Chinese Room is lodged in a robot that is running a brain simulation program. It is argued that here we would have to ascribe intentionality to the entire system, in that the whole behaves indistinguishably from a human.

---

[7] In general, syntax refers to language structure, and semantics to meaning. However, in the next section I show that there is a difference between the linguistic (natural) language and computer programming use of these two terms, where this is relevant to Searle's syllogism.

- **Other minds:** This objection arises out of the consideration that we know that others have minds (and so think) by inferring this from their behaviour. Thus, if one can legitimately attribute cognition to humans based on their behaviour, one is required to do the same in the case of the Chinese Room.

- **Many mansions:** This objection suggests other means than programming in order to confer intentionality and cognition on the Chinese Room. The implication is that these other means are non-computational.

Searle responded to each of these objections in turn, as follows:

- **Systems rebuttal:** Searle responds by imagining himself to internalise all the elements of the system, by memorising the instructions, and so on. However, he still understands nothing of the Chinese language and neither does the system.

- **Robot rebuttal:** Searle replies that such perceptual and motor capacities add nothing in the way of understanding. He imagines himself in the room inside the robot, computationally acting as the robot's homunculus. He argues that, by instantiating the program, he has no intentional states of the relevant type (relevant to the Chinese language).

- **Brain simulator rebuttal:** Searle replies that even getting close to the operation of the brain is still not sufficient to produce understanding. He envisages a system of valves and water pipes that simulate the neuronal structure of the Chinese reader. Instead of manipulating pieces of paper, the person in the Room operates the valves, where each water connection corresponds to a synapse in the Chinese reader's brain. Searle argues that the person still has no understanding of the Chinese language and nor have the valves and pipes. The key word for Searle seems to be 'simulator', in that all the system does is simulate intentionality and thought. This harks back to the discussion about Turing's test, which is a behavioural test only.

- **Combination rebuttal:** Searle replies, in effect, three times nil is still nil. He does concede that it is tempting to attribute intentionality to the robot combination if we do not know how it works. However, Searle argues, once we can account for the behaviour of the combination, we cannot attribute intentionality to it; that is, we now know what is occurring inside the Room and so must concede that there is no intentionality nor thinking nor understanding, despite appearances.

- **Other minds rebuttal:** Searle dismisses this as an epistemological worry beside his metaphysical point. 'The problem in this discussion,' he says, 'is not about how I know that other people have cognitive states. But rather

what it is that I am attributing to them when I attribute cognitive states' and 'It couldn't be just computational processes and their outputs because the computational processes and their outputs can exist without the cognitive state' (Searle, 1980a, pp. 421–422).

- **Many mansions rebuttal:** Searle replies that this 'trivializes the project of strong AI by redefining it as whatever artificially produces and explains cognition' (Searle, 1980a, p. 422). In conclusion, Searle advances his own thought that the brain must produce intentionality by some *non-computational* means that are 'as likely to be as causally dependent on … specific biochemistry … as lactation, photosynthesis, or any other biological phenomenon' (p. 424).

In his original paper (Searle, 1980a), Searle explains that he did not offer a proof that computers are not conscious. Rather, he offered a proof that computational operations by themselves, that is, formal symbol manipulations by themselves, are not sufficient to *guarantee* the presence of consciousness. The proof was that the symbol manipulations are defined in abstract syntactical terms and syntax by itself has no mental content, conscious or otherwise. Furthermore, the abstract symbols have no powers to cause consciousness because they have no causal powers at all. All the causal powers are in the implementing medium. A particular medium in which a program is implemented, a brain for example, might independently have causal powers to cause consciousness. However, the operation of the program has to be defined totally independently of the implementing medium since the definition of the program is purely formal and thus allows implementation in any medium whatever.

In a companion article (Searle, 1980b), Searle expands on his Chinese Room arguments, objections and rebuttals. In particular he elucidates the distinction between *intrinsic intentionality* and *observer-relative ascriptions of intentionality*, defining the former as the kind of intentionality that we humans have and the latter as the ways we have of talking about machines or similar entities that lack intrinsic intentionality. Searle argues that we cannot attribute intrinsic intentionality to machines (for example, carburettors and thermostats) because machines do not posses beliefs, whereas humans do. He also argues that the fact that he cannot explain how the brain works to possess intrinsic intentionality is not grounds for dismissing his view. No one can yet explain it, but this doesn't alter the fact that people possess intrinsic intentionality. While not subscribing to some 'numinous Cartesian glow' (as accused of by Rorty, 1980),[8] Searle

[8] Rorty questions Searle's claim that human mental phenomena are dependent on physico-chemical properties of the brain, arguing that this claim is a device for insuring that the 'secret' powers of the brain are pushed further and further out of sight every time a new brain model looks as though it might explain mental content.

argues that mental states are not dispositions in the sense that temperature responsiveness is a material disposition of a thermostat. He says that mental states cannot be explained in terms of dispositions as they are 'made' of qualia and partake of ontological subjectivity.

Searle explored these arguments further still in his Mind, Brains and Science (Searle, 1984), where, in chapter 2 ('Can Machines Think?'), he argues that his Chinese Room is a decisive refutation of the strong AI view that mind is to brain what program is to computer. In his book, Searle states his axiomatic premises in this way:

1. Brains cause minds.
2. Syntax is not sufficient for semantics.
3. Computer programs are entirely defined by their formal, or syntactical, structure.
4. Minds have mental contents; specifically, they have semantic content.

From these premises, Searle drew several conclusions:

- No computer program by itself is sufficient to give a system a mind. Programs, in short, are not minds and they are not by themselves sufficient for having minds.
- The way that the brain functions to cause minds cannot be solely by virtue of running a computer program.
- Anything else that caused minds would have to have causal powers at least equivalent to those of the brain.
- For any artefact that we might build which had mental states equivalent to human mental states, the implementation of a computer program would not by itself be sufficient. Rather, the artefact would have to have powers equivalent to the powers of the human brain.

As mentioned above, the CRA led to an on-going debate between those who supported Searle and those who opposed him. This debate is still going on into the 21st century. There is not the space in this paper to provide a full and detailed account of the more recent contributions to the debate, so I shall restrict myself to what I feel are the key contributions.

Wakefield (2003) reminds us that Searle's Chinese Room argument (CRA), in using the Chinese Room experiment (CRE), aims at refuting computationalism, hence strong AI. He believes that many readers will consider CRA to be refuted. As presented by Searle, CRA is a question-begging (where the evidence given

for a proposition contains the proposition itself) appeal to intuition, in that it relies on faulty intuitions. Wakefield discusses the systems objection and Searle's rebuttal, in which the person in the Chinese Room instantiates the room; that is, the operator in the room *is* the system. The challenge then, for strong AI, is to demonstrate that the operator understands Chinese, because the operator instantiates the selfsame programme as a native Chinese speaker. Wakefield points out that Searle argues that instantiating the right programme cannot be what confers understanding, because the operator has no such understanding despite instantiating the right program.

In Searle's original article (1980), in response to the 'Systems' reply, he instantiated the room within the operator, so that the operator internalises all the elements of the CRE. She memorises the rules in the ledger and the data banks of Chinese symbols, and does all of the calculations in her head. There isn't anything at all to the system that she doesn't encompass. Searle says that we can even get rid of the room and suppose that the person works outdoors. However, she still understands nothing of Chinese; thus, neither does the system. Wakefield points out that, later, in a republication of his original article (Searle, 1991), Searle amended this scenario, so that the manual has been rewritten to apply to input sequences of sounds rather than to written symbols. Rather than getting an input sequence of Chinese characters, the operator listens directly to the speaker and utters the Chinese responses herself. With practice, the operator can become flawless in her responses, so that there is no noticeable difference between her responses and those of a fluent Chinese speaker.

Wakefield argues that Block's (1998) objection to Searle's revised rebuttal stands or falls on Block's ability to draw a relevant distinction between the program's and the operator's meaning within a computationalist account of meaning. However, Searle's argument is subtler than Block allows. The program in the CRE exists as thoughts in the mind of the operator. Thus, for computationalism, the operator and program must possess the same content. As the operator understands only in a syntactic fashion (without meaning), the program cannot understand Chinese. If implemented syntax constitutes semantic content, then, because the operator does not understand Chinese, neither does the program. Wakefield asserts that Block is trying to imply a multiple personality condition, distinguishing between operator and program states. However, there is nothing in the CRE that might cause a difference in syntactic steps between program and operator. If computationalism is correct, there is no way for the program to understand Chinese unless the operator possesses the same understanding.

Wakefield cites Fodor (1991), who accepts that the CRE does not understand Chinese. However, Fodor argues that the program would understand Chinese without the intervention of the operator, because introducing the operator has

rendered the implemented system non-equivalent to the original program of the Chinese speaker. Searle assumes that introducing the operator preserves the Turing-machine equivalence. Fodor is challenging this assumption, in that transitions between program steps are not direct and involve further mediation (conscious, deliberate actions); that is, these mediating steps are part of the program. Hence, the program is not equivalent to programs lacking such mediating steps. Fodor argues thus because he claims that a transition from one state to the next is not directly and proximally caused by the prior state. Wakefield says that Searle responded that introducing causal relations does not affect the CRE because, whatever caused the inputs, the processing of symbols could still proceed without any understanding of the semantic content of the symbols. Searle also argues that the introduction of conscious implementation of the program by the operator does not cause the system to become a non-Turing machine. The operator is still following the program steps. In fact, conscious implementation ensures that one step is directly caused by another (Fodor's requirement). As Wakefield points out, in any case, the crucial issue is whether the operator understands the arriving sentences and not the causal relation to the emitter of these sentences. Also, because the modified CRE internalises the entire program within the operator, she is directly receiving input, is internally implementing all program steps involved in understanding the input, and is directly responding herself. Thus, the new CRE eliminates any causal chain deviance.

Wakefield turns to Hauser's (1997) argument about unconscious understanding on the part of the operator, in which Hauser claims that, although he accepts that the operator does not consciously understand Chinese, she does *unconsciously* understand Chinese and has unconscious Chinese semantic content. I will return to this when discussing Hauser's contribution to the debate.

Wakefield reports that, in relation to the unconscious argument, Searle (1991) says that a person could not have an unconscious understanding unless, at least in principle, the content could become conscious. In the case of the CRE, any first-person report of unconscious content becoming conscious would reveal only syntactic descriptions and not semantic content. In any case, there is no need to postulate an unconscious understanding in order to explain the system. Wakefield argues that the critical question for strong AI and the CRE is whether the operator really understands Chinese. There is a distinction between being 'unaware' of content one actually possesses unconsciously and plain ignorance, in which one does not possess the content at all, either consciously or unconsciously.

Wakefield says that the CRE yields the intuition that the operator does not understand Chinese, either consciously or unconsciously. The CRA uses this result to argue that computationalism cannot be true. Wakefield introduces an *essentialist* objection to the CRA, which argues that the common pre-theoretical intuition that the Chinese Room operator does not understand Chinese is not an appropriate reason for concluding that she does not in fact understand Chinese. Wakefield argues that computationalism is best considered as a theoretical claim about the essence of meaning. An essential description (for example, whales are mammals) contrasts with an intuitive description (whales are fish). Computationalism holds that the essence of standard cases of human intentional content is the running of certain formal programs. Thus, if anything shares this essence it also has intentional content. For Wakefield, the CRE presents a non-standard human instance that possesses that essence but violates our pre-theoretical intuitions regarding the possession of intentional content. The computationalists claim that the CRE essentially shares the human essence of intentionality and thought, even though it is non-standard. But this demands that computationalists successfully explain intentional content in typical human thought in computationalist terms.

Wakefield holds that the CRA poses a threat to strong AI, if reframed as an indeterminacy argument. Strong AI holds that the operator's intentional states are determined simply by the program she follows. It is difficult to counter this without question-begging (that is, there's no genuine intentionality in virtue of formal programming alone). However, Wakefield offers an indeterminacy test. If the Chinese-understanding program leaves claimed intentional contents indeterminate in a way that genuine intentional contents are *not* indeterminate, then we can say with confidence that the program does not constitute intentional content. Basically, this shows that computationalism is unable to account for how anyone can ever understand Chinese, even in standard cases of human thought that intuitively are clear instances of genuine Chinese understanding. The issue is that there is a distinction between manipulating the syntactical elements of a language and actually understanding that language at a semantic level. Wakefield cites Quine (1960), who claimed that semantical and intentional content remain indeterminate (that is, open to multiple incompatible interpretations consistent with all possible evidence) if the relevant evidence is limited to syntax alone.

Wakefield further argues that the computationalist claim that to think a certain semantic or intentional content is just to be in a certain syntactic state is mutually incompatible with the claim that, in the CRE, all the operator's thoughts and actions relate only to the program's syntactic structures and transitions. The indeterminacy consists, then, of the fact that, consistent with all

the syntactic and programming evidence that strong AI claims to exhaust the evidence relevant to fixing content, a person who appears fluent in a language may be meaningfully using the language or may be merely implementing which states are identified syntactically. Thus, there may not be any meaning to her utterances. For each brain state with a syntactic structure S that would be interpreted by strong AI as a thought with content T, the person could have T or could have the thought 'syntactic structure S'. For example, someone might make the utterance: 'Please pass the salt' and intend the meaning of that request, or they may be following some program in which they utter the noise: 'Please pass the salt'. There is no evidence in the program itself that could distinguish which of these two interpretations is correct. The indeterminacy argument might go as follows:

- There are determinate meanings that distinguish between thoughts and intentions, and thoughts about syntactic shape. Thoughts dealing with syntactic shape are different from those that possess semantic content as expressed by those shapes.

- All syntactic facts under-represent (hence leave indeterminate) the content of thoughts and intentions-in-action. The syntactic structure S is ambiguous between the meaning M of S and the meaning the program specifies to be the syntactic structure of S.

- Thus, the thoughts and intention-in-action cannot be constituted by syntactic facts.

This argument provides the support needed for Searle's third premise. In the CRE, every sentence in Chinese (with its usual semantic content) is translated into a sentence about the syntax of the original sentence. In this context, the challenge for computationalism is: 'What makes it the case that people who in fact understand Chinese do have genuine semantic understanding and that they are not, like the operator in the CRE, merely manipulating syntax where meanings are unknown to them?' Any claims that the operator does understand Chinese demand that the claimant distinguish between that and ordinary understanding of Chinese or at least explain why they are different. However, the indeterminacy approach concludes that computationalism cannot do this. Computationalism remains an inadequate account of meaning.

Teng (2000) is generally in support of Searle's CRA and analyses it within a cognitive model based on schemas and metaphors with the notion of conceptual blending. Teng focuses on Searle's assertion that syntax is not intrinsic to physics, and argues that the assertion is a modified version of the original CRA. Strong AI claims that implementing the right program is sufficient for having

a mind. That is, cognition is computation, which is a purely syntactical set of operations. As seen earlier, in the refuting of this claim, Searle's CRA has three steps:

1. Programs are entirely syntactical.

2. Minds have a semantics.

3. Syntax is not the same as, nor by itself sufficient for, semantics.

Therefore, QED, programs are not minds (Searle, 1990, 1997).

Teng argues that the CRA amounts to a blending of a number of cross-domain mappings. The generic space is the physical system capable of manipulating the symbols on the basis of their syntactical features, according to defined procedures. The source space is the person who understands English, but not Chinese. The target space is the computer for manipulating the Chinese characters. This is based on the computer-as-person metaphor, where Teng concludes that this metaphor does not understand Chinese solely on the basis of implementing the right program.

Teng then looks at the 'syntax-is-not-physics' argument of Searle, using the cognitive modelling as above in relation to the CRE. However, Teng is misapplying this argument, because, in the context of his 'syntax-is-not-physics' assertion, Searle was not talking about the CRA or CRE specifically in his 1990 paper. He was dealing purely with digital machines in relation to their physical construction. Nonetheless, Teng correctly draws the conclusion that a computer operates on physical patterns without 'knowing' that these physical patterns are symbols of some language. These physical patterns are merely physical patterns, which suggests (as Searle argues) that syntax is not intrinsic to physics. This being the case, computation, algorithms and programs are not intrinsic to physical systems. Thus, they are notions assigned relative to observers and users.

Teng (2003) also deals with the debate between Chalmers (1994) and Searle, in which Chalmers argues that the main problem with the CRA is that it fails to respect the role of implementation. He argues that implementation requires that the system have the right causal power built into it. The CRA does not rule out the possibility that semantics can arise from implementing the right program. Searle argues back that programs are purely syntactic and observer-relative. The issue for Searle is that of physical systems that are capable of having a mind. Brain processes bring about mental states, and since programs are observer-relative, there is no such thing as programs involved in the operations of brains that give rise to the basic features of minds. For Searle, the first-person point of view is essential to understanding a language. As the CRE lacks such

a viewpoint, it cannot judge for itself whether or not it understands Chinese. It doesn't matter whether a brain or other physical system carries out the implementation, as long as the right causal dynamic is secured, in that cognitive systems have the right mental properties by virtue of their causal organisation.

Hauser (1997) rejects Searle's CRA and describes it as an *ignoratio elenchi* (a logical fallacy that argues beside the point, reaching a conclusion that is irrelevant to the proof that was attempted). He regards the CRA as sophistry (a plausible but fallacious argument). He also accuses Searle of an *ad hominem* (personal rather than logical) appeal. Finally, he regards the CRA as logically and scientifically a dud, having mainly historical interest, which survives only as a sociological phenomenon. These are strong accusations. However, part of the strength of Hauser's arguments is that he chooses to see AI as synonymous with thinking. This is similar to Turing's original ideas in that Turing didn't distinguish between intelligence and thought. This seems fallacious, in that intelligent behaviour doesn't imply thought or thinking. After all, many mammals display intelligent behaviour in acting in a problem-solving and adaptive way. The debate engaged in by Searle is about thinking machines, not machines that exhibit intelligent behaviour. Most would agree that modern AI systems behave in an intelligent fashion. The Turing test is a test of intelligent behaviour (or at least the simulation of such behaviour). But, the key thing is that intelligent behaviour is not synonymous with thinking.

Hauser categorises the various arguments within the debate as follows:

- Possible AI (PAI): the claim that computers will one day think
- Actual AI (AAI): the claim that computers already do think
- Essential AI (EAI): this is computationalism in which program is identified with mind, where the claim is that anything that computes entails thinking. It is what Searle calls strong AI.

The first two above are, collectively, AI proper. Hauser distinguishes the above three metaphysical claims from epistemological or methodological claims that programs explain human cognition, where Hauser calls these latter claims *cognitivism*.

With regard to Searle's key syllogism (note that Hauser phrases this somewhat differently from either Searle or Teng, but it amounts to the same thing):

1. Programs are formal (syntactical).
2. Minds have mental contents (semantical).

3. Syntax by itself is neither constitutive nor sufficient for minds

4. Therefore, programs are neither constitutive of, nor sufficient for, minds.

Hauser argues that this syllogism's point is to show that you cannot get the mental content from syntax alone. He argues that it is not at all obvious that existing AI systems do not have causal powers equivalent to brains. It begs the question against AAI to assume computers lack these powers. He views the argument as little better than claiming: 'Human brains cause mental states'; 'No digital computers are human brains', therefore, 'No digital computers cause mental states'. He explores the notion of equivalence, looking at what it means for something to have the causal equivalence to a human brain. Does it mean having the equivalent brainpower? Is it equivalent to behave 'as if', or does it require true possession of equivalent mental powers? In this context, there is the issue of hierarchical levels of mental powers, where Searle has argued that many species display intentional actions, yet lack intentionality. But Hauser argues that we don't know what features of human brains determine their mental capacities. Searle has suggested that intentionality (hence mental powers) is dependent on biochemistry. But, this, in itself, doesn't disconfirm AI proper. For Hauser, Searle seems to be arguing that it is a certain kind of chemistry (X) that leads to intentionality. Thus, any non-human system lacking X lacks intentionality.

Hauser regards Searle's arguments as sophistry, in that Searle classes what Hauser has dubbed essential AI (what Searle calls Turing Machine functionalism) as strong AI while, in the same breath, contrasting strong AI to the thesis that computers merely simulate the mental abilities that they seem to manifest (what Searle calls weak AI). I find it hard to see the sophistry of Searle's arguments, in the above. Essential AI (computationalism by Hauser's own admission) is clearly what Searle is talking about in his definition of Strong AI (that computers can think). There is no falsity of argument to then say that this is in contradistinction to what Searle is calling weak AI (computers do not think).

Of interest, Hauser believes that, in part, AI's detractors are motivated by a fear of thinking machines (as raised in one of the objections to Turing's ideas). Turing held that those who regarded thinking machines as too dreadful are likely to be intellectuals, since they value the power of thinking and regard Man (sic) as superior. Among such intelligentsia, Hauser argues, are those who say that computers can think but don't. I'm not sure what this adds to the debate. Whether or not fear is a factor for the detractors (or optimism a factor for the supporters) does not alter the fact that they must rationally defend their viewpoint.

Hauser gives what he terms a bare bones version of Searle's syllogism as follows:

1. If instantiating a right program R is a sufficient condition for understanding Chinese, then anything that instantiates R understands Chinese.

2. While in the Chinese Room, Searle himself instantiates R but fails to understand Chinese.

3. Instantiating a right program is NOT a sufficient condition for understanding Chinese.

Hauser relies on arguments about the inability of a human to process the card shuffling fast enough to pass the Turing Test, to dispute the conclusion (3) above. However, the CRA does not entail a speed of response criterion, only that, to someone outside the room, there appears to be someone in the room who understands Chinese. To my knowledge, no speed criterion was ever mentioned by Searle. Hauser also argues that by internalising the rules (minor premise 2 above), Searle would become cognisant of the meanings of the symbols. He further argues that, even so, one would *unconsciously* understand. In this, Hauser distinguishes between seeming to oneself to not understand and to really not understanding. He accuses Searle of a Cartesian identification of thought with private experience, with its corollary of private access, saying that privileging the first person fatally biases the CRA. However, there is no circular argument entailed in the CRE, which attempts to demonstrate that third-person competence is not sufficient for content attribution and that the first-person perspective is always relevant to such attributions. In raising the conscious-unconscious distinction above, Hauser himself is invoking a form of Cartesianism. I believe that Searle has argued adequately, that one cannot ignore the first-person subjectivity issue when considering intentionality and cognitive processes, as these relate to thought in any structure, organic or otherwise. One gets the impression that Hauser will invoke whatever it takes to dispute Searle's CRA and is sometimes guilty of those very things of which he accuses Searle. It may well be that fear is a motivator in those who reject the strong AI argument, but then one could argue that fear also motivates those who, at almost any cost, want to prove that machines can think. In their case, the fear seems to be centred on some potential failing of AI.

Reber (1997) accepts that Searle's CRA is correct in its essentials, but that it does not go far enough, believing that it runs afoul of the problem of emergentism. Reber says that an interesting feature of AI is that the program is what counts, not the device on which it runs. This is the hardware independence assumption. No one has ever argued that the computer used to run existing

AI programs is, itself, an intelligent entity. Both strong and weak AI share the hardware independence assumption. However, strong AI claims that a properly programmed machine will exhibit subjectivity, be conscious and have mental states. That is, strong AI goes far beyond claims that an appropriately programmed machine can simulate various forms of complex behaviour, claiming that such an entity really does have a mind. Weak AI has had success in constructing useful models of human cognitive functioning, whereas strong AI has had no such success. In addition, while weak AI's claims and successes have raised no controversy, strong AI's claims has raised a storm of controversy, especially among philosophers. One reason that Reber puts forward is that some people seem insecure at the thought that our precious mental capacity could be merely the carrying out of a program. Another reason comes from those who feel that the concepts of strong AI are fundamentally wrong: deeply flawed. Reber argues that this belief is the basis of Searle's approach in his CRA, which he regards as the strongest assault on strong AI.

Reber discusses the revised CRA, in which the room and all of its operations are instantiated in a human operator that Reber calls Maxine. He points out that Maxine is simply following a set of syntactic computational guidelines, satisfying the linguistic constraints of the Chinese language. In this sense, Maxine qualifies as an AI device and would pass the original Turing Test (TT) and perhaps even TTT. She lacks understanding and true intelligence. She is likewise bereft of qualia, mental content, semantic content. In essence, she lacks consciousness.

Reber claims that the CRA is against the root assumption of hardware independence, showing that, instead, the system is extremely hardware-dependent. In fact, mind is hardware-dependent, being the property of certain classes of organic matter organised in a certain way. Reber argues that Searle (and any other opponent of the strong AI claim) has not taken this line of reasoning far enough: has not taken it to its logical conclusion. Searle, Reber claims, begins to approach this view at the end of his original 1980 article, where he says that strong AI tells us little about machines since it is only about programs. True mental states, argues Searle, are products of biological machines, not of mere programs. These properties of organic systems are not explicable simply by providing a formal instantiation of the processes that underlie them. They are inherent, intrinsic features of organic systems such as brains. While these systems may follow a set of instructions that can be instantiated as formal programs, they are not equivalent to the organic systems, and the carrying out of these programs is not equivalent to the carrying out of the biological functions of organic systems.

Strong AI's argument is that mind emerges from programs; that is, run the right program and mentation will appear. Searle argues that this position is fundamentally flawed, that consciousness and other mental experiences are intimately tied not to a program but to a particular device that just happens to have consciousness as one of its emergent properties, and the brain is one such device. Reber points out that, in a sense, this raises more questions than it answers, especially the question: how does consciousness emerge from brains? The seductive thing about strong AI is that, if correct, the problem of mind becomes much easier to solve. As it is, Reber says that we don't have a clue as to how mind can emerge from a brain. However, Reber believes that we have been asking the wrong question. The error has been to focus on the human brain and wonder how mind is produced in it. Reber proposes that phenomenal experience is an inherent property of all living organic forms. It doesn't emerge only when the structure gets as complex as a human brain. It is there all along. Philosophers have worried mainly about human consciousness. Reber argues that we no more have to worry about how consciousness emerges from a brain than we have to worry about how gravity emerges from mass.

Reber claims that the major problem with current theories of the origins of consciousness is that they are hoist by the petard of emergentism, in that they all require a miracle. He argues that his proposal of some kind of inherent ur-consciousness in all living organisms sidesteps the emergentist problem and avoids going down the dualist path that the likes of Chalmers (1996) feel we must travel. However, Reber admits that his proposal, although avoiding having to explain how complex biological structures give rise to consciousness, still needs to wrestle with the issue of how organic systems of varying complexity permit the evolution of systems of varying richness. He feels that the solution to this problem will not come easily.

Reber claims that his position is a physicalist-reductionist one, which avoids miracles and dualistic garden paths. He is not a vitalist because he argues that consciousness is a part of the natural world.

There are others who have contributed to this debate, such as Jacquette (1989, 1990), and I have no doubt that this debate will continue because Searle's CRE and CRA are so threatening to those who wish to cling to Strong AI. While I can understand the fear that appears to motivate the likes of Hauser (1997), it does not justify the almost ridiculous lengths that he and others go to in their attempts to debunk the CRA, hence weak AI. I am inclined to agree with Reber (1997) in his analysis of the CRA, where he argues that consciousness (hence mind and thought) is hardware dependent, and that mind is an inherent feature of biological systems. If he is correct, then the CRA has won the day, and the strong AI proponents will have to back down with their tails between their legs.

In other words, while we could possibly programme a machine to imitate the effects of intelligence, it would never be truly conscious. This criticism is probably valid when applied to the old style of rule-based AI systems, rather similar to Searle's Chinese Room. Such systems exist (using digital combinatorial binary techniques) and have met with some success (see the discussion above on expert systems). However, it is not clear that it truly applies to neural network-based AI, since there is no real concept of a set of rules determining a response. Intelligence is not envisaged as arising out of a machine obeying a set of rules but as some as yet ill-defined property of the natural functioning of billions of neurons. Later, we will look briefly at artificial neural networks (ANNs), seeing the extent to which they differ from fixed instruction set, stored program devices. In this way, it may become clearer as to whether an ANN can provide the basis of a thinking machine. However, in this, we must acknowledge the view put forward by Reber (1997) who, I assume, would say that even the most complex ANN cannot think because it can never attain consciousness. But this would depend on the nature of the structures used in future ANNs. If biological tissues were used as a basis, it may be that whatever gives rise to primordial consciousness in biological systems would arise in such ANNs. It might be that future ANN technology will be able to create artificial brains that have structures and functions not alike natural brains, and where true artificial thinking becomes a possibility.

As the issues of syntax and semantics are crucial to Searle's arguments, we need to look at them in greater detail.

## Syntax versus semantics

The terms *syntax* and *semantics* derive from the linguistic tradition, and so refer to human language and the ways in which language provides a means of communication with symbols (sounds and writing). Linguistics rigorously analyses the forms and meanings within a language, where language definitions can be grouped into three components: syntax, semantics and pragmatics. In this context, syntax refers to ways in which symbols may be combined to create sentences. Syntax defines the formal relations between the constituents of language and deals solely with form and structure, without reference to meaning. Semantics reveals the meaning of syntactically valid strings. In human languages, this entails the correlation between sentences/phrases and the objects, thoughts and feelings of our experience. Finally, pragmatics refers to psychosocial phenomena such as utility, scope of application and the effects upon a language's users. Here, I wish to focus only on the first two terms.

In a computing context, these same terms take on a somewhat different meaning. In the case of syntax, the meaning is not much different from its meaning in linguistics because one is dealing with a programming code in which sentence structure is relevant. However, semantics describes the behaviour that a computer follows when executing a program. In this rather narrow and specialised use of the term *semantics*, one could accept that computers have a semantical dynamic. In this acceptance, there is no admission that any kind of thought processes occur within the program or machine. However, the proponents of strong AI are not staying within this narrow definition, but appear to be invoking the wider, natural language definition, which entails thoughts and feelings. It is this distinction between computing language and natural language that lies at the heart of Searle's argument, as mentioned above, from which can be stated:

- Programs are formal, hence syntactical.
- Mental operations are semantical.
- Therefore, programmed machines cannot think.

Note that the syllogism pertains only to digital, stored program machines. It may not apply to machines employing other AI technologies, such as the ANN, as mentioned above. As promised, I will return to this issue of other technologies and the bearing of the maturation of machine science on the debate.

For Searle, his syllogism is the end of the argument. As machines lack a semantic content, they cannot think. But Searle goes beyond even this in a

more recent revisitation of his earlier work (Searle, 1990). He argues that there is nothing at all in a stored program digital computer that can be regarded as thinking, including the program itself. (He argues in the same way for machines that do not use von Neumann architecture, such as in massively paralleled devices.[9]) The program is merely a list of coding, created by a human programmer. It is not even syntactical. It follows syntactical rules, but these are those of the programmer. The program simply causes the combinatorial logic gates in the hardware to operate in certain ways. Searle argues that the digital machine no more thinks than, say, a thermostat or a carburettor (see his earlier arguments above). Using rather anthropomorphic language, we tend to loosely say that such devices are self-regulating in that they modify their own behaviour (this, note, was a key issue for Turing). But Searle rightly argues that they do no such thing. The thermostat (using the bi-metallic strip) simply obeys the laws pertaining to the expansion of metals, and the carburettor simply obeys the laws of fluid dynamics. This same argument is just as true even in the most sophisticated cybernetic machine.

Searle (1990) more clearly defines what he meant in his 'Chinese Room' paper by strong and weak AI and also introduces the term *cognitivism*. Strong AI is the view that all there is to having a mind is a program. Weak AI, he defines as the view that brain (hence mental) processes can be simulated computationally. Cognitivism is the view that the brain is a digital computer. He points out that Turing can be seen to have argued that anything a human can do algorithmically can be done by a universal Turing machine. However, Searle says the human does this consciously, whereas the Turing machine does it unconsciously. While I agree with his exposition, I am not comfortable with his use of the term *unconsciously* because it still implies something human (for example, in the Freudian sense). Perhaps the better term might be *mechanically*, as in any computational engine. Of interest, Searle accuses the cognitivists of a form of dualism, in that they subscribe to an homunculus fallacy: the view that there is some agent within the human brain that uses the brain; that is, cognitivists ascribe an homunculus to the digital computer. But, Searle argues, there is nothing in a digital computer that can act as an agent. He argues that there is nothing syntactical or semantical in physics, hence, at the level of programming code and electronic gates. The homunculus resides in the programmer. Searle further argues that there is no computation as such going on in the digital machine, no more so than in a nail being driven into wood. There is no algorithm in the nail that is computing how far to go with each hammer blow. It is simply obeying the laws of physics. The same is true of the digital machine.

---

[9] Von Neumann architecture refers to the structure and function of binary digital computers that have a central processing unit (CPU) consisting of a clock, registers, instruction set, and memory, with input-output ports, that can be programmed to perform a variety of computational activities.

## Digital machines, human brain processes and cognitivism

Those whom Searle calls *cognitivists* regard the human brain as working in a fashion similar to that of a digital von Neumann-type machine. This view enables the cognitivists to categorise brains into hardware (perhaps, more correctly, wetware) and software (the programs). The neurological structures and processes constitute the hardware, and cognition/mind constitutes the program. Thus, by specifying the program, we have specified the causes of cognition. The neural hardware merely provides the means of implementing the program. However, as the analogy is a digital one, it means that the program consists of a coding list that manipulates binary digits (1s and 0s). It is not clear what, in the human brain, is the equivalent of 1s and 0s. The neuron doesn't work in a discrete digital manner. However, leaving aside this difficulty for the moment, even in an actual digital computer there are no 1s and 0s as such. There are voltage states that represent 1s and 0s, where these are realised in a solid state substrate (usually doped silicon) as the movement of holes and electrons. The 1s and 0s are a notion in the mind of the programmer. They do not exist as such and certainly have no causal powers. The physical program itself resides in hardware (say, on a compact disc). It is an implementing medium in the same way as the silicon chips that constitute the NAND and NOR gates in the computer hardware[10]. At the hardware level, the program has neither real existence nor ontology. The causal power resides in the programmer who creates the code. So, in this sense, it is erroneous to talk of a distinction in level between hardware and programming within the digital machine. There is simply implementing hardware.

It might be argued that I have played down some aspects of cognitive science in this brief treatment, for example, how consciousness creates representations of the sensory world. However, such considerations are outside the scope of this paper, in that the focus of the paper is *whether or not* a machine can think, not *how* the brain thinks.

[10] NAND and NOR gates have their basis in Boolean algebra and derive from the original AND and OR functions as defined by that algebra.

## Intentionality and intentional states

In the 'human computer', there is a programming level that is intrinsic to the system and acts causally. This is so because we have consciousness and 'know' what we are doing when following an algorithmic rule in some computation; that is, we possess intentionality, which enables us to evaluate a set of rules and follow them consciously. No such thing occurs in a digital computer. It does not follow rules as such. It might appear to be following rules, but it is merely behaving as it was designed to behave. I suppose one could argue that the human brain is also behaving as it has evolved to behave. From a reductionist viewpoint, there is truth in this argument. However, few would deny that we posses intentionality.

The term *intentionality* goes back to the scholasticism of the Middle Ages, it derives from the Latin *intendere*, which means to point/aim at or extend toward. In 1874, Franz Brentano revived the term (Brentano, 1874/1995) and claimed that intentionality is the defining distinction between mental and physical states. He argued that all, and only, mental phenomena exhibit intentionality.

Brentano further argued that intentionality is an irreducible feature of mental phenomena and, since no physical phenomena could exhibit intentionality, mental phenomena could not be a species of physical phenomena. This has become known as *Brentano's thesis*. The essence of this thesis is that mind cannot be the brain. Brentano used the term *intentional existence*, meaning that the objects of one's perception are in the mind. This is clearly a dualistic view, although Brentano did not necessarily go so far as Descartes.

Without getting caught up in the dualism of Brentano's thesis, I consider that thought and intentionality seem very closely intertwined. To think implies intentionality and to possess intentional states implies thinking. Intentional objects can be objects of sense perception, empirically 'out there' in the physical world. Intentional objects can also be abstractions, such as beliefs and values, and entirely imaginary 'objects' such as unicorns and aliens from Epsilon Eridani. Thus, our thought processes entail intentional objects, manipulated by internalised speech patterns. In this, there are syntactical and semantical dynamics, where syntax provides the structure and semantics the content. This is not to argue that intentionality can be realised only in the human brain. This would be a form of species chauvinism and has been countered successfully by functionalist arguments.

Ned Block (1980) identifies three senses of functionalism. The first is simple decompositional functionalism. Functionalism in this sense refers to a research strategy that relies on the decomposition of a system into its components. The whole system is then explained in terms of these functional parts. Secondly, computation-representation functionalism is a special case of decompositional functionalism, which relies heavily on the *computer-as-mind* analogy. Psychological explanation under computation-representation functionalism is 'akin to providing a computer program for the mind' (Block, 1980, p. 171). Thus, mental processes are seen as being decomposable to a point where they can be thought of as processes, which are as simple as those of a digital computer or, similarly, a Turing machine. Lastly, Block identifies metaphysical functionalism. This form of functionalism is a theory of mind that hypothesizes that mental states simply *are* functional states. The metaphysical functionalist claims that mental states are the types of mental state they are because of the causal relations between inputs, outputs and other mental (that is, functional) states of the system, as in a Turing Machine. The physical implementation of the set of functions that implement a mind are irrelevant to what makes something a mind: it is the functional relations that count.

It is feasible to consider that the functionality of the human brain, with its intentional states, can be realised in other structures, such as gas giant planets, quantum foam or even silicon. However, the key issue here is that, regardless of the structure, the same functionalities must exist to be regarded as possessing intentional states.

## Artificial neural networks

As a lead-in into this topic, let us first consider the most important features of the neural networks found in the brain. The brain contains many billions of special kinds of cell: the nerve cells or *neurons*. These cells are organised into a complicated intercommunicating network. Typically, each neuron is physically connected to tens of thousands of others. Using these connections, neurons can pass electro-chemical signals among one another. These connections are not merely *on* or *off.* The connections have varying *strengths,* which allows the influence of a given neuron on one of its neighbours to be very strong, very weak (or absent altogether) or anything in between. Furthermore, many aspects of brain function, particularly the learning process, are closely associated with the adjustment of these connection strengths. Brain activity is then represented by particular patterns of firing activity amongst this network of neurons. It is this simultaneous, cooperative behaviour of many simple processing units that is at the root of the enormous sophistication and computational power of the brain.

Artificial Neural Networks (ANNs)[11] are computers whose architecture is modelled after the brain. They typically consist of many hundreds of simple processing units, which are wired together in a complex communication network. Each unit or *node* is a simplified model of a real neuron, which *fires* (sends off a new signal) if it receives a sufficiently strong input signal from the other nodes to which it is connected. The strength of these connections may be varied to enable the network to perform different tasks corresponding to different patterns of node-firing activity. This structure is very different from that of traditional digital computers. There is no central processing unit (CPU) following a logical sequence of rules; indeed, there is no set of rules or program. The only rule as such is that concerning the threshold at which an input will 'fire' the recipient node. In a sophisticated network, this threshold itself may be some function of firing patterns. Computation is related to a dynamic process of node firings. This structure then is much closer to the physical workings of the brain and leads to a new type of computer that is rather good at a range of complex tasks.

However, although ANNs come much closer to simulating what happens in the cortex of the human brain and can be made to adapt to and learn from their sensory inputs, there is no evidence that they are thinking. The arguments of Searle still apply. There is no consciousness-causing aspect of such a system. There is no inherent intentionality. To take these arguments further, it is now

[11] A good source on ANNs can be found at http://www.cs.stir.ac.uk/~lss/NNIntro/InvSlides.html

necessary to look more closely at what we mean by terms such as *intelligence*, *thinking* and *consciousness*, which I shall do shortly. But, one final word in relation to ANNS. One might want to invoke the modified forms of Turing's original test (TT), such as the TTT and even the TTTT, as mentioned at the beginning of this paper. A sophisticated ANN-based system might well be able to pass the TTT in that it could contain sensory inputs and outputs. However, in this case, it still does not satisfy the demands of strong AI so is still merely simulating human thought. If such an ANN were to pass the ultimate TTTT, then we have a new order of intelligence, where such a system might satisfy Reber's proposal and contain inherent consciousness. This is possible because the TTTT tests right down to the molecular level.

I now wish to move on to the relationship between AI, thinking and consciouness.

[13] Vernor Vinge, as well as being an academic in the Department of Mathematical Sciences, San Diego State University, is an author of science fiction novels (for example, Vinge, 1992).

## Artificial intelligence, thinking and consciousness

What do we mean when we say someone is intelligent? Is it that they are, for example, very good at mathematics or translating foreign languages? These people are certainly good at understanding and manipulating abstract concepts. But what about poets, novelists and musicians? They are clearly intelligent because they are creative. Indeed, intelligence is visible in almost every form of human activity; it is expressed in the ability to adapt, learn new skills, and form complex relationships and societies. Much of this intelligent behaviour appears to be unique to humans and differentiates us from all other species. We might say that all of our abilities and behaviour can be attributed to the fact that we are *conscious*.

This is the domain of cognitive science (CS), which has a long history (its intellectual origins go back to the mid-1950s) that has produced a vast literature (for example, Vygotsky, 1962; Chomsky, 1965; Gardener, 1975; Dreyfus, 1979; Edelman, 1987, 1989, to name but a few sources) in addition to those already mentioned throughout this paper (for example, Turing, 1950; Newell, 1961: Searle, 1980).

Unfortunately, there is no precise, widely agreed-upon definition of the word *consciousness*. However, most of us have an intuitive sense of what is meant by the term. *Consciousness*, or *cognition*, is a sort of awareness of self (self-awareness), of interaction with the world, of thought processes taking place, and of our ability to at least partially control these processes. We also associate consciousness with an inner voice that expresses our high level, deliberate, thoughts, as well as intentionality and emotion. It seems doubtful whether true intelligence can ever arise in the absence of consciousness. Perhaps, one might take the view that intelligent behaviour is the outward sign of a conscious being. If so, any machine that could display human-like intelligence qualities could be said to be conscious. However, this view is fraught with difficulties, as discussed earlier in regard to the Turing Test and Searle's arguments. We can only *infer* consciousness or self-awareness. The best test can only establish that external behaviours lead to the reasonable inference that self-awareness and consciousness lie behind them.

In the debate addressed here, the term that is used throughout the literature is *artificial intelligence*. The word consciousness does not appear very often, if at all, to my knowledge. I think that it is reasonably legitimate to talk of an artificial intelligence, depending on how we define intelligence. To talk about artificial consciousness is a very different story, one that I will come back to later.

Let us, for the moment, adopt a general definition of *intelligence* as an ability or competence to behave in a way that leads to results that are beneficial to the entity emitting the behaviour and to those who might be recipients of the behaviour. This is a broad definition and doesn't limit the behaviours to the problem-solving variety usually associated with intelligence. However, in one view, all intelligent entities are continuously faced with a problem to be solved or a decision to be made (the latter entails the former). Using this view, we can think of intelligence as the capability to solve problems.

Having decided on this definition, there remains the issue of degrees of intelligent behaviour, which must be bound up with levels of ability-competence; for example, an AI system's task might be simply to distinguish between two states, such as on and off. Such a task is so simple, one might argue, that it hardly requires AI because a simple stored program machine could do the task. Perhaps this example anchors the lowest end of the spectrum of ability-competence of intelligent machine behaviour. At the other end of this spectrum, if we constrain ourselves within present-day digital computing and ANN technology, a very complex task might be, for example, assisting a surgeon in the diagnosis of medical conditions and giving advice on curative procedures. At its highest level, such a task would entail the same level of ability-competence as that of a highly experienced and skilled surgeon.

It is not difficult to imagine that, with the advances in AI and ANN technology that are about to emerge today, we will see machine intelligence reach a very high level in the next decade or so. The level reached would not only satisfy Turing's test (TT and TTT) in the sense of behaviour indistinguishable from intelligent human behaviour but be demonstrably of a high order of intelligence. Using the arguments thus far, machine intelligence is a fact and is dependent for its future only on advances in machine technology (hardware, wetware and software). In this sense, we are talking about AI emulating human intelligence. Within this context, there is no reason to put limits on just how intelligent an AI system might become. It is possible that it might well exceed the ability-competence of any human. Already, AI technology is much faster and more accurate than any human cognitive processing. At present, humans are better at dealing with high levels of ambiguity and in creating novelty. But achieving such abilities in a machine is mainly a matter of processing power, entailing techniques such as massive parallel processing and interconnectivity.

So far, we have focused on intelligence, whether machine or human. I think that even the skeptic will admit that AI systems do behave intelligently and that the level of this behaviour is some function of future developments in the technology. If we introduce the concepts of thinking or consciousness and start to consider artificial thinking or artificial consciousness, we have entered

a wholly different domain of discussion and speculation. Let us first deal with thinking, and then move on to consider consciousness.

What is thinking or thought? We all do it and know what it is like to be doing it. While we will avoid getting caught up in the complexities of what came first, speech or thought, thinking is closely bound up with speech and seems to be an internalisation of it. When we think, we use words in our heads. It might be true that some forms of cognitive processing do not necessarily require the use of internalised words. For example, in considering the next move in a game of chess, the player doesn't necessarily use internalised speech. It is possible to think purely spatially in such circumstances, making decisions on the present and possible future relationships of the pieces on the board within the context of the rules of the game. However, even the most advanced chess players are likely to 'talk' themselves through the variety of possibilities facing them and their opponent. The same non-verbal processing could be argued for activities such as manipulating mathematical expressions. There is also a range of cognitive activities that do not normally entail internal speech, for example, sporting activities such as running, swimming or cycling, and musical performance. However, whilst they may not entail internalised speech, they all entail cognitive processing of some kind.

So, it is simplistic to equate thinking with internalised speech. Thinking has a subjective and internal feel to it and is not accessible to an observer. An observer can only infer my internal processing from my external behaviours. For example, if my partner asks if I can recall seeing her car keys and I say, 'Oh, just a moment', then go silent for a minute or more, it is reasonable for her to infer that I am 'thinking' about where I might have seen her keys. She makes this inference on the basis of what she would be doing were I to ask her a similar question. She has no access to my thoughts. And even when I answer that I saw her keys on the kitchen table, she cannot know what actually went on in my head. In fact, in such a case, my thought processing may be so automatic that even I could not report exactly what went on while I was trying to recall where I had seen her keys.

What would an AI system have to be doing to emulate my thought processes, if it were asked a similar question? It is possible even with the present state of AI to design a machine that could be asked such a question and could then come up with a valid answer as long as it had access to the same sensory data that I had had. How would we know that it was thinking in the way that we describe thinking? No matter how sophisticated its 'neural' circuitry might be, how could we know it was thinking? As it is a machine, its most sophisticated processes are reducible to the movement of charge carriers through conductors.

Even if, with the use of sophisticated ANN technology, simulating neurons, it could produce massive levels of interconnectivity, it remains a machine. We might counter with, 'But, ah, that is all we are — biological machines'. This may be a valid argument. However, we are machines with a strong sense of the subjective nature of our own cognitive processes; they are very distinctly ours and entail a measure of self-awareness. The machine, no matter how sophisticated, appears to lack this subjective sense. Even if it claimed to possess it, how could we ascertain this in an empirically testable way? What would constitute a test for machine self-awareness? The Turing test is not an appropriate test as it tests only for the adequacy of the response by criteria for human responding. The Turing test can tell us that, as far as we know, we cannot distinguish between a machine's response and a human's response under the test conditions. It tells us nothing about self-awareness.

In bringing in the notion of self-awareness, we have already started to consider consciousness. In fact, one might regard self-awareness and consciousness as synonymous. While it is possible to imagine a machine whose behaviours and responses are so sophisticated that it appears to be thinking, in the sense that it is cognitively processing data, I find it difficult to credit this same machine with self-awareness. In fact, it would not need to be self-aware even if this were possible. It would be able to function effectively and efficiently without self-awareness. This raises the interesting question of why we humans have self-awareness. What does it add to our ability as a biological machine? In the view of some, our human self-awareness is simply a superfluous byproduct of our neural activity. We do not understand why it came about (but see Reber's proposal discussed: Reber, 1997). It doesn't appear to serve an evolutionary purpose, at least, not in the narrow Darwinian sense of diversifying the gene pool. It appears to have arisen as a result of crossing some threshold of neural organisational complexity (Melnechuck, 1980).

As far as we know, only humans possess self-awareness. Even the most highly evolved of the non-human animals (for example, chimpanzees) do not appear to have self-awareness. They certainly have awareness. They are very aware of their environment and deal with it in very effective ways. But they do not appear to be aware that they are aware. Self-awareness appears to demand a higher level of cortical abstraction than even the primates possess. In this, I may appear to be arguing that, were a machine intelligence to achieve an equivalent level of neural abstraction, it would become self-aware. Perhaps, so. But there remains the problem of establishing this as a fact. What would constitute an empirical test of self-awareness? As far as I know, no such test currently exists. In fact, why would we need it at present? We know that we are self-aware and reasonably infer this to be true of other humans.

The issue of self-awareness should not be confused with the experiencing of *qualia*: the feeling states. Non-human animals experience qualia; of this we can be fairly certain. (Who has not seen an angry dog, or the pleasure of a cat stretching out in the sun?) Thus, the feeling and expression of emotional states is not an exclusive criterion for being human. It might be possible to arrange for an AI system to express emotional states and claim to feel them. But this does not imply self-awareness. However, to claim to feel them is quite different from simply expressing these states. Perhaps this is the clue to a test of self-awareness — being able to claim it. But could we infer the possession of it simply from the claim? In the way that we are able to infer its possession by other humans (we cannot truly know that another is experiencing it), could a self-aware machine reasonably infer the possession of self-awareness in another machine?

It would appear that the issue of self-awareness is bound up with that of human consciousness, as it is implicit in what it is to be conscious. Descartes said, 'Je pense, donc je suis' (*I think, therefore I am*). This was the one thing that Descartes could not doubt, even though he put himself into a state of universal doubt. The key point here, however, is that he could declare this because he was self-aware. The declaration does not imply or demonstrate self-awareness in itself. It means that, because an entity is thinking, that same entity must exist. This is simply a matter of logic. Thus, a thinking machine can be said to exist because it is thinking. However, to be able to make this declaration an entity must be self-aware, by definition. That is, the entity is aware that it is thinking, hence self-aware.

Any test of self-awareness must probe an entity's (animal or machine) ability to operate at a level of abstraction beyond its deterministic data processing; that is, it must be aware beyond its programming. A test of this kind would force the entity into a response that would demonstrate an insight outside the rules of its behaviour. When a cat sees itself reflected in a mirror, its sees another cat. When still a kitten, it tends to look for this 'other' animal (for example, behind the mirror if it is free-standing). However, as the kitten grows to be an adult cat, this tendency disappears. The reflection comes to be taken for granted like so many other things in the adult cat's life. It is accepted as a phenomenon that carries no threat, reward or interest. The experiences of a very young child with a mirror are very similar to those of the kitten. In the same way, as the child grows, the reflection comes to have a different significance. However, the key difference in the case of the growing child is that it comes to see the reflection as itself. It recognises its own image. Thus, self-recognition is entailed as a criterion for self-awareness. It does not come about because the child is given some set of rules about images. It comes about naturally and internally as the child's cognitive processing matures. A point is reached where the child is not only able to differentiate itself from the rest of the world but also becomes self-aware.

There appear, then, to be a number of key issues relating to self-awareness. It would seem that self-awareness is a major aspect of being conscious. Thus, any machine that claims to be conscious must also be self-aware, which entails going outside the algorithms that determine its internal behaviour. Godel's theorems[12] imply that no finite system of algorithms (for example, mathematics) can go outside itself. It is constrained by the rules/logic of its own system; that is, no mathematical or algorithmic system can provide meaning or explain itself. Thus, Godel's theorems are to do with meaning. Any system of algorithms is devoid of meaning. It is simply a set of rules. Here we look again at the difference between syntax and semantics. An AI system may have a wide array of sophisticated syntactical rules, which enable it to perform tasks that fully meet the criteria of Turing's test. However, there is an absence of semantical processing. The machine does not possess a value set by which it can attribute meaning to what it does.

One could argue that our cognitive processes arise purely out of such a set of syntactical rules or algorithms. It is true that these are extremely complex and we are far from knowing them. However, the assumption of a finite rule set is not unreasonable. If this is true, then how are we able to come to self-awareness and attribute meaning to our sense data? Without invoking supra-physical concepts that would attribute to consciousness something beyond neural processing, movement outside a finite algorithm set does not seem feasible. Yet we clearly achieve this. The feasibility seems to come back to the issue of levels of cognitive abstraction, wherein a high enough level of abstraction enables us to move outside the rule set. The implication here is that the rule set does not govern neuronal interaction at the higher levels of abstraction. Or, if it does so govern, then it governs in a qualitatively different way from that of the lower levels.

A hierarchy is implied, at the lowest level of which there are simple rules that must be obeyed. At this level we have the hard wiring of instinctual behaviour, which is characteristic of simple animals, such as insects. The organism is entirely reactive to its environment. At a high enough level of processing, an organism can do more than simply react instinctively to its environment. It is able to analyse situations, attribute priorities and make decisions. However, these are still syntactical and don't entail semantics. At the highest level, the neural structures are abstracted from the sensori-motor functions and can operate entirely independently of the organism's environment. This seems to permit a movement outside the rule set. There are no longer rules so much as guiding principles, which may be obeyed or not. At this level of the hierarchy, values and meaning appear to arise, along with self-awareness.

[12] See the earlier discussion under the Church-Turing thesis.

As stated above, humans have no choice but to infer the presence of self-awareness in other humans. We cannot directly access another's internal processes. We know that we possess self-awareness. We know what it is like to be self-aware, and we make the reasonable assumption that this is true for other humans. There is no test that I can think of that demonstrates self-awareness in an objective manner. It is a subjective property and can only be approached tangentially. Those interested in measuring psychopathology in humans have faced a related problem. Objective tests are unable to probe the subconscious levels, and projective (subjective) tests are used. The idea here is that the patient projects from the subconscious levels, revealing the maladjusted dynamics at that level. However, in this article we are talking about self-awareness, which is not at all subconscious — quite the reverse, in fact.

In the case of an AI system, we do not have to make inferences or approach the problem tangentially. We can empirically observe what is going at the 'neuronal' level. We can make direct measurements of the states of the ANNs or other 'circuitry' that the system employs. It should be possible, at least technically, to make comparisons between an AI system that clearly doesn't have self-awareness and one that does appear to have it. For example, let us say that we have two AI machines, identical in their initial structure and rule set, and both with the potential for self-awareness. Let us further assume that the acquisition of self-awareness is dependent on the degree of environmental stimulation of the machines. Let us say that we expose one machine to a rich environment that is likely to produce self-awareness and keep the other machine in a limited environment that, at best, will encourage only simple, rule-governed behaviours. It should be possible to make an objective comparison of the internal changes in each machine, noting the differences in internal behaviour that match the changes in external behaviour. By doing so, we should be able to see what happens as the environmentally enriched machine acquires what we judge to be self-awareness. Is there evidence that the self-aware machine has moved outside its rule set? If so, what form does this take? Does it question its existing rule set? Does it generate its own rules and values? How does it deal with ambiguity, uncertainty and novelty in its environment? In this way, it is theoretically possible to establish a test for machine self-awareness.

However, despite these possibilities, there remains the vexed question of machine consciousness. Is being conscious synonymous with being self-aware? Many theorists would argue that consciousness is a prerequisite for being self-aware. They might not, however, agree to their synonymy. There appears to be something more to being conscious than simply being self-aware. Consciousness appears to be a meta-property that embraces a range of features, including self-awareness. But, despite these considerations, if a machine demonstrated self-awareness, would we be required to class it as a being with

ethical rights, which included treating it with respect and dignity? For me, this is an even bigger question than the subtleties of consciousness. Much depends on how we define *sentiency*. If we restrict the definition to biological organisms, we cannot really regard a manufactured machine as being sentient, even though such a machine might contain 'circuitry' of an organic nature. However, the word *sentient* doesn't imply life or living so much as awareness; that is, a sentient being is one that is aware and acts on that awareness. If we accept this definition, a machine possessing self-awareness is surely sentient. This being the case, such a machine would be deserving of the same rights and treatment as all other sentient beings; that is, it would be immoral to cause harm or injury to such a machine. Perhaps the term *machine* is problematic in this consideration. It has connotations that seem at odds with sentiency. However, I argue that sentiency is not defined in terms of life or biology. It is a property that any number of structures might have, which obviously includes biological organisms but doesn't exclude non-biological structures.

## The future of AI

It is hard to predict how AI systems will develop over the next decade or so. We have already seen enormous changes within AI over the past six decades or more, these changes appearing to be exponential rather than linear. If this trend continues, we are likely to see even more dramatic changes in the next few decades. At present, despite their impressive capabilities, most AI systems are based on digital processes and owe an allegiance to the basic von Neumann architecture employed by more conventional digital computers. Work is going on with ANNs to simulate the behaviour of human neurons, but it has a long way to go yet. I am sure that we will see a dramatic increase in the power of AI machines into new areas beyond that of the 'expert system', with its data base and inference engine. AI is likely to be applied with increasing effectiveness to the solution of complex problems, such as weather prediction, where the processes are essentially chaotic, and analysing social processes. Whatever the newly emerging capabilities might be, it is fairly certain that AI will play an increasingly important role in our lives. It has the potential to create a dependency that is even more far-reaching and subtle than that we already have on digital machines.

## Summary and tentative conclusions

It is clear to me, following Searle's reasoning discussed above, that digital von Neumann machines could not possess intentional states. There is nothing in such a machine that is intrinsically causal. If thinking machines are to become a possibility, their realisation cannot be in stored program digital machines. At present, the other main avenue of realisation lies in artificial neural networks (ANNs). These devices are presently in their infancy by comparison with the current generation of digital machines (either von Neuman or massively paralleled machines). ANNs have been developed to emulate certain functions of the human brain, such as the recall process in memory, where one particular design is able to recover a complete (perfect) memory when presented with a very partial stimulus (for example, a small portion of text).

To date, ANN technology has been used to emulate only very specific brain functions, such as aspects of memory, as mentioned above, or visual recognition. There is a long way to go before these devices could even begin to approach the complexity of just one structure within the human brain. However, in saying this, I am not saying that this will not one day be the case. Assuming that this is simply a question of sufficient maturation of ANN technology, it would appear to be merely a matter of time before we have truly thinking machines. However, it is not clear, with the current nascent state of this technology, where the threshold of an ANN system's complexity lies. This threshold has to be crossed before thinking occurs. How much of the human brain has to be emulated before this occurs? Can we ignore the sub-cortical structures and functions wholly or in part?

We do not know enough about brain functioning to be able to confidently answer this question today. While it seems reasonably understood that what we regard as thought processes occur mainly within the abstracted cortical regions of the brain, there is a dependency on certain sub-cortical functions. It may be that pure abstract thought, such as in higher mathematics, doesn't entail sub-cortical functions at all. However, this seems to ignore the excitement that goes with operations where the creation of mathematical novelty is involved, which appear to entail at least some degree of sub-cortical functionality. Without reviving the old arguments explored in the theses of James-Lange, Cannon-Bard or Schacter, emotional states are a part of such processing. Thus, it would seem that, even for the most abstract level of thought, some degree of sub-cortical function is necessary. It would entail far more than the emulation of neuronal structures. Other structures exist within the sub-cortical regions of our brains that do not have the more familiar neuron-axon-dendrite structure. Could we

emulate these using ANNs? Seemingly not, because the term ANN implies emulating neurons. So, we would need some other 'machine' element for these sub-cortical structures.

There is a thought experiment attributed, I understand, to Jaron Lanier, the inventor of virtual reality. It goes like this. Let us say that we could systematically and slowly replace the neurons in a human brain with a silicon device (some collection of NAND/NOR logic gates plus a small Read-Only Memory (ROM) program that performed a neuron's functions. Initially, the brain would continue functioning as normal, with its conscious and intentional states intact. However, when most, if not all, of the neurons had been replaced with their silicon counterparts, would there come a point at which consciousness disappeared? People like Searle would argue that consciousness would disappear because there is nothing in each pseudo-neuron that looks remotely like consciousness. Thus, replacement will not yield consciousness, because quantity is not quality. Strong AI adherents would disagree.

This thought experiment raises some interesting issues in the thinking machine debate. As the replacement is occurring slowly and systematically at each stage, we can assume that there is no massive disruption to normal brain functioning. The strong AI adherents argue that complete replacement will not lead to any difference in brain functioning, leaving consciousness and intentional states intact. The person undergoing this 'operation' will not notice anything different. The weak AI adherents, however, seem to be arguing from a basis of complexity, in which they envisage that, at some threshold of complexity, consciousness (along with thinking) will disappear. Weak AI would seem to say that, initially, replacement would have no effect. This is reasonable, because how could one pseudo-neuron have much of an effect? But, what about ten or a hundred or ten thousand replacements? I suspect that weak AI would say that it is not just quantity but complexity, which implies structure.

However, there is more to the weak AI argument than this, because adherents say that there's nothing in the pseudo-neuron that is causal. So replacement will eventually disrupt brain functioning. But this argument implies that there is something causal in our individual neurons. Surely, at this level of reduction, there is no causality, intentionality or consciousness? If there is, where does consciousness kick in? Is it simply an issue of crossing some level of complexity, probably at the cortical level? If it is, then why does replacement not work? The implication is that there is some mysterium in the human brain that gives rise to consciousness (see the earlier discussion about Reber's (1997) views).

David Chalmers (1996) has dealt with this issue, where he argues that the mind-body problem has easy and hard parts. The easy part deals with the psycho-technical dimension (this includes psychology, as in cognitive studies), whereas the hard part deals with the phenomenological dimension (the philosophical aspect, the mind-body problem in particular). Chalmers is saying that mind is not reducible to brain; that is, the subjective element cannot be explained in terms of neurons firing. Mind does not supervene on matter. It is something wholly different. Thus, in Chalmers view, there is a mysterium that cannot be explained in terms of brain physiology. If this is true, then Searle is right, and the case for strong AI is doomed. No matter to what level of sophistication we might take AI machines, we shall never be able to impart the mysterium.

While I accept that, within the mind-body debate, the first person, subjective sense of being me cannot be explained in terms of neurons firing, there are no arguments that persuade me that mind is something other than an aspect of the brain's functioning. It may well be that mind is holistic in the sense that it is greater than the sum of its parts. We cannot pin mind down to any specific location within the brain. It appears to be all-pervasive. But this is not to say that it doesn't arise from brain functioning. I believe that mind is a part of the natural order of things, in that it has arisen within a physical organ. It seems likely that, with the maturation of sciences such as neuropsychology, we shall one day come to understand just what mind is. If this is true, then we might also produce machines that think. As I fully agree with Searle's weak AI argument in relation to digital von Neumann-type machines, the platform for thinking machines is going to be something wholly other. What kind of technological platform these machines will have remains pure speculation at present, but it seems likely that it will be quasi-biological (see the earlier discussion on the views of Reber, 1997).

While the Turing test (whether TT, TTT or TTTT) is very severe and could lead to a combinatorial explosion in even the most powerful present-day digital AI machine, it does not test for thinking. It is a behavioural test, merely testing for the simulation of thinking. Thus, it does not really serve the thinking machine debate, and success in it does not support the strong AI position. Searle, for me, has captured the problems inherent in the strong AI argument, showing that the digital machine does not possess even syntax, never mind a semantic dynamic. If thinking machines are to be realised, then the technological platform must be something other than a digital von Neumann machine. Artificial neural networks might be a possible platform. However, at present, ANN technology is in its infancy. Some (for example, Chalmers, 1996) imply that mind/consciousness has a mysterium component, further implying that AI machines, of any variety, can never be conscious or think. I do not share his pessimism. While mind and consciousness remain in a very hard category, I believe that they arise from the physiological substrate of the brain. If this is true, then we might approach the thinking machine via a quasi-biological platform.

One final piece of speculation. If we ever do create truly thinking machines, consider some of the ethical and other consequences. These machines will, by definition, be sentient. This means that they will have rights. We, as their creators, will have a special relationship with them and a high degree of responsibility. It seems likely that such machines will ultimately surpass us in cognitive ability. Vinge[13] (1993) has touched on this possibility in his notion of singularity, where he wonders how these far superior machines will regard us and treat us. Will they hold us in the same regard that we hold our pets? Life might not be too bad, were this the case. However, what if we were to them as ants are to us: for the most part a nuisance? Could we expect benevolence? Are benevolence and compassion the obvious outcome of a far superior intelligence? Perhaps the compassion we show to other species is some function of how close we are to them on the phylogenetic scale. We can feel a kinship with warm-blooded mammals as we are of that phylum. But we do not feel this same level of kinship with insects, or even reptiles. Yet, all these phyla are of the order of animals. A sentient machine is of a wholly different order of sentience, so is it likely to feel any kinship with us whatsoever? Most likely not, even where it is fully aware that we were the creator of its forebears. If this day ever comes about, then perhaps the best we can expect is a benevolent retirement, in which we can face final redundancy as a species!

These speculations complete the circle, taking us back to my opening comments regarding the prophetic role of science fiction. Science fiction is predicting that intelligent and conscious machines will make their appearance in a few decades. If this is likely, we current *thinking machines* might well partake in this, by a process of gene engineering and nanotechnological neural implants, in order to share in that post-humanity.

[13] Vernor Vinge, as well as being an academic in the Department of Mathematical Sciences, San Diego State University, is an author of science fiction novels (eg, Vinge, 1992)..

# References


Baum, E. B. (2004). *What is thought?* Boston, MA: The MIT Press.

Block, N. (1980). Introduction: What is functionalism? In N. Block (Ed.), *Readings in philosophy of psychology*. Cambridge, MA: Harvard University Press.

Block, N. (1998) The philosophy of psychology: Classical computationalism. In A. C. Grayling (Ed.), *Philosophy 2: Further through the subject*. New York: Oxford University Press.

Brentano, F. (1995). *Psychology from an empirical standpoint* (Rancurello et al., Trans.) London, England: Routledge. (Original work published 1874)

Chalmers, D. (1994). On implementing a computation. *Mind and Machines, 4,* 391–402.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford, England: University Press.

Chomsky, N. (1965). *Cartesian linguistics*. New York: Harper and Row.

Church, A. (1936). An unsolvable problem in elementary number theory. *American Journal of Mathematics*, *58*, 345–363.

Dennett, D. C. (1988, Winter). When philosophers encounter AI. *Daedalus: Proceedings of the American Academy of Arts and Sciences*, 283–295.

Dennett, D. (1990). *The age of intelligent machines: Can machines think?* Retrieved November, 2004, from http://www.kurzweilai.net/articles/art0099.html

Dennet, D. (1995). 'The unimagined preposterousness of zombies': Commentary on T. Moody, O. Flanagan and T. Polger. *Journal of Consciousness Studies*, *2*(4), 322–326.

Dreyfus, H. L. (1979). *What computers can't do*. New York: Harper Colophon Books.

Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books.

Edelman, G. M. (1989) *The remembered present: A biological theory of consciousness*. New York: Basic Books.

Fodor, J. (1991). *Afterthoughts: Yin and Yang in the Chinese Room*. In D. M. Rosenthal (Ed.), *The nature of mind*. New York: Oxford University Press.

Gardner, H. (1975). *The shattered mind*. New York: Knopf.

Gödel, K. (1965). On undecidable propositions of formal mathematical systems. In M. Davis (Ed.), *The undecidable*. New York: Raven.

Harnad, S. (1990). The symbol grounding problem. *Physica D, 42*, 335–346.

Hauser, L. (1997). Searle's Chinese box: Debunking the Chinese Room argument. *Minds and Machines*, 7, 199–226.

Jacquette, D. (1989). Adventures in the Chinese Room. *Philosophy and Phenomenological Research*, *49*(4), 605–623.

Jacquette, D. (1990). Fear and loathing (and other intentional states) in Searle's Chinese Room. *Philosophical Psychology*, *3*(2/3), 287–305.

Kalonder, J. (1983). Reconstructive memory: A computer model. *Cognitive Science, 7*, 281–328.

McCarthy, J. (1956). Inversion of functions defined by Turing machines. In C. Shannon & J. McCarthy (Eds.), *Automata studies*. Princeton, NJ: Princeton University Press.

Melnechuck, T. (1980, October) *Consciousness and brain research*. A paper presented at the Spring Hill Conference on the Assessment of Consciousness Research, Wyzata, Minnesota.

Millican, P. J., & Clark, A. (Eds.). (1996). *Machines and thought: The legacy of Alan Turing*, (Vol. 1). New York: Open University Press.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Newell, A. (Ed.). (1961). *Information processing language V manual*. Englewood Cliffs, NJ: Prentice-Hall.

Penrose, R. (1989). *The emperor's new mind*. Oxford: Oxford University Press.

Preston, J., & Bishop, M. (Eds.). (2002). *Views into the Chinese Room: New essays on Searle and artificial intelligence*. New York: Open University Press.

Quine, W. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Reber, A. S. (1997). Caterpillars and consciousness. *Philosophical Psychology, 10*(4), 437–350.

Rorty, R. (1980). Searle and the special powers of the brain. *Behavioral and Brain Sciences*, *3*, 445–446.

Searle, J. (1980a). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417–424.

Searle, J. (1980b). Intrinsic intentionality. *Behavioral and Brain Sciences*, *3*, 450–456.

Searle, J. (1984). *Minds, brains and science*. London: British Broadcasting Corporation.

Searle, J. (1990). Is the brain's mind a computer program? *Scientific American*, *262*(1), 26–31.

Searle, J. (1991). Minds, brains, and programs. In D. M. Rosenthal (Ed.), *The nature mind*. New York: Oxford University Press.

Searle, J. R. (1997). *The mystery of consciousness*. New York: New York Review of Books.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *Series 2*, *42*, 230–265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*.

Vinge, V. (1992). *A fire upon the deep*. London, England: Millenium.

Vinge, V. (1993). *The technological singularity*. Retrieved November, 2004, from http://www.kurzweilai.net/articles/art0092.html.

Vygotsky, L. (1962). *Thinking and speaking*. Boston, MA: MIT Press.

Wakefield, J. C. (2003). The Chinese Room argument reconsidered: Essentialism, indeterminacy and strong AI. *Minds and Machines*, *13*, 285–319.

Wiener, N. (1972). *Cybernetics: Or, control and communication in the animal and the machine* (3rd ed.). Cambridge MA: MIT Press.