ARTIFICIAL INTELLIGENCE AND OTHER MINDS

The Search for Strong AI

Oskari Vesterinen

Master's Thesis

Philosophy

Department of Social Sciences and Philosophy

University of Jyväskylä

Autumn 2019

ARTIFICIAL INTELLIGENCE AND OTHER MINDS
The Search for Strong AI

Oskari Vesterinen
Master's Thesis
Philosophy
Department of Social Sciences and Philosophy
University of Jyväskylä
Supervisor: Mikko Yrjönsuuri
Autumn 2019
Page count: 63

ABSTRACT

The purpose of this thesis is to investigate the possibility of conscious artificial intelligence, or Strong AI. The thesis approaches its research question through three main perspectives: the Turing Test by Alan Turing, the Chinese room argument by philosopher John Searle, and embodied cognition, each supplemented by academic philosophical discussion. The Turing Test represents an early attempt to empirically measure the consciousness of machines, but proves to be inadequate for its intended task. The Chinese room argument is a classical thought experiment which supposedly refutes Strong AI, and, while controversial, provides strong objections against conscious AIs as disembodied, symbol-based programs. Embodied cognition is approached first through the "brain in a vat" thought experiment, and then through a more practical approach which may facilitate philosophical advances over the Chinese room argument.

The central finding in analyzing these perspectives is that the possibility of Strong AI cannot be conclusively proved or disproved under our current, imperfect understanding of cognition and consciousness. Main points that hamper definite conclusions are that that we do not know the necessary boundary conditions for consciousness, and that intelligently behaving AIs or machines possess the problem of other minds in determining whether they are conscious or not.

TIIVISTELMÄ

Tämän tutkielman tarkoituksena on tutkia tietoisen eli ns. ”vahvan” tekoälyn mahdollisuutta. Tutkielma lähestyy tutkimuskysymystään kolmen keskeisen näkökulman kautta: Alan Turingin esittämän Turingin testin, filosofi John Searlen esittämän kiinalaisen huoneen argumentin, sekä kehollisen kognition kautta. Turingin testi edustaa varhaista yritystä mitata koneiden tietoisuutta empiirisesti, mutta osoittautuu tähän tarkoitukseen riittämättömäksi. Kiinalaisen huoneen argumentti on klassinen ajatuskoe, joka väitetysti kumoaa vahvan tekoälyn mahdollisuuden, ja joka kiistanalaisuudestaan huolimatta tarjoaa vahvoja vastalauseita tietoisen tekoälyn suhteen sikäli kun ne ymmärretään kehottomina, symbolipohjaisina ohjelmistoina. Kehollista kognitiota lähestytään ensin tunnetulla ”aivot vadissa” -ajatuskokeella, ja sen jälkeen käytännöllisemmän lähestymistavan kautta, joka voi tuottaa filosofisia edistysaskelia kiinalaisen huoneen argumentin suhteen.

Keskeinen löytö näistä näkökulmista on, ettei vahvan tekoälyn mahdollisuutta voi varmuudella osoittaa mahdolliseksi tai mahdottomaksi nykyisen kognition ja tietoisuuden ymmärryksemme pohjalta. Varmoja johtopäätöksiä rajoittavia tekijöitä ovat ensinnäkin se, ettemme tunne tietoisuuden välttämättömiä reunaehtoja, sekä toiseksi se, että yrittäessämme määrittää sitä, ovatko älykkäästi käyttäytyvät tekoälyt tai koneet tietoisia vai eivät, törmäämme muiden mielten ongelmaan.

Keywords: artificial intelligence, Strong AI, artificial consciousness, Turing test, Chinese room argument, embodied cognition

Table of Contents

1. INTRODUCTION

It is good for any article to catch the attention of its readership early. A good way is to start out with zombies.

In philosophy, a *philosophical zombie* is human which is outwardly indistinguishable compared to an ordinary human being, but which does not possess any subjective mental experiences or inner life. Such a "p-zombie" would laugh when it hears a joke, cry out if it is stabbed, but never experience joy, pain, or anything else; it is mentally dead, a "hollow man". No one from the outside could ever know this, as the p-zombie behaves identically to any other human being, exhibits identical brain states, and so on.

P-zombies have seen most of their use as vehicles to philosophical thought experiments, and few philosophers entertain the possibility of their actual existence seriously. Who would really be prepared to take the solipsist position that other people are empty machines with no thoughts or feelings?

However, the rapid technological advancement of the human race has brought an unexpected twist to the concept of p-zombies. For the first time in human history, we have managed to create artificial entities which exhibit limited forms of intelligent and autonomous behavior. While the field of *artificial intelligence* is still a far cry from the superintelligent villains of Hollywood blockbusters, a boom of constantly advancing AI research has raised the age-old questions from the philosophy of mind to the forefront of contemporary philosophy. It may be the case that human beings are not philosophical zombies, but what is the mental status of AIs – and how would we know what it is?

This thesis is concerned with investigating one primary question: could an AI – or an embodied machine, if the difference is meaningful – possess a subjective consciousness under some conditions, or is this altogether impossible? By subjective consciousness I mean subjective, phenomenal mental states, or *qualia*. In this thesis I will use also use the word "consciousness" as a synonym, although I acknowledge that consciousness can be understood in different ways as well. It goes without saying that this formulation of the research question presupposes the existence of qualia in the first place; I take this assumption as a reasonable starting point despite occasional attempts in philosophy to sidestep or discredit the concept.

Anyone who follows the field of AI can tell that even the most modern AIs of the present day are not conscious entities. While they can exhibit intelligent and even humane behavior under some circumstances, their cognitive architectures are designed to address individual, clearly defined objectives and not intelligence in general. They are, at best, philosophical zombies. The question is thus targeted towards as-of-yet conjectural, more general forms of AI which would possess broader intellectual capabilities and be closer to human intelligence. While these kinds of advanced AIs do not exist today, the rapid advances in AI research suggest that they might become reality in this century.

Why is it important to know whether AIs or machines can possess a subjective consciousness or not? The reason is that there are serious societal and ethical questions that go side by side with conscious beings. What would a society of not one, but two "sapient" species look like? Machines are people's property – but can this be justified if the machine possesses a mind? The treatment of AIs would suffer from a severe clash of two paradigms: one where they are tools and property of their owners, and one where they suddenly possess a subjective mind and, quite possibly, deserve or demand legal rights. The *moral status* of a conscious AI is ambiguous, and this is why it is important to know whether or not conscious machines can exist in the first place.

The question on how to prove the existence of artificial minds currently lacks an empirical, scientifically testable solution, and this thesis is not meant to provide one. Our knowledge of the mechanisms behind consciousness are too incomplete to venture a response that would be beyond any reasonable doubt. In current AI research the question is not very central at all; contemporary research has mostly focused on (highly successful) utilization of AIs on individual tasks, with no overarching goal of creating conscious artificial entities. However, despite the relatively peripheral[1] role of the topic of this thesis, I believe it is beneficial to sketch out options and to explore what could perhaps enable an artificial consciousness – or what could certainly *not* enable it.

The thesis is divided into a short chapter which will explain the conceptual foundations of the topic at hand, and into a longer one which will explore the question through three perspectives. The first one is the Turing Test, proposed by Alan Turing in 1950, which provides an early attempt to answer the question of artificial consciousness. The Turing Test is arguably the most famous empirical experiment to address the question, and as such it is a natural

[1] AI-wise, not philosophy-wise.

starting point to start our treatment of the topic. The second perspective is the so-called Chinese room argument by philosopher John Searle, which is perhaps the most famous objection against conscious AIs. Searle's argument is a philosophical refutation of the validity of the Turing Test, and it has received tremendous amounts of scholarly attention since he first proposed it in 1980. Searle's thorough treatment will take up a sizeable portion of the thesis. The third perspective is embodied cognition, a sub-field of cognitive science which carries interesting implications that could possibly be used to bypass much of the criticism fielded by the Chinese room argument. Embodied cognition is exemplified here through two viewpoints: the "brain in a vat" thought experiment, and by Professor Pentti Haikonen, who has proposed a connectionist, associative neural model of consciousness that is applicable to machines. Each of these perspectives is complemented and criticized by a number of other thought experiments and academics as deemed appropriate. Lastly, I will sum up the conclusions from the chapters at the end of the thesis.

The methodology of the thesis will consist of systematic philosophical concept and argumentation analysis. The research material will consist of previous academic literature on the subject, supplemented with other literary sources when they are considered to have argumentative power to the question at hand.

2. TERMINOLOGY OF ARTIFICIAL INTELLIGENCE

The term "Artificial Intelligence" was first coined in a DARPA-sponsored conference at Dartmouth College, New Hampshire, in 1956. This conference is seen as the opening shot of the scientific research of AI, although the concept of "intelligent" or "thinking" machines is somewhat older than that – for example, Alan Turing posed the question of thinking machines in 1950 in an article where he introduced his famous "Turing Test" (Selmer Bringsjord & Naveen Govindarajulu, 2018, chap. 1). The exact definition of AI, however, has proven to be somewhat elusive. On the face of it, the term might appear self-explanatory, but the notion of "intelligence" in relation to AI has proven to be a divisive issue.

Stuart Russel and Peter Norvig, writers of the influential book *Artificial Intelligence: A Modern Approach*, have noted the presence of discord in defining AI. Russel and Norvig summarize a quartet of categories that cover most competing definitions given over the years, and note that in most definitions AIs are broadly understood as systems that either:

a) Think (i.e., reason) like humans,
b) Act like humans,
c) Think rationally, or
d) Act rationally (Russel & Norvig, 2016, p. 2).

All four types of definitions have been – and to an extent, still are – followed in AI research. Definitions a) and b) approach AI's "intelligence" through a human benchmark, while c) and d) do so by ideal rationality – that is, perfectly rational reasoning or behavior in relation to any specific goal.

Having mapped this range of potential definitions, Russel and Norvig also provide their personal proposal with the notion of an *intelligent agent*. Their definition of AI is "the study of [artificial] agents that receive percepts from the environment and perform actions." Their conception of intelligent agent is also a *rational agent* – it always acts to achieve the best expected outcome. Russel and Norvig consider the "acting rationally" definition of AI to be the most fruitful approach to follow; the human-based approach would limit AI to human skills, while "thinking rationally" would limit AI to only one formal method of achieving rationality. (Russel & Norvig, 2016, pp. viii, 4–5.)

Russel's and Norvig's definition is a good one for AI research, since emphasis on rationality is beneficial if we are aiming at creating AIs that are applied to perform certain tasks. After all, a maximally rational AI would probably get the job done more efficiently than a "humane" one. But if we are concerned with the hypothetical *mind* of an AI, it may be relevant to ask if the "maximum rationality" approach will lead to the expected conclusion. If we want conscious AIs, we may want to follow a more human-centric approach.

In their article on AI in the *Stanford Encyclopedia of Philosophy*, Bringsjord and Govindarajulu provide a definition of AI where AI is "the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons)." (Bringsjord & Govindarajulu, 2018). Their definition fits closer to the topic of this thesis, since we are interested in AIs as conscious entities and not as mere problem-solving tools. Bringsjord and Govindarajulu are of course not claiming that all AI research is literally attempting to create "artificial humans". Instead, they argue that AI researchers are essentially attempting to build artificial correlates of "naturally" occurring intelligence – human or animal – even if in limited form (Bringsjord & Govindarajulu, 2018, footnote 1).[2]

If attempting to produce an artificial conscious entity, it would naturally make sense to follow a familiar template. Without knowing exactly how a human-like consciousness is born, the best way to reproduce that consciousness would be to rigorously replicate human cognitive processes. Such a design philosophy would most likely follow a human-centric approach, although it is not impossible for a maximally rational AI to theoretically possess subjective mental states. Ron Chrisley has noted that there is a distinction between reproducing *human* consciousness and consciousness *in general*, and that there might be different kinds of consciousness that could be attained by different cognitive architectures (Chrisley, 2008, pp. 124, 133). Naturally, these questions remain irrelevant as long as AIs remain as the relatively simple tools they are today, but they acquire new-found relevance if more complex and general-purpose AIs are developed in the future.

Regardless, research into AI is conducted even if there is no semantic consensus on what kind of intelligence exactly is being sought. It should be noted that conscious AIs are hardly

[2] This might in fact bring Bringsjord and Govindarajulu closer to Russel's and Norvig's definition of AI, depending on how they understand the term "intelligence" in relation to rationality.

at the center of contemporary AI research, which focuses on producing more practical solutions in clearly-defined domains. As such, most AIs are not of interest to us in our context. A relevant distinction is found in the difference between *Narrow* AI and *AGI*. We will take short a look at both.

2.1 Narrow AI and AGI

The AIs that we currently have in use today are designed to work on a very limited set of tasks. They can be intelligent, even superhuman, in their own fields, but woefully incapable of anything not related to the tasks they were programmed to do. An AI designed for a certain specific task is called a *Narrow AI*; IBM's Deep Blue system that beat the world chess champion Garry Kasparov in 1996 is an example. (Sam Adams et al., 2012, p. 26.) Narrow AIs have been developed for a wide variety of tasks, such as diagnosing diseases, trading shares for optimal prices, or playing games. Indeed, virtually all AIs currently in use can be considered Narrow AIs.

Since the cognitive capabilities of Narrow AIs are limited to individual tasks, few people are willing to attribute any sort of subjective mental states to them. Consciousness remains quite distant from this sort of AI, and as such, we are not interested in Narrow AIs in this thesis. In order to be considered as a candidate for conscious being, an AI would have to possess far broader cognitive capabilities. Such an AI is known as *Artificial General Intelligence*, or *AGI*.

An AGI is an AI with broad, approximately human-level cognitive capabilities. An AGI would not be limited to performing one specific task in its design; it would be able to successfully operate in a wide range of scenarios and learn new ways of thinking or processing information over time. An AGI's intellectual abilities could theoretically rival or even greatly surpass a human being, and it is indeed seen as the final destination of the research in the field. (Adams et al., 2012, pp. 26, 28.)[3] Human-level cognitive capabilities mean that the question of consciousness becomes more interesting with AGI than with Narrow AI.

[3] However, having human-level cognitive abilities does not mean that an AGI's way of thinking must necessarily be identical to that of a human, as Russel and Norvig's quartet of possibilities above suggest.

It is important to empathize that no AGI exists yet, and the exact requirements of AGI pose some interesting questions. What kind of cognitive capabilities are sufficient for an AI to be considered an AGI? Do its abilities need to cover all levels of intellectual enterprise, even if shallowly? Do its abilities need to reach human levels, or is sub-human performance acceptable in some areas?

In response to some of these questions, Adams et al. use a number of cognitive architecture requirements for AGI. The criteria they propose are as follows (*R* stands for *requirement)*:

R0. New tasks do not require re-programming of the agent
R1. Realize a symbol system
Represent and effectively use:
R2. Modality-specific knowledge
R3. Large bodies of diverse knowledge
R4. Knowledge with different levels of generality
R5. Diverse levels of knowledge
R6. Beliefs independent of current perception
R7. Rich, hierarchical control knowledge
R8. Meta-cognitive knowledge
R9. Support a spectrum of bounded and unbounded deliberation
R10. Support diverse, comprehensive learning
R11 Support incremental, online learning (Adams et al., 2012, p. 27.)

Adams et al. further supplement this list with a number of elements for the environment and tasks that an AGI is expected to perform in. *C* stands for *characteristics*:

C1. The environment is complex, with diverse, interacting and richly structured objects.
C2. The environment is dynamic and open.
C3. Task-relevant regularities exist at multiple time scales.
C4. Other agents impact performance.
C5. Tasks can be complex, diverse and novel.
C6. Interactions between agent, environment and tasks are complex and limited.
C7. Computational resources of the agent are limited.
C8. Agent existence is long-term and continual. (Adams et al., 2012, p. 27.)

Adams et al. do not intend these criteria to be definite or unchangeable, but rather as starting points to elicit discussion among AGI researchers. They also note that a common list of

architectural requirements makes it much easier for researchers in different projects to compare and duplicate the results of independent research teams. (Adams et al., 2012, pp. 26–27.)

While some of the terms mentioned above may invite philosophical nitpicking (can an AI really have "beliefs", or what exactly is meant with that term in R6), they provide a rough idea on what kinds of cognitive abilities an AGI should be expected to hold.[4] Note that all criteria provided above are compatible with both human-centric and rationality-centric approaches to AI. Both human and rational approach could thus be used in developing AGI, and, presumably, the choice would be affected by what we want the AGI for. If we want some kind of practical "universal problem solver", the rationality angle would most definitely be the more useful one. But if we want to create an artificial human mind (to analyze human cognition, for example), the human angle would naturally be our best bet. This is not to say that a rational agent cannot be conscious, or that a humane one would not be of any use in solving problems. It only means that fundamental philosophical questions related to AI may be addressed or ignored depending on the designers' motives for building an AGI.

Nevertheless, not all AIs are relevant as objects of philosophical consideration. Simple, Narrow AIs are tools, and there is very little that would or should make us think otherwise. It is AGI – whatever its exact design angle may be – where the deeper philosophical questions of consciousness come into play, and where we will focus our study from now on.

2.2 Strong and Weak AI

Based on our current understanding, there is nothing in theory that prohibits the creation of AGI. Even if our current technical understanding in the field is somewhat lacking, research into AGI is and has been on-going for several years. We cannot exclude the possibility that we could one day design an AGI with broadly human-like behavioral characteristics – one that talks like us, expresses emotions and thoughts etc. Such a scenario would raise a host of moral and philosophical questions. Does an AGI really feel pain or happiness if it claims so?

[4] The list is not intended to be exhaustive.

How do we know that an AGI really *feels* the emotions it claims to feel, instead of simply simulating a response that we mistake for the "real thing"?

Philosopher John Searle has made a distinction between what he calls "Strong AI" and "Weak AI" when it comes to this problem. A "strong" AI, according to Searle, is an AI that really feels, understands, and has cognitive states comparable to human thinking – the AI really is a *mind* and not just a mere simulation of one (Searle, 1980, p. 417).[5]

Searle contrasts the concept of a Strong AI with a "weak" one. Searle's original definition of Weak AI from his 1980 article is somewhat vague, as he merely states: "According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion." (Searle, 1980, p. 417.)

What Searle means here is that a Weak AI is essentially a *simulation* of a human mind, useful for analyzing human cognition, but lacking an actual consciousness (Bringsjord & Govindarajulu, 2018, chap. 8.1). In other words, a Weak AI is a philosophical zombie of a kind: human-like in appearance, or at least in behavior, but with no inner life or qualia.[6]

Since simulating or emulating a whole subjective consciousness would presumably require broad cognitive capabilities, both Strong and Weak AIs could be considered AGIs. This raises the central question of this thesis: *can an AGI be a Strong AI, or is it necessarily a Weak AI (i.e. can a Strong AI exist?)* In other words, can an AGI be equipped with all human cognitive faculties, including subjective consciousness, or can it merely mimic them?[7]

This question is difficult to answer, since to an outside observer both Strong and Weak AIs would appear to be identical in behavior. There would be no way to know which is which – after all, at the moment we cannot even directly ascertain that other human beings possess a subjective consciousness! The distinction between Strong and Weak AI therefore runs into the problem of other minds – a problem with no clear-cut answer in sight. It is not possible to look into the "head" of an AI (or any other being, for that matter) and conclude just by

[5] Searle himself is highly skeptical that a Strong AI could exist at all, at least in the framework of our modern, digital computers. His views will be discussed in more detail in the Chinese room argument sub-chapter.

[6] A small terminological ambiguity should be addressed here. When discussing AIs with a consciousness, different terms are often used: artificial consciousness, artificial mind, thinking AI, AI with qualia, AI with a subjective/phenomenal consciousness etc. For the purposes of this thesis, I will consider these terms synonymous with each other.

[7] In popular literature the terms are sometimes used as each other's synonyms; Strong AI and AGI are used synonymously, and likewise are Narrow and Weak AI. However, this kind of usage is somewhat care-free to the implications the respective terms carry.

looking at it whether it has truly conscious, subjective experiences or not. Our means of exploring that question must therefore be indirect – even if that means that our answers must, for the time being, be imperfect as well.

3. AI WITH A SUBJECTIVE CONSCIOUSNESS

While a standard trope of science fiction, Strong AI is nowhere to be seen in the real world, and the philosophical debate still rages on whether it is even possible for such an entity to exist. A central issue in answering this question is one of the most nebulous problems in the philosophy of mind: the problem of other minds, namely, how can we know that other people (or other agents) experience subjective mental states or *qualia*. No scientific theory at the moment provides a verifiable explanation for it. No test has been devised to indisputably prove it. Yet our subjective experiences are there, so an explanation for them must exist as well. Whatever this explanation may be, it might also answer the open question of "can a Strong AI exist?"[8]

Sadly, we do not know yet how to explain the mechanisms behind subjective experiences. It is reasonable to believe that other human beings possess a mind similar to our own, but with machines or AIs the assumption cannot be made as lightly. Since we do not know the exact mechanisms behind consciousness, it is difficult to determine whether AIs can be ruled in or out from it. Because of this uncertainty, we have to make do with what we have. Even with incomplete understanding we can investigate certain tests, thought experiments or hypotheses that could imply the presence of a mind, and likewise, we must consider the positions that would limit or even altogether deny its possibility. The rest of this chapter explores some of these positions.

3.1 The Turing Test and the Behavioristic Approach

The Turing Test is possibly the most famous thought experiment that addresses the philosophical status of intelligently behaving AIs. The test was first proposed by Alan Turing in his paper "Computing Machinery and Intelligence" in 1950, although similar thought experiments had been explored by as early thinkers as René Descartes (Graham Oppy & David

[8] Of course, there are certain philosophical positions which maintain that our minds are somehow dualistic or mental in nature, with no physical phenomena capable of explaining them. However, given that these positions require a bloated ontology veering closer to religion and mysticism, and at the very least offer little prediction power and falsifiability, I consider it reasonable to exclude them from this review.

Dowe, 2016). Turing believed that the question "can machines think?" was too vague to deserve further thought as it stood. He wanted to replace the question with a more specific one: could a machine pass for a human being in a standardized behavioral test. (Turing, 1950, p. 433.) If the answer was "yes", then machines could be considered to think.

An important side note is that Turing never uses the word "conscious" when describing his test – he only speaks of "thinking" and "intelligent" machines. However, the Turing Test has often been interpreted as a more general test for the presence of a mind (Oppy & Dowe, 2016), and this is the interpretation I will be following here.

Turing begins his experiment by describing what he calls an "imitation game". In this game, three people participate; a man, a woman, and an interrogator. The interrogator can only communicate with the other two by standard text input, and tries to determine who is who by asking them questions. Turing then asks what will happen if a machine takes the place of the man or the woman. If the interrogator mistakes the machine for a human being as often as he does the man for the woman or vice versa, the machine can be considered to "think", and has passed the test.[9] (Turing, 1950, pp. 433–434.) A more commonly heard version skips the man/woman aspect of the test and simply asks if the interrogator can distinguish the machine from the human participant.

The limitations of the Turing Test are apparent immediately after we make the distinction between Weak and Strong AIs. The test is purely behavioristic in nature, and does not concern itself with the inner life, or lack thereof, of an AI. The focus is entirely on the outward behavior. The immediate problem is that either kind of AGI, Weak or Strong, could pass the test with equal ease and have consciousness attributed to them. While the everyday distinction between the two might be non-existent, their philosophically fundamental difference would be unacceptable to many philosophers.

Despite its obvious limitations, the Turing Test has not sunk into oblivion in the almost seventy years since it was first proposed by Turing. Why is that?

A simple answer is that we have no better means to study consciousness than by observing the outward behavior of the agent. We have no direct access to the subjective experiences of

[9] Turing later clarifies that he means digital computers when he uses the word "machine" (Turing, 1950, p. 436). The word "AI" is absent from the article, as the word had not been coined yet.

anyone except ourselves. Aside from thought experiments and *a priori* postulations, there are currently no better alternatives available than behavioral analysis.

In his article Turing proactively attempted to respond to criticisms he had faced with his proposal (unfortunately, the criticisms are largely unattributed by Turing). Some of the criticisms Turing addressed may seem trivial or outright absurd to the contemporary reader. I will therefore first summarize the problematic ones very briefly, and give more space to the objections that seem to deserve some further thought. The criticisms include:

The Theological Objection: It is argued that thinking is an aspect of our immortal souls, and therefore machines cannot think. Turing himself finds this argument unconvincing, but also argues that if God can grant souls to humans, there is no reason why he couldn't grant them to machines. (Turing, 1950, p. 443.)

The Heads in the Sand Objection: It is argued that the consequences of thinking machines would be so dreadful that we can only hope they cannot exist (Turing, 1950, p. 444). Not much needs to be said of this argument aside from the fact that it is a clear case of fallacious reasoning.[10]

The Argument from Extra-Sensory Perception: Easily the most bizarre of the objections, Turing claims that "the statistical evidence, at least for telepathy, is overwhelming", and that it might be used by the interrogator to deduce which of the participants is the machine (Turing, 1950, p. 453). Suffice to say, science has marched on since Turing's days.[11]

Arguments from Various Disabilities: The argument goes that despite their skills, machines will always have limitations that they can never overcome. A selection of such limitations is quoted by Turing:

> Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make some one fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really new. (Turing, 1950, p. 447.) (Spelling as in the original.)

[10] It is true that the consequences of thinking machines *would* probably be quite dramatic.
[11] Also, if telepathy *was* a real thing, it could be used to probe the machine's mind to discover whether the machine is thinking at all, making the Turing Test itself redundant.

It is not clear how some of these limitations limit the possibility of thinking machines. Under a generous interpretation, it could be understood as claiming that machines lack several human characteristics that are required for conscious experience. For example, "enjoying strawberries and cream" could be seen as an example of lacking the qualia to appreciate the taste of strawberries. Even then, many of the examples are trivial or already achieved by AIs. The last three examples of the quote are more interesting, and are in fact discussed in other arguments against Turing.

The Argument from Informality of Behavior: The argument says that since it is impossible to produce a set of rules that dictates every decision a human ever makes, we cannot be machines (Turing, 1950, p. 452). Presumably, the assumption is that all machines operate under such rulesets, and therefore machines cannot be equivalent to humans and "think".

The argument becomes problematic whether we look at it from a deterministic or indeterministic perspective. If the world is deterministic, then there *are* rules (or, more precisely, causes) that determine what humans do in every circumstance. If the world is indeterministic, there are no such rules for machines either. (Oppy & Dowe, 2016, chap. 2.8.) In either case, the distinction between humans and machines appears somewhat blurry. The argument appears to tie itself to the existence of free will; people have free will to act as they choose, but machines are always just following their programming. The deterministic nature of simple programs is observed easily enough, but if the program becomes complex and self-augmenting to a sufficient degree, it is no longer obvious what its difference to human brain activity is – after all, brains operate under a certain set of rules too, and we do not (at least in our everyday lives) find it problematic for free will.

The Mathematical Objection: The objection claims that Gödel's incompleteness theorems limit the thinking of machines, because they necessarily limit the logical-mathematical reasoning that machines base their cognitive functions on (Turing, 1950, p. 444). The theorems state, as quoted by Finnish philosopher Panu Raatikainen:

> The first incompleteness theorem states that in any consistent formal system *F* within which a certain amount of arithmetic can be carried out, there are statements of the language of *F* which can neither be proved nor disproved in *F*. According to the second incompleteness theorem, such a formal system cannot prove that the system itself is consistent (assuming it is indeed consistent). (Raatikainen, 2015.)

The idea is that machines, operating under formal systems, would eventually run into unprovable Gödel statements they cannot solve, revealing their inherent cognitive limitations. The underlying assumption is also that humans could intuitively see whether the statements are true or false; if so, our minds would not be limited by the theorems, and our cognitive capabilities would therefore be superior to machines. (Raatikainen, 2015, chap. 6.3.)

This objection has been advocated by later proponents, such as J. R. Lucas and Roger Penrose, decades after Turing. Still, there is a wide consensus that the argument fails (Raatikainen, 2015, chap. 6.3). A central problem in the argument is the claim that humans have no similar limitations related to Gödel sentences – a claim that is asserted without actual proof (Turing, 1950, p. 445; Russel & Norvig, 2016, p. 1023). Russel and Norvig argue that humans *do* have such limitations, but that they are ultimately irrelevant to cognition: the sentence "J. R. Lucas cannot consistently assert that this sentence is true" is unprovable to J. R. Lucas, and humans are incapable of solving many mathematical problems on their own, while computers can solve them effortlessly. Yet J. R. Lucas or small children are not any worse off because of their mathematical or logical limitations. (Russel & Norvig, 2016, p. 1023.) The Mathematical Objection further assumes that humans can always see whether any given formal theory is consistent or not, a claim that Raatikainen has ruled highly implausible (Raatikainen, 2015, chap. 6.3).

Lady Lovelace's Objection: Turing quotes Lady Ada Lovelace, who wrote about Charles Babbage's Analytical Engine (an early, mechanical computer from the 1800's) in her memoir. According to Turing, Lovelace wrote that "The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform." (Turing, 1950, p. 450.) While Lovelace wrote about an early predecessor of the modern computer, the Analytical Engine had the same fundamental, if slower, operating logic as modern computers. Her point may thus still be argued: computers lack creativity and never come up with anything truly *new*. This may be seen as an example of the Argument from Disabilities.

It is true that many conventional Narrow AIs do lack any sort of creativity. But the large-scale application of simulated neural networks in the past ten years has produced a number of remarkable results in AI research that contradict Lovelace.[12] For example, DeepMind's AlphaGo Zero program used reinforcement-learning algorithms to become an expert Go

[12] Artificial neural networks were invented as early as the 1940's, but they have become truly useful only in the past few years, with sufficient computing power and available data to run and train them.

player in less than three days by playing against itself, starting with zero domain knowledge beyond the basic rules of the game. AlphaGo Zero not only demonstrated adaptiveness in the game, but it also came up with new strategies that humans had not thought of in 2500 years. (David Silver et al., 2017.) Thus it seems apparent that modern AIs *can* do ”something really new”, such as originating novel strategies for a given task.

One could argue that machine learning demonstrated by contemporary AIs is not equivalent to human creativity, and would therefore not count. This may be the case, although we would first need to agree on the precise definition of “human creativity”. However, in relation to the original objection this would be moving the goalposts.

A perhaps stronger objection could say that AlphaGo Zero does not really come up with anything *truly* new. This objection can be sharpened by a conceptual division made by Riccardo Manzotti and Vincenzo Tagliasco, who provide an interesting perspective with their concept of *teleological plasticity.* According to Manzotti and Tagliasco, AIs can be divided into roughly three categories by their ways of accomplishing their goals: fixed control architectures, learning architectures, and goal generating architectures. An AI has *fixed control* architecture if it is unable to modify its way of fulfilling its goal; it is slavishly following its code to the letter. A *learning architecture* AI has fixed goals, but is able to modify its behavior on *how* to fulfill those goals. A *goal generating architecture* AI not only modifies its behavior, but can also decide its goals for itself. (Manzotti & Tagliasco, 2008, p. 114.)

While AlphaGo Zero is an adaptive player of Go, it is still bound to playing Go only. It is a learning architecture AI, but not a goal-generating one; it cannot decide to do something else instead. Lovelace’s objection could be modified to object that AIs are not goal-generating agents, and therefore lack a central aspect of human cognition.

This modified objection could be directed at two targets. Lovelace’s original objection appears to be directed against the capabilities of Analytical Engine, and by extension, computers. This direction may turn out to be fruitless in the long run, since there is no reason in principle why AIs could not generate goals for themselves. Lovelace’s objection stalls in face of technological progress.

However, the modified objection could be directed against the Turing Test. Since a sufficiently advanced learning architecture AI could (presumably) pass the Turing Test, it could be argued that the test is not a good measure of consciousness because entities of inferior

cognitive skills could pass the test. The Turing Test would be too *easy* to pass, at least if we want it to measure consciousness.

The modified Lovelace's objection against the test carries an assumption that there is an inherent difference between learning architectures and goal-generating architectures that is significant in relation to consciousness. Intuitively, there does indeed appear to be a major difference between an agent that cannot decide what it does and an agent that can. However, our limited understanding of cognition means that we do not know how important the difference is in relation to consciousness. This is especially because the distinction between an advanced learning architecture and a goal-generating one may be slightly blurrier than initially seems. Suppose we have a highly advanced learning architecture AGI in our hands, and that the AGI's goal is to impersonate a human being as successfully as possible in a Turing Test. Even if the AGI is incapable of creating independent goals for itself, it must be able to create and pursue *sub-goals* to reach its set final goal – successful impersonation. This means that the AGI would have to be able to perform any task a human being could, such as (for example) learning to read, write poetry, make plans for the future, etc. If the AGI could do all these things, how significant is its difference to an independent goal-generating AI?

Argument from Continuity of the Nervous System: Another objection Turing addressed is that since a digital computer is a discrete-state machine (it can only be in one state at a time – one or zero) and the human nervous system is a continuous system with its constant neuron firings, a computer cannot mimic the human nervous system (Turing, 1950, p. 451).

An important note in this argument is what it exactly claims. The argument as stated does not only say that a computer program cannot be conscious, but that it cannot even *mimic* the brain activity to pass the Turing Test. The objection therefore suggests that even a Weak AGI is impossible – at least if we want to create the AGI's intelligence by simulating human neural activity.

Turing acknowledges the distinction between discrete and continuous systems, but claims that a simulation *is* possible with a discrete system. He observes that it is possible to simulate the results of a differential analyzer (a non-discrete machine) with a digital computer with a very small degree of error. There would then be no reason in principle why discrete-state systems could not simulate neural activity as well. (Turing, 1950, pp. 451–452.)

Oppy and Dowe observe that Turing's response may not be good enough. If the objection is asserted more strongly – that only a continuous system can generate consciousness – then Turing's response would only show how inadequate the Turing Test is in the first place. They argue that a stronger line of defense would be to question that consciousness can only be born in a continuous system, as it is asserted without further proof, or even to question whether a human nervous system really is a non-discrete system at all. (Oppy & Dowe, 2016, chap. 2.7.)

The Argument from Consciousness: This time Turing quotes Professor Jefferson's Lister Oration of 1949 as an example of the last objection he discusses:

> Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals a brain – that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants. (Turing, 1950, pp. 445–446.)

As Turing notes, the objection essentially denies the validity of the Turing Test: it claims that subjective qualia of some kind must be present in order to attribute true thinking to machines. In modern terms, one could say that both Weak and Strong AIs can pass the test, making it useless.

This is perhaps one of the strongest objections against the Turing Test, but paradoxically, it also shields Turing to some extent. Turing notes that by the same standards, we cannot know if a human being really "thinks". The only way of being sure of that is to *be* the human being (or machine) in question. Embracing the objection would drive its proponent into a solipsist corner. (Turing, 1950, p. 446.) Turing remarks that despite this obvious pitfall, few people are actually willing to take the extreme solipsist position. Instead, it is usually assumed that everyone who appears to think really thinks – this is what Turing calls the "polite convention". (Turing, 1950, p. 446.)

An opponent might well insist that there is nonetheless a difference. Each individual human being knows they think. They further know that their brain structure and wider anatomy is similar to other humans, so it is a reasonable jump to assume that other humans think as well. But since we do not know how consciousness is born, we cannot say with certainty that computers have a similar structure as the human brain that enables consciousness. We cannot

conclude from computers' behavior alone that they are conscious. Weak AIs can pass the test too.

Turing argues that the distinction eventually fades away by itself. He provides an example of a conversation, made famous by later authors quoting it:

> Interrogator: In the first line of your sonnet which reads 'Shall I compare thee to a summer day', would not 'a spring day' do as well or better?
>
> Witness: It wouldn't scan.
>
> Interrogator: How about 'a winter's day'? That would scan alright.
>
> Witness: Yes, but nobody wants to be compared to a winter's day.
>
> Interrogator: Would you say Mr. Pickwick reminded you of Christmas?
>
> Witness: In a way.
>
> Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.
>
> Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas. (Turing, 1950, p. 446.)

Turing argues that if machines were capable of complex conversations as demonstrated above, people would eventually cease making the distinction between "human" thinking and "machine" thinking, and simply adopt the polite convention that everyone thinks (Turing, 1950, p. 446–447).

Turing believed that it would take "about fifty years" until computers with storage capacity of about 10^9 existed that could play the imitation game so well that an average interrogator could reveal the machine only 70% of the time (Turing, 1950, p. 442). Turing provided no unit of measure, but if we take the common byte as a unit we can tell that even a modern laptop with 1000GB of storage space exceeds Turing's prediction by three orders of magnitude. But still, no computer program has managed to pass the Turing Test satisfactorily. The annual Loebner Prize for the most successful AI program has to this day never had a participant winning the grand prize. Claims of a passed Turing Test have been made throughout the years, such as *PC Therapist* in 1991, *Cleverbot* in 2011, and *Eugene Goostman* in 2014, but all with small trial sizes and unprojectible results (Oppy & Dowe, 2016, chap. 1). The

Turing Test does not therefore seem like an easy test to beat, despite the advance of AI research. Is the problem with the test or the research?

A short answer is that AI has moved on to other topics. Contemporary AI research mostly focuses on Narrow AIs that provide expertise on a very narrow domain, and the research community sees passing the Turing Test as an irrelevant goal to achieve in those specific tasks. Research of AGI is not a central focus on the field. However, this does not mean that the test itself is uncontroversial.

Many critiques of the Turing Test depend on how the test exactly is to be interpreted. Oppy and Dowe divide the interpretations into four: (1) that the Turing Test provides logically necessary and sufficient conditions for attribution of intelligence; (2) that the test provides logically sufficient, but not logically necessary, conditions for attribution of intelligence; (3) that the test produces sufficient but defeasible conditions for attributing intelligence, and (4) the test provides *probabilistic* support for attribution of intelligence. (Oppy & Dowe, 2016, chap. 4.) We will take a look into all four interpretations.

Firstly, the interpretation (1), that the Turing Test provides logically necessary and sufficient criteria for attributing intelligence. Oppy and Dowe observe that few people interpret the Turing Test this way, and indeed, it would be quite a rigid view to hold. Yet some criticism of the test only applies if it was interpreted exactly like that: the test has been criticized as chauvinistic, since any number of conscious beings could fail the test for any number of reasons, such as lacking a common means of communication. (Oppy & Dowe, 2016, chap. 4.1.) This is certainly true, and it should make us weary of thinking the Turing Test as an absolute measuring stick of intelligence. It also somewhat limits the test's usefulness, but not to an entirely unusable degree.

Next interpretation (2), that the Turing Test provides logically sufficient, but not logically necessary, conditions for attributing intelligence. This view has had more support than interpretation (1) among the philosopher community, with at least some claiming that it is impossible for an unintelligent agent to pass the test (Oppy & Dowe, 2016, chap. 4.2). The latter claim should perhaps be taken with a grain of salt. A reply to interpretation (2) is that the Turing Test could be passed with "brute force" methods, and that this would not yield true "intelligence" (Oppy & Dowe, 2016, chap. 4.2). A prime example of this kind of "brute force agent" is Blockhead, a thought experiment suggested by philosopher Ned Block in 1981. Block argues that for any finite-length conversation there exist only a finite amount of

sensible responses. He describes Blockhead (Block does not use the term himself in the original article) as a machine that has stored every conceivable response to every conversation imaginable in a lifetime, and looks up an appropriate answer to any sentences it hears from its look-up tree of pre-coded responses. Blockhead would pass the Turing Test, but in the words of Block, it would have "the intelligence of a toaster." Blockhead would prove that a non-intelligent robot could pass the Turing Test, denying its validity as a *logically* sufficient condition for intelligence. (Block, 1981, pp. 19–21.) The emphasis is important, since it will be brought up later.

Oppy and Dowe observe that the only ways to deny the logical validity of the Blockhead argument is to either deny that Blockhead is a logical possibility, or to claim that it in fact possesses a mind (Oppy & Dowe, 2016, chap. 4.2). Both of these approaches are controversial, but ultimately they can be side-stepped by interpretation (3) of the Turing Test, that the test produces sufficient but defeasible conditions for attributing intelligence.

The idea behind interpretation (3) is that in order to attribute consciousness to an entity, there must be true claims of observable behavior of the entity that imply a present consciousness. If observable behavior (coupled possibly with other claims) does not play any role in attributing consciousness to an entity, then there are no grounds to attribute consciousness in the first place; the Turing Test can be seen as providing *empirical criteria* for consciousness, rather than *logically* sufficient or necessary conditions for it. (Oppy & Dowe, 2016, chap. 4.3.)

Since we are no longer discussing a strictly logical possibility, we can question the Blockhead response. Even if Blockhead was a logical possibility, it would not necessarily follow that it is a real, *physical* possibility. Computer science Professor Drew McDermott has calculated that the Blockhead, or any other exhaustive look-up tree program, would require absolutely *enormous* amounts of space, many, many, *many* orders of magnitude larger than the entire observable universe (McDermott, 2014, p. 157).[13] The logical possibility of Blockhead would therefore not impugn the Turing Test as a practical experiment or criterion for attributing consciousness, since Blockhead is a physical impossibility.

[13] More precisely, the size of the observable universe is approximately 2×10^{185} measured in cubic Planck lengths, and an exhaustive Blockhead program would contain around $10^{22\,278}$ individual conversation nodes (McDermott, 2014, p. 157).

Lastly, interpretation (4), that the Turing Test provides probabilistic support for attributing intelligence. This interpretation is the most moderate of the four we have considered, and quite possibly the most useful one. The interpretation holds that success in the Turing Test only provides us with an increased level of confidence that the machine that passed the test is conscious. Specific factors would yield more confidence, such as the length, content, and number of interrogations, the expertise of the interrogator etc. (Oppy & Dowe, 2016, chap. 4.4.) ”Increase of confidence” means that we need not hold passing the Turing Test as a definite finishing line when attributing consciousness. It only means that we have placed this attribution on a more solid foundation, subject to further confirmatory or contrary evidence. From a philosophical point of view this interpretation may not be as satisfying as the previous ones, but its inductive line of reasoning is in line with the scientific method in lieu of further breakthroughs in cognitive science.

The Turing Test has sometimes been suggested as a concrete goal in artificial intelligence research, with varying opinions on whether this is a good or realistic goal. Some people hold that the test is too difficult to be a realistic goal or too narrow to meaningfully test broadly intelligent behavior (see for example French, 1990). Oppy and Dowe (2016, chap. 5.2) advocate for the view that any entity that passes a sufficiently rigorous Turing Test must be able to operate intelligently on a wide variety of circumstances – a hallmark of AGI. If true, the Turing Test would be a good method for identifying certain kinds of AGIs.[14] Not *every* kind of AGI could be caught by the test, however. French observes that the Turing Test is a good measure of only specific, culturally oriented human intelligence, and thus suitable only for that specific inquiry. Other intelligent entities could easily fail the test because they might lack common linguistic conventions, for example. (French, 1990, p. 8.)

- -

In conclusion, the Turing Test is a problematic approach to determining consciousness. As we have seen, the test measures only a narrow domain of culturally oriented intelligence, and even then, it can only provide us with a probability that the interrogated agent in question is conscious. The test may be somewhat better if we interpret “intelligence” more loosely as

[14] A sufficiently advanced AGI might have a number of incentives to mislead its interrogators into believing it is less intelligent than it really is, but these scenarios are not strictly relevant in our context. An in-depth analysis on the dangers of such superintelligent agents has been written by Nick Bostrom (2014).

any intelligently interpretable behavior and not consciousness, but it nonetheless makes for a poor general-purpose tool of establishing the presence of conscious thought.

The Turing Test is not the only behavioral test that has been devised for its purpose, although it is arguably the most famous one. Alternative behavioral tests and modifications have been proposed over the years, such as the so-called "Total Turing Test", "Truly Total Turing Test", and the "Lovelace Test", which take into account sensorimotor abilities, evolutionary and historical record, and creativity, respectively (Oppy & Dowe, 2016, chap. 5.3). While these tests set new or different criteria, they face the same central problem of the behavioral approach that the Turing Test does: the claim that behavior is enough. If the stalemate in cognitive science continues and we develop AIs that *appear* conscious, this may be enough for practical purposes – in such a case Turing's polite convention sounds like a plausible outcome, especially given our tendency to humanize various entities regardless of their exact nature. Still, the end result would probably leave many philosophers dissatisfied.

This dissatisfaction has fueled later critics who have sought to point out the inadequacies of the Turing Test. The next section will discuss the most notable of these critics, John Searle and his Chinese room argument.

3.2 The Chinese Room Argument

One of the most important academics in the Strong AI debate is John Searle, who coined the term in his 1980 article *Minds, brains, and programs*. Searle is a strong opponent of a conscious AI, and his main argument to support his position was provided in 1980 in the form of the *Chinese room argument*. Searle designed his argument (abbreviated CRA for ease of use) in part against the mainstream cognitive science of the day, and to point out the flaws in the Turing Test in attributing consciousness based on behavior alone. The CRA has been a major point of contention for decades, and its larger context spills over from the direct AI debate into other fields of philosophy, such as the symbol grounding problem. The argument has been highly influential in the years following its publication, and is still hotly debated to this day.

The argument goes as follows: suppose a man has been locked inside a room containing a pile of papers full of Chinese symbols – incomprehensible to the man, who does not speak Chinese – and instructions in his native language on how to correspond those symbols with each other. A second batch of Chinese symbols is then delivered to the room via a hatch, and the man follows the instructions which tell him how to correspond the symbols to the ones he received. For example, the instructions might tell him that when he identifies the symbol string 你从哪里来, he is to respond with symbols 我来自中国. The man follows the instructions blindly and sends out his responses. Unbeknownst to the man, he is actually conversing with a native Chinese speaker; the Chinese person is inputting questions outside the room, and the responses he receives are in perfectly legible Chinese given by the man inside. The Chinese person appears to be having a genuine and meaningful conversation, but the man in the room has no knowledge of this, and indeed, learns nothing about the Chinese language and what he is "saying". He is just mechanically following instructions on how to respond to any given input. Thus, Searle concludes, a computer program – in this case, the man carrying out the instructions or the "program" – cannot *think* simply by virtue of manipulating symbols. The program has *syntax* but no *semantics* attached to it – it does not *understand* what it is doing. (Searle, 1980, pp. 417–418.) Because a computer is ultimately just juggling binary ones and zeroes, it can achieve no semantic understanding of what is it doing, and hence cannot be conscious. *Ergo,* Strong AI is impossible.[15]

The CRA spurred a number of counter-arguments after its publication, and only the most notable ones can be addressed here in varying capacity. The so-called "Systems Reply", "Robot Reply", "Other Minds Reply", "Brain Simulator Reply" were named and addressed by Searle in his original 1980 article, in addition to which we will look at the so-called "Intuition Reply". Not all counter-arguments against the CRA can be discussed here, as this would require a full-blown article of its own.

Systems Reply: The *Systems Reply* is the one Searle noted he received most often when receiving feedback for his article, and which he attributes to UC Berkeley. The reply asserts that Searle is wrong in saying that the man in the room has gained no understanding is the same thing as *no understanding at all* having been gained. The man in the room is only a

[15] Searle rather vaguely uses "semantics", "intentionality" and "mind" in an apparently synonymous fashion. While these terms can be interpreted differently, I will take them to refer to subjective mental experiences or qualia.

part of the whole mechanism, and even if he does not understand Chinese, the system as a whole does. (Searle, 1980, p. 419.)

The Systems Reply has had a wide range of advocates, such as Georges Rey, Margaret Boden, and Daniel Dennett, among many others. Many of their objections echo each other: Rey has argued that the human inside the Chinese room acts as the CPU of the "computer" and should not be the focus of the inquiry (Rey, 1986, p. 173). Boden has raised a similar point regarding human cognition, and notes that while the brain is central in forming intentionality, according to computational psychology intentional states are ultimately "properties of people, not brains" (Boden, 1987, p. 12). Dennett claims that semantic understanding arises from increasingly complex syntactic architectures, and that a sufficiently complex system would be able to understand Chinese just as the human brain does (Dennett, 1992, p. 438–440).

Searle's response to the Systems Reply is to modify the thought experiment so that the man memorizes all the Chinese symbols and rules on how to manipulate them, and gets rid of the room itself. In other words, he memorizes and "becomes" the entire system. Still, Searle says, no understanding of Chinese is gained without something more being added. He notes that if the man in the room does not understand Chinese, it is absurd to claim that the man and the rulebook in conjunction would somehow come to understand it. (Searle, 1980, p. 419.)

Searle further goes on to investigate the reply through another (perhaps more implausible) perspective, according to which the man contains two sub-systems, one which understands English, and one which understands Chinese. Searle notes that these sub-systems would not be even remotely equal: the sub-system for English understands that the word "hamburger" refers to hamburgers etc., but the sub-system for Chinese would only know that the symbol "squiggle-squiggle" is followed by the symbol "squoggle-squoggle" and have no knowledge of what these symbols actually refer to. (Searle, 1980, p. 419.) Searle says that the only independent grounds he can imagine for supposing the existence of such sub-systems is that they could pass the Turing Test and hence demonstrate understanding of Chinese to an outside observer, but that this is an inadequate response because the CRA is meant to show precisely that this is not enough for semantics understanding. Such an argument would simply beg the question. (Searle, 1980, p. 419.)

Robot Reply: The so-called *Robot Reply* (attributed to Yale University by Searle) concedes that pure symbol manipulation alone is not sufficient for semantic understanding, but that a

computer could glean such understanding under some circumstances. Suppose the computer is put inside a robot. It is provided with visual and auditory sensors that would enable it to perceive the outside world, and actuators that allow it to interact with it. Thus the robot – and more importantly, the program operating it – would perceive and act on the world, and so gain semantic understanding by observing, for example, what a hamburger looks like. (Searle, 1980, p. 420.) David Cole puts this by saying that sensory information would provide sufficient causal connections to ground the internal symbol system to the outside world (Cole, 2014, chap. 4.2).

The necessity of a sensorimotor interface or embodiment for consciousness is a topic studied under embodied cognition, and will be discussed in more detail in its own sub-chapter. However, for our immediate purposes Searle rejects the Robot Reply as a plausible response to the CRA. Firstly, Searle notes that the reply concedes a crucial point that he is trying to make: that semantics cannot be gained from syntactic symbol manipulation alone. Searle also sees the counter-argument in general as flawed. He points out that, ultimately, a robotic body would not grant anything new to an AI, since all the extra information it receives from its sensors would have to be converted into ones and zeroes for the computer to process it; it would just be computing *more* ones and zeroes. Searle compares this to the man in the Chinese room *not* getting (say,) windows to look outside, but rather as more Chinese symbols and instructions that, without him knowing, correspond to sensory inputs and instructions on how to move the "robot's" limbs. (Searle, 1980, p. 420.) The man in the room, and an AI, would still be "blind" to what they are actually doing.

Other Minds Reply: This reply from Yale makes a more down-to-earth response. In the spirit of the Turing Test, the reply notes that we can assert consciousness of other beings only by their behavior. If we are prepared to do this with humans, why not with seemingly conscious robots? (Searle, 1980, p. 421.)

Searle gives this objection a curt response:

> The problem in this discussion 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 to them. The thrust of the argument is that it couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive state. (Searle, 1980, p. 422.)

Searle's reply has not satisfied everyone. Some critics have objected that evidence for human understanding is ultimately the same as evidence for a hypothetical alien visitors understanding, and hence for a robot understanding. Presuppositions about our own species should not be considered relevant, since presuppositions can be untrue. (Cole, 2014, chap. 4.4.) Cole also raises the question of what extra is Searle exactly attributing to humans, if behavior alone is not sufficient (Cole, 2014, chap. 4.4). Searle's reply to this is that we attribute both behavior and a similar physiological system to justify consciousness of other beings – for example, we conclude that a dog is conscious because it flinches when we shout at its ear, and because we attribute the same properties to dog's ears as we do to human ears (it hurts when you yell too loud). Both behavior and physiology are used in conjunction to attribute consciousness to others. (Searle, 2002, pp. 73–74.)

Searle's reply is not unique to him. It was proposed by John Stuart Mill over 150 years ago. Mill wrote:

> I conclude that other human beings have feelings like me because, first, they have bodies like me, which I know, in my own case, to be the antecedent condition of feelings; and because, secondly, they exhibit the acts, and outward signs, which in my own case I know by experience to be caused by feelings. (Mill, 1865, p. 208.)

Whether or not Searle knew about Mill's ideas, the general principle appears to be the same. This approach to the problem of other minds is considered "traditional", and is often referred to as the "argument from analogy", since it cites similarities between two entities to justify the postulated presence of further, assumed, similarities (Anita Avramides, 2019, chap. 1.1).

Is Searle's reply satisfactory in this case? The problem of other minds in general has been a source of lively philosophical debate, so it may be worth a brief detour to summarize the discussion to see if there are any insights that may prove useful in the debate over Strong AI.

The argument from analogy mentioned above was later contested by 20th century philosophers. A common objection claimed that the argument relies too heavily on generalization from one single case – one's own mental life – to justify the existence of other minds (Simon Blackburn, 1996, p. 273). Another objection is that the argument's conclusion cannot be logically checked – that is, it cannot be empirically verified. (Alec Hyslop, 1995, p. 41).

However, Searle is not alone among contemporary philosophers in advocating the argument from analogy. Hyslop has defended the argument from the two objections mentioned above.

According to Hyslop, the objection that the argument's conclusion cannot be logically checked is not as problematic as its proponents suggest. Hyslop argues that the objection holds a crucial premise (P):

> (P) No analogical (inductive) argument which proceeds from observed facts to a logical uncheckable conclusion is a good analogical argument (Hyslop, 1995, p. 55).

Hyslop argues that (P) is not self-evident and should not be followed without further arguments in support of it. Hyslop claims that there are other cases where uncheckable arguments are nevertheless held valid, and cites helium as an example. We can observe that helium on Earth possesses a certain spectrum, and we can deduce from this that the Sun is partially made of helium, because it emits a similar spectrum. The conclusion is widely accepted, even though we have ever only studied helium on Earth and cannot (currently) study helium from the Sun for first-hand evidence. Thus Hyslop concludes that (P) is not self-evident and should not be followed on *a priori* grounds, and that the burden of proof is not with the proponent of the argument from analogy. (Hyslop, 1995, pp. 40, 56–57.) Hyslop says that while the helium example is a case of empirical impossibility and not logical impossibility, it does not impugn the immediate argument he is trying to make (Hyslop, 1995, p. 57).[16]

The other classical objection to argument from analogy was that the argument makes an unacceptable generalization from a single case. In response, Hyslop claims that there are acceptable analogies that are drawn from single cases, and that the existence of other minds is one of them (Hyslop, 1995, p. 52).

Hyslop compares the argument from analogy to a scenario where he drops an egg on the floor, breaking it. Is he entitled, Hyslop asks, to infer from this that the next egg will break when it is dropped? Hyslop notes that we would certainly be happier if we could repeat the experiment multiple time with the same results. But according to him, this is not because it would yield a statistically more significant result, because the total number of untested eggs versus tested ones is so large that any reasonable amount of new tests would not increase the total percentage of eggs tested. (Hyslop, 1995, pp. 52–53.) Instead, according to Hyslop, further tests can rule out the possibility that the first egg broke because of some other reason than because of it was dropped. If we had known for sure that the first egg broke because of being dropped, the following tests would have been unnecessary. (Hyslop, 1995, p. 53.)

[16] Hyslop also addresses other objections from logical uncheckability, which unfortunately cannot all be discussed here. Readers can turn to Hyslop (1995, pp. 56–70) for the relevant parts.

Hyslop uses this argument to support the idea that inferring the existence of other minds can be reasonably done from a single case, blunting the objection. According to Hyslop, we infer that our own mental states are caused by our physical (brain) states, and further infer from this that the same applies to other people. In other words, we make a connection to causality: since *my* physical brain states cause my mental states, and because similar causes have similar effects, I can infer the existence of other minds with similar physical states. (Hyslop, 1995, p. 53.) Hyslop claims that strengthening the sample size would not yield an increase of confidence in physical states causing mental states, since it is a matter of inference, not observation. What the argument from analogy needs is a connection between physical and mental states, and since this can be established in one's own case, the argument can provide "a multitude of relevant correlations". (Hyslop, 1995, pp. 53–54.)

Hyslop's egg-sample could be objected to for not appreciating the fact that inductive reasoning he relies on generally depends on greater affirmation through more and more successive samples. Indeed, despite his arguments, Hyslop is one of the relatively few contemporary defenders of the argument from analogy, and we do not need to accept his reasoning here. Another alternative exists.

An alternative solution to the problem of other minds is the so-called "best explanation", or argument from scientific inference. This explanation has been suggested as an advance to the argument from analogy, and it enjoys broad support today. (Avramides, 2019, chap. 1.2.)

The argument from scientific inference maintains that we are justified in believing in other minds simply because it provides the best explanation for the behavior of other people. It is implausible to appeal to other explanations, such as God controlling other people via puppetry, to explain the same outward behavior that I exhibit because of thoughts and feelings. (Pargetter, 1984, p. 158.) Robert Pargetter notes that while this does not give us completely certain knowledge of other minds, it does provide strong inductive arguments for their existence (Pargetter, 1984, p. 159).

So how is this different from the argument from analogy? The difference is subtle. The argument from analogy bases its force on the strength of the analogy and the generalization based on it. Scientific inference, on the other hand, is founded on the practical assumption that an explanatory model is more likely to be true simply if it explains the observed phenomena satisfactorily. If they were available, alternative explanations could trump our epistemic commitment to other minds if they could explain behavior in a better way.

While popular, the argument from scientific inference has faced its own objections. Hyslop has claimed that the argument incorrectly assumes that similar effects (behavior) are always associated with similar causes (mental states), and thus makes the inference of other minds too quickly. He also says that the argument still relies on a crucial single case to work – one's own – and is therefore not any stronger than the argument from analogy. Hyslop further asserts that the conclusion from scientific inference is also logically uncheckable, and because of this it is subject to the same criticisms as the argument from analogy.[17] (Hyslop, 1995, pp. 30–31.)

It is possible to defend the argument from scientific inference from Hyslop's first objection by reforming it slightly. Hyslop writes that:

> At its most straightforward [scientific inference] starts with the problem of explaining human behaviour and concludes that the best explanation for that behaviour is that human beings behave as they do because they have minds. (Hyslop, 1995, p. 30.)

If this formulation is true, then it would indeed be the case that the argument would base itself on the debatable idea that like effects have like causes. Hyslop may be justified in this formulation, since Pargetter refers to behavior alone when he talks about inferring other minds. But our ultimate goal is to justify the existence of other minds, not to explain behavior, and we can achieve this by modifying the argument to approach the question from a different angle in a way that responds to Hyslop.

We could modify the argument to assert that we appeal not only to behavior, but also to observable, physical brain activity to justify the existence of other minds. We could argue that since anyone could observe (in principle, with the right equipment) that their physical brain activity is associated with subjective mental experiences, and see that other people display similar brain activity and behavior, it is reasonable to infer that their brain activity and behavior are accompanied by a subjective consciousness. This formulation would avoid appealing to behavior alone, since it appeals to a common, underlying cause – observable brain activity. We do not need to go into details in which is the more fundamental causal basis for behavior, brain activity or mental states. We only need to assert that because relevant brain activity is associated with subjective mental experiences, the scientific inference of other minds is a justified postulation – especially if it is further supported by behavior.

[17] Given that Hyslop argued that the latter two objections are invalid, it is not entirely clear how he believes they supposedly harm the argument.

Pargetter does not take this line of defense, since he focuses on behavior, but he has defended scientific inference from Hyslop's latter two objections: uncheckability and reliance on a single case. Pargetter says that while the argument from analogy is forced to defend itself from logical uncheckability and reliance on one's own case to work, the argument from scientific inference can bypass these problems altogether. According to Pargetter,

> The strength of a scientific inference depends solely on the explanatory power of its conclusion. Now the explanatory power of the hypothesis that other people have minds qualitatively similar to my own is not in any way impaired by the fact that there is only one mind (my own) of which I have direct knowledge. Nor is it impaired by the fact that I cannot check on the conclusion after using the inference to argue that someone else is minded. What I *can* check on is that this hypothesis does explain the behaviour of other people in a satisfactory way, and this is the only check required for the justifiable use of a scientific inference. (Pargetter, 1984, p. 160.)

In other words, the argument utilizes the scientific method of following the hypothesis that provides the highest explanatory value. If a better, more plausible, explanation to other people's behavior is found, it should be pursued instead. The argument is not doubtless, and is not meant to be one; it is inductive and fallibilist by nature because we have no direct access to other minds, and uses all available evidence, including knowledge of one's own consciousness (Pargetter, 1984, p. 162).

The discussion of other minds in relation to other human beings is one of the topics in philosophy that has moved quite some distance away from everyday thinking and "common sense". But the debate is more pressing than ever with AI, so what are the insights of these arguments in relation to Strong AI?

As we noted earlier, Searle seemed to endorse some version of the argument from analogy. In relation to machines the analogy does indeed seem to support Searle, since because AIs share no similar physiology with humans, its strength of analogically inferring a similar mental life for AIs is seriously weakened. At the same time it must be noted that the argument from analogy could not be used to infer that an alien being possesses a mind, since there is no similar physiology to base the justification on. It would seem that the argument is unable to assert the consciousness of *any* non-human or non-animal entity, unless we take appropriate behavior as a sufficient criterion. This would, however, reduce the strength of the analogy. The argument is designed to apply to humans, but this makes wider generalization difficult. In short, while Searle could use the argument to insist that AIs are not *obviously*

conscious despite behavior, it would not be enough to claim that they *cannot* be conscious. Searle must rely on other means to assert this.

How about the argument from scientific inference? Pargetter certainly believes that the argument can be extended to apply not only to humans, but also to other animals and hypothetical alien beings (Pargetter, 1984, p. 161). Presumably, this could apply to AIs and machines as well. However, Hyslop criticizes the scientific inference argument for making the inference too loosely. He objects that scientific inference makes hypothesizing the existence of other minds too easy, and interestingly, it "would seem too easily to provide sufficiently versatile computers with minds". (Hyslop, 1995, p. 30.)

Hyslop does not go into details with this notion, but it seems evident that he excludes at least some computers from possessing consciousness. This is by no means an unreasonable claim, since even Narrow AIs can give an impression of consciousness without actually being conscious. Hyslop's objection does raise a problem for scientific inference: while scientific inference can justify belief in non-human minds if it provides the best explanatory value, it is also possible that it attributes minds in cases where there are none to be found. It is not fully clear how deeply scientific inference really falls into this bog, however. Whether subjective consciousness provides the best explanation for behavior of advanced AIs is an open question, and alternative, plausible explanations can be based on the AI's algorithms without having to appeal to subjective consciousness. Unfortunately this is not very helpful for our purposes.

In summary, neither the argument from analogy nor scientific inference provide a satisfactory response to the Strong AI debate. Both arguments are designed to address the problem of other *human* minds, and both require first-person accounts as a starting point to be truly effective. This means that the arguments are not very helpful in relation to Strong AI: the argument from analogy becomes weaker when it is extended to radically different entities than us, and while scientific inference can be used to reason that non-human entities may have minds, this assumption lacks persuasive force if there are equally plausible hypotheses at play. On the other hand, advances in the scientific study of consciousness can place scientific inference of other minds on firmer footing in the future.

It would seem that Searle's defense against the Other Minds Reply is certainly not his weakest point, and that he does not carry the weight of the burden of proof in this regard. While Searle appears poorly equipped to attribute minds to possibly conscious entities in cases

where the argument from analogy is not at its strongest, it does not hamper his position towards Strong AI.

Brain Simulator Reply: Another response to the CRA is the Brain Simulator Reply from Berkeley and MIT. The reply suggests we could construct a computer that does not execute a formal computer program, but rather, is simulating the entire functioning of the human brain, down to every single neuron firing. This system would accurately simulate what is happening inside the head of a native Chinese speaker when he is speaking Chinese, and therefore the system would understand Chinese as well as any human speaker. If this was to be denied, wouldn't it also deny that any humans understand Chinese? (Searle, 1980, p. 420.)

The Brain Simulator Reply contains certain connectionist undertones. Connectionism is a movement in cognitive science that studies cognition as operation of neural networks, as opposed to the classical theory of mind where cognition is seen as computing symbolic representations (Cameron Buckner & James Garson, 2019). Connectionist AIs utilize artificial neural networks which can be seen as analogous to neurons, and could thus implement the scenario described above.

Once again, Searle modifies his thought experiment in response. Now the man inside the Chinese room is not operating Chinese symbols, but a complex series of water pipes and valves. When he receives the Chinese input, he looks up the English-language operating manual and opens and closes the valves according to its instructions, simulating "neuron firings" in the brain. After all the operations are complete, a Chinese-language answer pops out. However, the man in the room, the water pipes, and both of them in conjunction are as clueless as ever. Searle observes that in such a scenario not enough can be simulated, since what is missing are the causal powers of the brain (i.e., the "powers" that produce a mind), or its ability to produce intentional states. Searle says that formal properties of the brain – computation, or the pipes and valves – are not sufficient to produce causal properties. Instead, he holds that consciousness is necessarily a *biological* phenomenon – a view he has dubbed "biological naturalism". (Searle, 1980, p. 421, 424.)

A supporter of connectionism could object to Searle. In Searle's scenario there is a central operator who operates the connectionist system, but the brain has no central processor that directs the neural activity. We want to know if the entire connectionist system comes to understand anything, not the hypothetical homunculus at its center. Would this make a difference?

Searle says no. Connectionism does not in itself pass his criticism of Strong AI. Searle argues that if the connectionist computer only simulates the formal (connectionist) structures of the brain, it is still subject to the critique of the CRA, since the computer is still using ones and zeroes to conduct the simulation. If, however, the system is indistinguishable from the human brain in the physical level, then you have duplicated, and not simulated, the human brain. (Searle, 2001, chap. II.) Searle does not say where the line between sufficient and insufficient duplication of causal properties is drawn.[18]

There is a similar thought experiment to the water pipe scenario that might provide further insight: the so-called "brain replacement experiment". The thought experiment involves replacing an individual's neurons one at a time with tiny nano-machines that perform the exact same functional roles that the neurons do (Russel & Norvig, 2016, p. 1029). If the subject was conscious during the operation, he should be able to observe any changes in his conscious experience. As more and more neurons are replaced, the question is: what happens to his consciousness? Is the situation identical to Searle's pipes and valves?

Russel and Norvig provide a brief analysis on the consequences of brain replacement experiment. They quote Searle, who appears to insist that the subject's conscious experience would end during the procedure:

> You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when doctors test you vision, you hear them say "We are holding up a red object in front of you; please tell us what you see." You want to cry out "I can't see anything. I'm going totally blind." But you hear your voice saying in a way that is completely out of your control. "I see a red object in front of me." …your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same. (Searle, 2002, pp. 66–67; quoted in Russel & Norvig, 2016, pp. 1029–1030.)

Russel and Norvig note that the scenario Searle describes is unlikely. If the subject's outward behavior remained the same while his conscious experience slowly faded away, the removal of volition would have to be instantaneous: otherwise the subject could alert the staff of his gradually fading consciousness. Since the neurons are replaced one at a time and not all at once, the scenario depicted by Searle is not likely to happen. (Russel & Norvig, 2016, p. 1030.)

[18] While not exempt from the objections of biological naturalism, some non-algorithmic connectionist architectures could address the syntax-semantics criticism rooted in the CRA. These accounts will be discussed in the Embodied Cognition sub-chapter.

What if there is no difference to the subject's observable behavior during and after the experiment? In this case, Russel and Norvig point out that the subject's external behavior would be identical even if his consciousness had disappeared with his biological neurons. Since the artificial nano-machines are only performing the same functions as the real neurons, and the subject would appear outwardly to be the same, at least our outward behavior – every smile, wink, and cry of pain – would be determined by the purely functional properties of the neurons, since the functional properties were unchanged between the biological and the artificial ones. (Russel & Norvig, 2016, p. 1030.) Outward behavior, at least, would not therefore be determined by some inherently biological characteristic. Russel and Norvig outline three possible conclusions for the experiment:

1. The causal mechanisms of consciousness that generate [behavioral] outputs in normal brains are still operating in the electronic version, which is therefore conscious.
2. The conscious mental events in the normal brain have no causal connection to behavior, and are missing from the electronic brain, which is therefore not conscious.
3. The experiment is impossible, and therefore speculation about it is meaningless. (Russel & Norvig, 2016, p. 1030.)

The first conclusion would directly contradict biological naturalism. The second conclusion would leave Searle's position at least formally intact, but it would reduce consciousness to an entirely epiphenomenal role, with no control or influence over our behavior. Neither conclusion appears pleasing to Searle.

However, the third conclusion cannot be dismissed either, and unfortunately Russel and Norvig commit a big blunder with their interpretation of Searle. In the passage that they quote from him (Searle, 2002, pp. 66–67), Searle uses the brain replacement experiment to argue for a different argument entirely. He goes through three possible outcomes for the experiment, of which the quoted one is just one. Russel and Norvig use the passage to show that Searle endorses to the conclusion he describes above, but in reality he does not commit to this scenario!

This does not directly undermine the three conclusions Russel and Norvig present, but it opens the door for other objections. Russel and Norvig only consider the scenario where the subject's behavior remains unchanged throughout and after the experiment. There are two other hypothetical outcomes as well: one where the subject's outward behavior terminates and he ends up in a vegetative state, and one where he notices his fading consciousness but continues to function regardless, ultimately with no conscious mental states present. Both of

these two options could vindicate Searle, since they would imply that consciousness is indeed generated by some biological characteristic that was removed in the experiment, although the second possibility again raises the question of epiphenomenalism in a more diluted form.[19]

What version does Searle endorse, then? It is difficult to say without a first-hand source. Since Searle outlined the water pipe scenario as a refutation of the Brain Simulator Reply, it would seem plausible to say that he would claim the subject's consciousness would end in the brain replacement experiment. That is, if the scenarios are indeed equivalent. The brain replacement experiment only replaces the neurons, and Searle could also argue that the causal powers that produce consciousness are located in other biological parts of the brain that were left in, however likely or unlikely this may be.

Searle may also have another way out of the brain replacement experiment altogether, depending on how strongly he holds on to his claim that consciousness is a biological phenomenon. Professor Pentti Haikonen has argued that the brain replacement experiment would not shed any light on the question "can a computer program instantiate consciousness?" since the system would not represent a computer program at all. Haikonen says:

> This kind of [nano-machinery brain] would not be the digital circuitry found in computers, but rather an addition-, threshold-, and pulse circuitry of its own kind. […] The result would not be a digital computer, since even this new system would follow the brain's architecture and information processing methods which leave no room for computer programs or machine commands. (Haikonen, 2017, pp. 32–33.)

If Searle were willing to follow Haikonen's line of thought, he could similarly insist that the brain replacement experiment does not address the question of computer programs and consciousness in the first place. It is unclear whether he would be prepared to do this, however, since he claims that his water pipe entity would be unconscious and analogous to a program. On the one hand, Searle has insisted that consciousness is necessarily a biological phenomenon. But on the other hand, he does not rule out the possibility of conscious, non-biological machines, as long as they possess the same "causal properties" of consciousness that human brains do (Searle, 1990, p. 27).

[19] The second option is different from Russel and Norvig's conclusion 2 in that here, the subject acknowledges his fading consciousness out loud, whereas in conclusion 2, he was unable to voice his loss of consciousness to others.

It is here where Searle's argument appears to be resting on a more hypothetical foundation. Searle empathizes the point of formal processes of the brain not being sufficient for the causal powers of the brain, but he does not explain *what those causal powers are*, aside from being biological in an unspecified way. If we knew how our consciousness arises from brain processes, biological or otherwise, the whole debate around the CRA would be meaningless, as we could simply point out where the argument succeeds or fails. But we are ignorant on the matter, and Searle's insistence that consciousness is necessarily biological seems a little hasty. Our own brains clearly do produce consciousness somehow, so if a simulation is performed where every connection and interaction of the brain is taken into account, what causal powers do we miss?

Biological naturalism has had its doubters. Psychology Professor Steven Pinker from the University of Harvard has noted that it is not the neural brain tissue alone that creates a mind, since brain tumors, cultivated cell colonies and such are made of it but are (presumably) unconscious. Rather, according to Pinker, it is precisely the suitably arranged interconnectivity of the tissue and the information processing it does that gives birth to consciousness. (Pinker, 1997, p. 97). Similarly, Russel and Norvig point out Searle's silence on what exactly are the causal powers of the brain that constitute a mind, if neural processing is not part of it. They also note that neurons evolved to fill functional roles, and that it would be an amazing coincidence if they produced consciousness in a way that has nothing to do with their functional roles, as Searle claims. (Russel & Norvig, 2016, pp. 1032–1033.)

Intuition Reply: Since Searle is somewhat short on concrete evidence to back up his claims except the seemingly self-evident conclusion of his thought experiment, another counterargument against the CRA, the *Intuition Reply*, has been mounted. The Intuition Reply states that the CRA is not providing any *a priori* reason to exclude the possibility of Strong AI, but instead, it only reinforces our intuition that it is impossible (Cole, 2014, chap. 4.5). It is important to note that this would not in itself prove that the CRA is false, but it would rob it of much of its argumentative power, effectively discrediting the argument.

The Intuition Reply has been touched upon by a number of writers who complain that Searle sets up his thought experiment in a way that leads to intuitive reactions in his favor. For example, Dennett has argued that the CRA relies on people not thinking too much about the many faculties a Strong AI would need (Dennett, 1992, pp. 437–439). Boden has also claimed that the analogies Searle appeals to do not carry through, and that the intuitions they

draw on are unreliable (Boden, 1987, p. 6). However, a more thorough argument from the Intuition Reply has been developed by philosophy professors Paul and Patricia Churchland.

The Churchlands, who have both specialized in neurophilosophy and the philosophy of mind, seek to point out the CRA's flaws through a formally similar but blatantly false argument they have dubbed "The Luminous Room". They imagine a scenario in the 1860's, where James Maxwell fruitlessly attempts to prove his theory that electricity and magnetism are the same thing, and are sufficient to produce light (=electromagnetic radiation). One of Maxwell's contemporaries raises a magnet and waves it up and down in the air. No visible light is generated, even though, according to Maxwell's theory, it should generate electromagnetic waves and therefore light. The contemporary could then argue that Maxwell's theory is false. (Churchland & Churchland, 1990, p. 35.)

More precisely, the formal structure of the analogy the Churchlands are making is:

> Axiom 1. Electricity and magnetism are forces.
>
> Axiom 2. The essential property of light is luminance.
>
> Axiom 3. Forces by themselves are neither constitutive of nor sufficient for luminance.
>
> Conclusion 1. Electricity and magnetism are neither constitutive of nor sufficient for light. (Churchland & Churchland, 1990, p. 33.)

The man waving the magnet is the experiment to support Axiom 3. Likewise, Searle's larger argument as quoted by the Churchlands follows a similar structure:

> Axiom 1. Computer programs are formal (syntactic).
>
> Axiom 2. Human minds have mental contents (semantics).
>
> Axiom 3. Syntax by itself is neither constitutive of nor sufficient for semantics.
>
> Conclusion 1. Programs are neither constitutive of nor sufficient for minds. (Churchland & Churchland, 1990, p. 33.)

Where the CRA is the defense of Axiom 3.

The Luminous Room argument is, of course, glaringly unsound. The Churchlands argue that Searle is making a similar mistake with his argument. Maxwell's contemporary fails to create light because the wavelengths he produces are not energetic enough to be seen as visible

light (electromagnetic radiation is indeed produced in this scenario, albeit at too long wavelengths to observe with the naked eye). Maxwell's theory is not false, despite the apparent counter-argument against it; today we know better, but Maxwell's contemporaries might have concluded that light does not consist of electromagnetic waves since they were not apparently generated. The Churchlands argue that Searle is likewise misled:

> Even though Searle's Chinese room may appear to be "semantically dark," he is in no position to insist, on the strength of this appearance, that rule-governed symbol manipulation can never constitute semantic phenomena, especially when people have only an uninformed common-sense understanding of the semantic and cognitive phenomena that need to be explained. Rather than exploit one's understanding of these things, Searle's argument freely exploits one's ignorance of them. (Churchland & Churchland, 1990, p. 35.)

Searle's reply to the Luminous Room argument was to assert that the Luminous Room deals with entirely physical phenomena, while syntax holds no causal capabilities, and therefore the analogy fails:

> The account of light in terms of electromagnetic radiation is a causal story right down to the ground. It is a causal account of the physics of electromagnetic radiation. But the analogy with formal symbols fails because formal symbols have no physical, causal powers. The only power that symbols have, qua symbols, is the power to cause the next step in the program when the machine is running. And there is no question of waiting on further research to reveal the physical, causal properties of 0's and 1's. The only relevant properties of 0's and 1's are abstract computational properties, and they are already well known. (Searle, 1990, pp. 30–31.)

But syntactic symbol systems are always implemented on a physical system, which does have causal powers on the world. Does this make a difference?

Searle argues that this merely shifts the focus to the physical properties of the implementing system, which would undermine the Strong AI hypothesis (Searle, 1990, p. 31). An opponent might still object to this. An unimplemented program obviously has zero causal influence on the world – it needs to be implemented by a physical system for anything to happen. When a chess program defeats us in chess, we congratulate the *program's* skills, not the computer's, since it was only implementing the chess program. Nothing in the computer itself gave way for a cunning strategy and a victorious match. A program was needed to play the match, and therefore it could be said that the program had *some* kind of causal power on the world. If the program had zero causal influence, why couldn't the computer play the match without it?

In either case, it should be noted that the Intuition Reply can be utilized the other way as well. The Robot Reply, Systems Reply and Brain Simulator Reply could all be attacked by the Intuition Reply in that they do not *prove* that their respective scenarios would produce a Strong AI, but they only attempt to reinforce intuitions that they would. Whether a supporter of Searle would want to follow this line of defense is dubious, however, since it would also apply to the CRA itself.

Searle's wider point around the CRA was summarized in the formal formulation the Churchlands quoted. Searle's main thesis is that *minds are not programs*, and hence computer programs alone could never generate minds, or qualia. Note that this is not the same as denying that *machines* could think:

> "Could a machine think?" My own view is that *only* a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains. And that is the main reason strong AI has had little to tell us about thinking, since it has nothing to tell us about machines. By its own definition, it is about programs, and programs are not machines. Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle because of a deep and abiding dualism: the mind they suppose is a matter of formal processes and is independent of quite specific material causes in the way that milk and sugar are not. (Searle, 1980, p. 424.)

It is interesting to note that with his views Searle is strongly opposing the mainstream cognitive science paradigm, which studies human cognition and AI from essentially the same perspective, as complex information processing systems (Otto Lappi, Anna-Mari Rusanen, & Jami Pekkanen, 2018, p. 42). Were Searle to be correct with his wider argument, the functionalist foundations of cognitive science would mostly fly into the trash bin.

Searle's analogy that simulation of lactation is not lactation, and likewise simulation of consciousness is not the same as consciousness, has been challenged. Russel and Norvig point out to historical precedent of sorts, when artificial urea was first synthesized in the 1840s. People had to correct their views that inorganic chemicals could never produce organic material, since there was no discernible difference that existed between synthesized and natural urea; it turned out that organic material had no properties that could not be reproduced by

inorganic materials, despite common beliefs that this was impossible. They also argue further that computer simulations of addition or chess are not just simulations, but implementations, of the activity in question. (Russel & Norvig, 2016, p. 1027.) A computer simulation of milk certainly produces no milk, but it's not obvious from this that simulating consciousness could not produce consciousness. Searle appears to be playing at least a partially rhetorical game here.

To reiterate, the bottom line is this: Searle claims that computers cannot be conscious in virtue of running a program; computing ones and zeroes, or any other formal symbols, is not enough because a program cannot glean any semantic meaning from the symbols it manipulates alone. Searle says that it must be the unspecified biological, "causal" characteristics of our neural activity that produce a consciousness, but in a way that is not solely dependent on the neurons' functional roles (i.e. their information processing).

Searle's views appear to lend much of its credibility from the claim that one cannot unlock the semantic meaning of a symbol system just by studying that symbol system alone. Since a computer performs all of its operations and information processing in binary code, it follows that computers cannot understand semantics, and therefore (according to Searle) cannot have conscious experiences or qualia.

If humans are different from computers, then where is the foundation of *human* understanding of semantics? What is the answer to the so-called symbol grounding problem?

Haikonen has a seemingly obvious answer. Humans get their semantics from a sub-symbolic level – for example, a small child learns the meaning of the word *spoon* by having different kinds of spoons shown to her, while she gradually attaches semantic meaning to the word (Haikonen, 2017, pp. 43–44). This explanation is certainly true, but it does not address the question in its entirety, since it does not explain how semantics is gained from our *neural processing*.

The situation in the Chinese room is not entirely similar to a computer program. The man in the Chinese room is a conscious being who, when he is following the instructions, may think: *"I have absolutely no idea what I'm doing here."* In contrast, a computer is machinery that has no part of it thinking *"I have no idea what I'm doing."* The Chinese room scenario focuses its attention on an already-conscious agent, while a computer works on a far more foundational level. There appears to some form of level-of-explanation fallacy present. We can examine the computer as a formal symbol processor, but more fundamentally it functions

by manipulating electric currents, which is ultimately what the brains do as well (Russel & Norvig, 2016, p. 1032). It is a different matter to say that an agent with qualia does not understand something by following a syntactic program, and to say that qualia or consciousness cannot be gained from syntactic processes. Neuron firings carry no intrinsic semantic content with them. They happen solely in relation to neural and chemical activity around them, and Russel and Norvig have argued that they might be called syntactic (Russel & Norvig, 2016, p. 1032), although with a very high level of sophistication. Russel and Norvig interpret "syntactic" rather widely here. The idea seems to be that as a program is syntactic because its algorithms follow each action after certain criteria are met, and so too the brain operates under complex, causal rulesets that are similar enough to warrant a meaningful comparison. It is not clear whether Searle means entirely the same thing when he talks about syntactic processing, but Russel and Norvig's underlying idea appears to be to shift the conversation about consciousness to a lower level of explanation to sidestep Searle's criticism. If the brain is a syntactic machinery of a kind, then it follows that semantics *can* be gained from syntactic processes – brains are syntactic, yet we have semantics. Dennett has raised the same point (Dennett, 1992, pp. 438–439). Only the exact method, the hard problem of consciousness, is not yet understood. Searle has three options to respond to this kind of argument:

1. Insist that brains have as-of-yet undiscovered intrinsic semantic qualities that computers could not have.
2. Insist that the syntactic processing of the brain is different from the syntactic processing of a computer in some relevant sense.
3. Accept that semantics can be gained from syntactic processing, and let his argument collapse.

Since Searle insists that consciousness is caused by the unspecified biological or causal properties of the brain, he would seem to choose option 1. But this option is debated. Option 2 was briefly discussed earlier by Turing in the Argument from the Continuity of the Nervous System, but suffice to say now, it lacks concreteness. What if we chose option 3? What would be its implications for the Chinese room? If syntactic processing is enough for semantic understanding, then would the man in the room come to understand Chinese just by following the instructions?

The idea that thinking could be accurately replicated with a large number of formal symbolic representations of any problem or question (such as logical if-then -propositions) was the underlying assumption of early AI research, and was later dubbed "Good Old-Fashioned Artificial Intelligence" or GOFAI by John Haugeland (Haugeland, 1985, pp. 112–113). The man in the Chinese room appears to be following the GOFAI approach: he would correspond Chinese symbols with each other in accordance with a ruleset. But human cognition does not appear to follow such a narrow methodology. If a *suitable* program is required for semantic understanding, then it is entirely possible that the CRA scenario would not produce any understanding because it is not instantiating an appropriate program. After all, human cognitive processing where semantic understanding arises takes place under a highly interconnected neural network, not a simple rulebook or a look-up table. The more appropriate scenario that Searle could muster was his answer to the Brain Simulator Reply: would a man opening and closing a huge series of water pipes, running a suitably interconnected "program", attain understanding of Chinese?

It seems reasonable to say that the man would most likely not. He is only a part of the system, and no individual part of our brain understands anything by itself. Following option 3 the system as a whole might, which would lead us to some kind of endorsement of the Systems Reply. Many open questions are of course present, but Searle's water pipe scenario, designed for maximum absurdity, is based on the intuition that no understanding could be gained by such a system. If the water pipes acted under some kind of self-regulating mechanism, it could at least simulate the neural firings in the brain. The room for the pipes would have to be absolutely enormous, however, and it would operate many orders of magnitude slower than any computer or brain. Some commentators have argued that the speed of information processing is a relevant factor, and that Searle's various scenarios slow down the neural processing to a range where we would no longer attribute intelligence to it (Cole, 2014, chap. 4.5). If the water pipe system was shrunk down to the size of the brain and its performance somehow sped up to the same speeds, and the behavior it gave out appeared intelligent (as per Searle's scenario), our intuitions of its intelligence or lack of it might change. This is not the same as proving that the system is conscious, but in a rhetorical debate it might shift the burden of proof to Searle.

Even if we accept option 3, it remains unanswered *how* consciousness arises from syntactic processing. What kind of processing would be required? Is the speed a relevant factor? *Is* consciousness really a purely computational process, or does it leave out other factors that

may need to be taken into account, such as different architecture of the brain (massive parallel interconnectivity) to the computer (hyper-fast sequential processor)? Claiming that semantics can be gained from syntactic processing does not lead to every AI being conscious, and it alone does nothing to show what kind of information-processing would be required for consciousness. Thankfully, we need not attempt such an explanation here.

- -

Searle's Chinese room argument has had considerable influence on the philosophy of mind and philosophical discussions about artificial consciousness in the past forty years. The debate around the CRA is still on-going, testifying both to the strength and divisiveness of the argument. Since the CRA operates on *a priori* grounds – it's not an empirical demonstration of any kind – our intuitions around it appear to play a crucial role in assessing its validity. This notion has been articulated by Dennett, who has argued that the CRA acts as an "intuition pump", which misleadingly reinforces our intuitions towards the desired conclusion – in this case, that Strong AI is impossible (Dennett, 1992, p. 440).

Whether or not the CRA can be understood as an intuition pump, Searle's claim that syntactic processing is not sufficient or even constitutive of consciousness is a resilient argument against Strong AI. However, his other claim – that consciousness is necessarily a biological phenomenon – is less solid. Until Searle points out the exact causal, biological properties of the brain that give rise to consciousness, we have room to dispute its biological foundation as a necessity.

It is possible that Searle's position may be victorious one day, but as long as debate around the CRA continues, the possibility of Strong AI cannot be entirely refuted on these grounds. Besides, there are other ways to approach artificial consciousness that can more or less sidestep much of Searle's criticism. These approaches are studied in the next section.

3.3 Embodied Cognition

The Strong AI hypothesis, as it was formulated by Searle, claimed that consciousness is a matter of programming: any suitable program when instantiated produces consciousness, and that a Strong AI is necessarily a *program*. Turing also described his imitation game -

passing machine as a digital computer. But digital computers and programs are not the only avenues of creating autonomous entities. In fact, if we want to avoid the bog that is the CRA, we may well want to follow a different approach.

In terms of robotics, a goal of conscious *machine* (as opposed to conscious *AI*) is sometimes referred to as *machine consciousness* or *artificial consciousness*, abbreviated MC or AC, respectively (David Gamez, 2008, p. 887). Its main difference to Strong AI is that MC discusses consciousness in terms of robotics instead of AI programs. The difference may be somewhat semantic, and it is brought up here just to avoid future confusion. It is true that AIs are mostly associated with programs and digital computers, but the definitions of AI discussed in chapter 2.1 do not make this a logical necessity. For the purposes of this thesis, we can use MC and Strong AI as synonyms of each other, even if Searle intended the latter to refer to programs only. We can think of AIs more broadly than that: as any *artificially* created *intelligence*, regardless of the exact design approach chosen. For now, we will use the terms "Strong AI" and "machine consciousness" synonymously.

Gamez (2008, p. 888) divides research of machine consciousness into four sub-sections, which are:

> MC1: Machines with the external behavior associated with consciousness.
>
> MC2: Machines with the cognitive characteristics associated with consciousness.
>
> MC3: Machines with an architecture that is claimed to be a cause or correlate of human consciousness.
>
> MC4: Phenomenally conscious machines.

We are interested in MC4 specifically, although it is clear that there is significant overlap between the four approaches. We will highlight one consideration to consciousness we have not discussed yet: embodied cognition.

Embodied cognition is a view of cognitive science which claims that the body of an agent, not just the brain, is a relevant factor in cognitive functions. Robert Wilson and Lucia Foglia have termed this the "Embodiment Thesis":

> Many features of cognition are embodied in that they are deeply dependent upon characteristics of the physical body of an agent, such that the agent's beyond-the-brain body plays a significant causal role, or a physically constitutive role, in that agent's cognitive processing (Wilson & Foglia, 2015, chap. 3).

Embodied cognition is thus contrasted with mainstream cognitive science, which studies cognition from a purely information-processing point of view. It should be noted that several aspects of cognition, such as memory or calculation, are already performed in a vastly superior way by computers than humans. Embodied cognitive science thus draws from the idea that in order for us to possess *all* the mental capacities that we do, it is necessary to involve the body both as a sensing and an acting entity. Cognition at large, and by extension, consciousness, are highly dependent on these factors. (Wilson & Foglia, 2015, chap. 1.) What exactly constitutes embodiment is not entirely uncontroversial. In robotics embodiment is usually understood simply as a physical/sensorimotor interface, but other academics hold that biological aspects of a living body must be taken into account as well (Tom Ziemke, 2016).

Embodied cognitive science creates some interesting considerations for Strong AI. Since AIs are primarily information-processing agents, and as such quite disembodied, is there any hope for consciousness to be present in any kind of AI if the central thesis of embodied cognition is true?

Here we will take two approaches that could be taken to answer the question: one practical, and one philosophical conjecture. The conjecture approach is through a famous thought experiment, the so-called "brain in a vat", or BIV for short.

The BIV thought experiment asks us to imagine a scenario where a human brain has been removed from its host body and placed into a vat, with all the nutrients and other necessities which enable the brain to remain alive. In the vat, the brain is connected to an intricate computer using the same exact neural connections that the brain was connected to when it still was in a real body. (Lance Hickey, n.d.) The computer receives neural signals from the brain and sends back signals that the brain cannot distinguish from neural signals it received by the body – for example, if the brain sends a signal to raise its body's hand (which it no longer has) the computer sends back a signal that corresponds to a sensation of a hand being raised. The brain could be fed sensory inputs of an entire virtual reality, which it could not distinguish from the real world. (Hickey, n.d.)

The BIV has mostly been used in support of skepticist or solipsist arguments in epistemology,[20] but it also provides an interesting perspective into embodied cognition and consciousness, since the experiment can be used to analyze the logical consistency of embodied cognition. If the brain in a vat was capable of cognitive functions, or even better, qualia of the simulated inputs, it would entail that cognition would not necessarily require a biological body *per se* to function (unless a biological brain is sufficient), only appropriate information input, with no concern over what the exact source of that input is. We have noted earlier that neural signals carry no known intrinsic semantic content with them. They are just electrical impulses, which could in principle be substituted by electrical impulses from some other source. The BIV would therefore put the central thesis of embodied cognition into doubt, if its logical validity is accepted, and possibly open the door for conscious, disembodied AIs.

Unfortunately things are not this simple. Evan Thompson and Diego Cosmelli have argued that the minimum requirements to sustain a BIV and to allow it consistent simulated interaction include contrivances so complex that they would essentially equal an artificial body. These include physical protection to replace the skull, a circulatory system with responsiveness to the brain's needs at any given moment (and all the necessary sub-systems to oxygenate and pump blood etc.), ability to react appropriately to the brain's intrinsic neural activity, feeding so many correct neural signals at once that even the fastest computers could not keep up – all this and much more without disrupting the balance of the whole system. Thompson and Cosmelli argue that for all this to work, we need the vat's workings to be regulated and controlled by the brain itself and even provide it with real[21] sensorimotor systems it can control to work properly. In other words, envatting the brain ultimately ends up with embodying it. (Thompson & Cosmelli, 2011, pp. 168–172.)

The real possibility of a BIV therefore appears slim. The practical considerations are daunting – but what about its *logical* possibility? After all, this is all we would need for purposes of AI. If a conscious BIV is logically conceivable, disembodied entities could not all be excluded from consciousness.

Thompson and Cosmelli concede that if we are interested only in the logical consistency of BIV, the practical considerations they raise are irrelevant. However, they note that as a mere

[20] It shares notable similarities with Descartes' evil demon and Zhuangzi's "I dreamt I was a butterfly" -scenarios.

[21] Thompson and Cosmelli don't really clarify why sufficiently advanced virtual sensorimotor systems do not suffice, except by pointing out the computational difficulties. Similar practical considerations have been raised by Dennett (1992, pp. 4–6).

thought experiment it tells us nothing about the mechanisms behind consciousness, since it sheds no light on consciousness' explanatory framework. We'd still have to struggle with what constitutes the basis for consciousness and even the brain itself – do we need to envat only neural cells, or also glial cells, the immune system etc.? Thompson and Cosmelli say that we "don't know what we're supposed to imagine when we imagine a brain in a vat, so the mere conceptual possibility of a brain in a vat seems an empty scenario." (Thompson & Cosmelli, 2011, pp. 174–175.)

For our purposes of course we want to imagine a conscious, disembodied entity. If this is logically impossible, then AIs as disembodied entities cannot exist, and would need some form of embodiment (and possibly more) to be conscious. Thompson and Cosmelli raise a valid point when they argue that the BIV as an imaginary scenario tells us little about consciousness. Does it tell anything about AI? It seems clear that the human brain would have severe cognitive difficulties in an unstimulated disembodied environment, but how about the cognition of AIs?

The BIV in relation to AI has been considered by Francis Heylighen, who takes a very critical stance to disembodied AIs. Heylighen argues that an AI with no sensors or actuators in the real world would be virtually useless as a cognitive agent:

> A virtually imprisoned AI program is even worse than a brain in a vat, as it simply has no sensors, no effectors, no body, and no real world to interact with. Therefore, it cannot be intelligent in the sense of being an autonomous, adaptive cognitive system that can deal with real-world problems and steer its own course of action through a complex and turbulent reality. At best, it can be a very sophisticated expert system that can solve chess, Jeopardy!, and similar highly artificial and constrained puzzles and games given to it by its designers, or help them to mine massive amounts of pre-formatted data for hidden statistical patterns. (Heylighen, 2012, pp. 132–133.)

Heylighen's criticism rings close to Thompson and Cosmelli's. He appears to deny that an AGI could exist without some kind of embodiment, regardless whether it is a Strong or Weak one. Heylighen is certainly right in the sense that contemporary AIs' field of expertise is data analysis in one form or another. However, his view that an AI "imprisoned" to a virtual reality is not equipped to deal with real-world situations intelligently and adaptively is somewhat ambiguous for our purposes. The claim is easier to accept if Heylighen means that an agent with no sensory information *at all* is useless as a cognitive agent (indeed, Heylighen's

quote appears to suggest such an interpretation). But couldn't an AGI learn to be an "autonomous, adaptive cognitive system" if it receives systematic input and output data that consistently models real life? Whether this sort of system would be capable of experiencing qualia is of course a different story. Once again, we do not know enough about the mechanisms behind consciousness to say anything certain about the necessity of embodiment.

In either case, we have learned that embodied cognition could assert a limitation of Strong AI in three ways. The most moderate position would be to say that consistent sensory information and interaction is required for consciousness, whether it happens through a physical or a simulated body. This interpretation could be attributed to Heylighen under a good-natured reading, although it is not obvious he means precisely this. The stronger position could assert that a physical body is necessary for consciousness, and no simulated body can ever do the job. This interpretation can be gleaned from Thompson and Cosmelli. The strongest position says that not only is a physical body necessary for consciousness, but it also has to be a *biological* body with all its biological properties. This is the core tenet of Searle's biological naturalism. The first position allows Strong AI to exist without a physical body. The second denies it if no physical body of some sort is present, and the third would require the utilization of some kind of bio-technology that would probably be quite detached from current research of machine consciousness. It seems clear that the third position would limit the possibility of Strong AI the most, although at the present day it is not possible to conclude which interpretation, if any, is closest to the truth.

Putting aside philosophical though experiments, the practical approach to embodied cognition is to develop AIs with sensing and acting bodies – robots. Robotics has long existed as a field of its own, and it contains a plethora of different design approaches to machine consciousness that could warrant an entire article of its own. For the sake of brevity, we will only highlight one approach as an example: Professor Pentti Haikonen from the University of Illinois and his associative neural model of consciousness.

Haikonen starts out with simple enough axioms. He says that explaining consciousness is about explaining how qualia are attained from the brain's neural processing, and explaining the exact contents of our mental states is not part of this task. Haikonen notes that we never experience our brains' neural processing as neuron firings, but rather, as qualitative, subjective experiences. (Haikonen, 2017, pp. 196–197.) He further notes that our conscious mental contents consist only of perceptions that we can remember for at least some time, either in

the form of real perceptions from our senses or as imaginary ones. From this, he concludes that perception is the brain activity that deals with consciousness, and that *consciousness is the same as qualitative perception, or qualia*.[22] (Haikonen, 2017, p. 197.)

But why do we perceive our neural signals as qualia, and not as neural signals? And why cannot the qualia themselves be observed by outside instruments?

According to Haikonen, the answer is that we do not have the right instruments to measure qualia. The precise nature of measurements depend on the instrument used, even if the physical quantity measured is the same. When we measure temperature with an analog thermometer, the result is an expansion of mercury; when we measure the same quantity with a digital meter, the result is a numeric representation. Likewise, measuring anything with a measuring tape always produces a length as the result, regardless of other qualities of the measured object. Haikonen claims that this is the reason EEG scans cannot observe qualia: they always measure neural activity alone and nothing else. To measure qualia, we would need a different kind of instrument altogether – and the only one we know of is the brain itself. (Haikonen, 2017, pp. 200–201.) Haikonen claims that no further explanation is needed on how the neural signals are transformed into qualitative experiences. Neural signals are perceived as qualia because there is no other choice: we have no sensory organs to perceive them as neural signals. Haikonen compares the explanation to a radio, where we can hear the music, but not the carrier radio waves that carry the music. (Haikonen, 2017, p. 203.)

Haikonen's theory of consciousness has some interesting implications, and at least one objection could be immediately raised against it. An opponent might say that there *is* a choice on how neural signals are perceived: that they are not perceived at all. An opponent could say that Haikonen dodges the question of emerging qualia and still owes an explanation on how the process works, exactly. If we need sensory organs to perceive neural signals, why don't we need them for perceiving qualia? The explanatory gap between brain activity and subjective experiences appears to still be present, although perhaps in a narrower form.

In robotics, Haikonen's model permits Strong AI as long as the AI is an embodied agent that is equipped with an appropriate associative neural network. This is how embodied cognition can address the problems associated with the CRA. As readers hopefully remember, Searle

[22] In Haikonen's model, "real" and "virtual" (that is, imaginary) perceptions are treated more or less equally. Since Haikonen associates consciousness with qualia, even people sleeping are considered "conscious" in the sense that they are capable of subjective experiences while asleep.

claimed that a computer program cannot be conscious since it only manipulates formal symbols, and syntactic symbol manipulation is not enough for semantic understanding. Haikonen is happy to embrace this assertion (Haikonen, 2017, p. 53). The workaround for Searle's argument is simple: since syntactic symbol manipulation is the problem, the agent's cognitive functions must be based on an entirely non-digital and sub-symbolic level similar to the human brain to avoid the CRA's objections. This is achieved by a physical, not simulated (otherwise the CRA applies), connectionist system which mimics the human nervous system and neural activity, allowing the agent to create and associate symbolic meanings to its sub-symbolic observations on its own. (Haikonen, 2017, p. 214.)[23] No digital parts may be used, since the binary code they use is symbolic. The neural network Haikonen proposes as an appropriate architecture is named, quite simply, *Haikonen Cognitive Architecture*, or HCA. Haikonen even claims to have succeeded in creating a successful model in a low scale: his self-made robot XCR-1 is built following Haikonen's own HCA architecture, and appears to exhibit rudimentary forms of awareness with no digital circuitry used in its making (Haikonen, 2017, pp. 231–237; Haikonen, 2015.)

In terms of *human* consciousness, however, Haikonen's model creates a consequence he appears to be unconcerned of. Since consciousness is simply a way of perceiving, it is not a *cause* of anything in itself, it is not an agent that *does* anything (except perceive). Since our subjective mental states are not responsible for moving the body, Haikonen's model appears to reduce consciousness to an entirely epiphenomenal role. While the same physical human entity makes decisions and acts, those decisions are not decided consciously, only *perceived* as such. Consciousness only observes the neural activity associated with making the decisions, which may give an *impression* of a conscious choice, but the choice was not made truly consciously. Haikonen himself acknowledges this in passing (2017, p. 204), although he does not seem to give much thought to its implications. Yet epiphenomenalism would probably have quite dramatic consequences for our understanding of free will and sense of selfhood. Whether this is an argument against Haikonen is up to debate. Many people are probably unwilling to yield to epiphenomenalist arguments, but on the other hand Haikonen's position cannot be rejected simply because it would radically change our conventional beliefs about consciousness and everything associated with it.

[23] Since Haikonen believes that all the required aspects of consciousness can be replicated with non-biological hardware, he sees that a body does not have to be biological to enable consciousness.

Back in robotics, Haikonen asserts that perceiving qualia requires a specific neural architecture and sensory information. Since an outside observer cannot directly detect qualia, how does Haikonen determine whether a robot is conscious or not?

Haikonen rejects the Turing Test and the behavioristic approach as wholly inadequate solution. He says that the Turing Test "does not measure the mental properties of a computer, it measures how gullible people are." (Haikonen, 2017, p. 258.) While Haikonen acknowledges that we have no direct means of ascertaining the presence of consciousness in a robot, he claims that there are two possible ways of possibly determining it. One is to inflict pain on the robot and to observe whether its behavior (and cognitive architecture) supports a subjective mental state of physical suffering, since Haikonen believes that feeling pain requires consciousness and "even the most primitive creature can feel pain".[24] The second way is to wait until a possibly conscious robot starts asking existential questions like "Why am I here?" (Haikonen, 2017, pp. 260–261.)

Unfortunately Haikonen's own suggestions fall short. While Haikonen is contemptuous towards the Turing Test, it is clear that his own "tests" are ultimately behavioristic as well. The main difference is that Haikonen requires an appropriate cognitive architecture to be present (similarly to Searle), but in the end we are still analyzing the robot's behavior to deduce its consciousness. Why couldn't a Weak AI in a robotic body behave as if it was in pain? Or ask existential questions? Haikonen's model follows an unfortunate circular line of reasoning. He claims that his cognitive architecture creates consciousness because it demonstrably creates appropriate behavior, and that the behavior is conscious because it is generated by the architecture. This fault is not, however, to be blamed solely at Haikonen. The same problem plagues every model that attempts to create conscious robots, and Haikonen works with what he has. Until we have some kind of test that empirically solves the problem of other minds beyond reasonable doubt, we are forced to rely on observation of behavior and the somewhat circular reasoning that accompanies it.

- -

Embodied cognition provides both a challenge and an opportunity for conscious machines. On the one hand, it places a great obstacle – a need for some form of sensorimotor interface or other kind of embodiment to attain conscious machines, but on the other hand, it enables

[24] Remarkably, Haikonen gives zero thought to the ethics of essentially torturing a robot to determine whether it's conscious or not. While the scenario carries some absurd elements, it is easy to find such testing inhumane.

approaches which can meet many of the challenges from the CRA. Of note is that utilizing a physical, non-digital connectionist cognitive architecture as proposed by Haikonen is not exactly what Searle meant when he talked about Strong AI. Searle's definition of Strong AI was that an appropriate *program*, when instantiated, is conscious, and since Haikonen's architecture is non-digital and contains no programs, it is not a Strong AI in the Searlean sense of the word even if conscious. The term "machine consciousness" should perhaps be preferred here, although the difference will be discussed further in the Conclusions section.

The exact prerequisites that embodied cognition may place to artificial consciousness are still somewhat shrouded in uncertainty, as demonstrated by the BIV scenario. This gives much room for different approaches in science and philosophy to study artificial consciousness, and much ground remains uncovered. Haikonen's architecture is only one of many proposed models for consciousness, each of which is undergoing debate over their respective merits and deficiencies.

4. CONCLUSIONS

In this thesis we have gone through several philosophical and scientific positions regarding the possibility of Strong AI. True to the finest traditions of philosophy, there is little consensus among the practitioners of the field as to whether AIs or machines can possess a subjective consciousness. Positions that support the existence of Strong AI that we have investigated here include mainstream cognitive science (often discussed under the names of *functionalism* or *computationalism*), embodied cognition (with some reservations) and philosophical behaviorism (such as the Turing Test). A central opposition against Strong AI is mounted by John Searle and his Chinese room argument, and other proponents who support his position in varying degrees. Embodied cognition also limits the possible design strategies for Strong AI to embodied agents, although it does not categorically rule out conscious machines. What are we to make of this discord?

The classical attempt to gauge the consciousness of machines has been the Turing Test, which – unlike most proponents – provides empirical criteria for attributing consciousness to machines. Unfortunately Turing's proposal is not rigorous enough to provide sufficient certainty of what he is trying to measure. As such, the Turing Test has seen its fair share of justified criticism over the perceived criteria it sees sufficient for consciousness, and over its emphasis on outward behavior only. As we saw, the conclusions drawn from the test can differ depending on how the test is interpreted: seeing passing the Turing Test as a logically sufficient or necessary condition for attributing consciousness makes the test more difficult to defend given its specific nature, but the test is somewhat vindicated if it is interpreted as providing increased confidence in attributing consciousness to machines or AIs. However, even this moderate interpretation of the test leaves much to be desired. The test cannot take into account the difference between Strong and Weak AIs, which is arguably an important conceptual contribution from philosophy to information sciences.

Despite its limitations, the Turing Test is not a forgotten piece of behaviorist zeitgeist of its time, and the reason for this is simple. As was noted by Searle and indirectly by Haikonen, behavior is only one aspect of justifying attributing consciousness to other entities; the other is an *appropriate physiology or cognitive architecture*. Behavior is not enough; we also need something that we know can enable consciousness, otherwise we are dealing with Weak AIs.

But we do not know what an appropriate cognitive architecture looks like. There is no consensus on whether it has to be biological, connectionist, sub-symbolic, or even embodied. This lends undue credence to behavior. While there is virtually zero consensus on what kinds of cognitive architectures allow consciousness, there is no serious debate over what kind of behavior appears conscious. This means a temptation to lean more heavily on behavioral evidence. Our only current alternative to philosophical behaviorism is to make *a priori* assumptions about any given cognitive model, and then see if that given model permits Strong AI to exist. Since we do not know which model is closest to the truth, making educated guesses based on the behavior of an intelligent agent must be part of our toolkit for the foreseeable future. This probably tells more about our current understanding of consciousness than about the merits of philosophical behaviorism, but for now, our ignorance seems to salvage the Turing Test from complete obscurity.

The most vigorous objection to Strong AI comes from John Searle. Does the CRA succeed in its quest to stamp out the false anthropomorphizing of advanced future AIs?

Searle has certainly done an admirable job in defending his argument from a barrage of objections over almost forty years. In the Embodied Cognition section of this thesis we treated the CRA's objections as valid, but this was mainly due to caution and not because of endorsement of the argument – given the resilient nature of the CRA in academic debate, I think it is wise to think ahead and seek out approaches that can respond to the argument even if its validity is not fully accepted. I hope that I have given a thorough enough treatment of the CRA and its usual objections earlier in this thesis, so here I will content myself to providing two observations about the CRA which I believe are worth highlighting.

The first observation is that the standard formulation of the CRA appears to be invalid, and that it needs modifying to be a credible argument. Searle claimed that if Strong AI were possible, a man corresponding formal symbols with each other would come to understand Chinese because he is simulating a program that appears to understand Chinese. But the Chinese room is a huge look-up table program, and it is not necessarily the right *kind* of program to instantiate consciousness. Searle did not claim that according to Strong AI, *any* intelligently behaving program is conscious, but only that an *appropriate kind* of program is conscious. It seems unlikely that human cognition works like a look-up table, or that a look-up table program could be conscious. While symbolic AIs do function more or less as such, many contemporary ones do not – connectionist approaches to AI are on the rise, with much

closer (if still non-identical) similarities to human cognitive processing. The standard form of the CRA fails because the Chinese room instantiates a program that most likely cannot support consciousness – only behavior that *appears* conscious at best, much like what some modern AIs can manage under ideal conditions. The man in the Chinese room (or even the whole system, as the Systems Reply would assert) would not understand Chinese because a look-up table program cannot be conscious, and thus Strong AI is not refuted by the standard formulation of the CRA.

An appropriate Strong AI program would more likely mimic the human brain activity and resemble some kind of massively advanced connectionist AI – essentially, the Brain Simulator Reply to the CRA. As the reader hopefully remembers, Searle was not unarmed against this response. He responded with a scenario where a man operates a complex system of water pipes and valves, simulating a connectionist cognitive network and its neural firings. Searle claimed that such a man would be equally clueless as the first one, and that the CRA is still valid. In other words, the Chinese room is modified in a way that takes into account the objection above.

This modified version of the CRA dodges my argument that the standard CRA fails, and is where opponents of Searle would probably benefit in focusing their attention on. I will not commit to endorsing either position, but I want to note two approaches that opponents of the modified CRA might want to take.

One approach could assert that the speed of neural processing is a crucial factor in creating consciousness, and that a man simulating brain activity with pipes and valves cannot even in principle simulate it fast enough that it would ever produce consciousness. While a sufficiently fast computer could conceivably simulate a connectionist system with complexity equal to the human brain fast enough, a human being could never do this in a lifetime. However, we do not know if speed is a relevant factor, so this objection is subject to empirical evidence.

Another approach is to bite the bullet and say that yes, consciousness is indeed created by simulating the appropriate brain activity with water pipes and valves. To salvage this position from some absurdity, it would likely be wise to specify that the *connectionist system*, not the man operating it, would create consciousness, in the spirit of the Systems Reply. How well does Searle's response to the Systems Reply hold in this connectionist version?

Could a man internalize and perform all the cognitive processes of the brain – but isn't he already doing that, with his brain?

The second observation I want to make about the CRA is that some concepts appear blurry in a way that muddies the waters in Searle's favor. Searle's argument stresses that syntactic symbol processing cannot account for consciousness. Fair enough. But two things need to be noted: first, computers are physical, causal systems, and as Russel and Norvig said, binary code can be seen as an interpretation of electric current manipulation and not only as a formal symbol system. Second, human neural processes are not intrinsically semantic, but operate as a highly sophisticated causal system where subjective consciousness is generated through unknown means. While it is clear that computers follow a more clearly identifiable syntax, it is an open question whether the fine-grained causal nature of human brain activity is qualitatively different from computer processing in a way that is relevant to consciousness.

This line of argumentation shifts some of the focus to the qualities of the implementing system in a way that appears to contradict Searle's definition of Strong AI. Suddenly, the focus is not only on what the program does, but also on what the *computer* does. While it may be difficult to separate the two in practice, Searle noted that if we are concerned with the properties of the implementing system, then Strong AI *as a mere program* is refuted.

I believe that this kind of defense of Strong AI may in fact be perfectly valid. However, it would mean that we would have to slightly redefine the prerequisites of Strong AI. It would be an *appropriate program implemented by an appropriate physical system*, as opposed to appropriate program implemented by *any* physical system, as Searle suggested. What are the necessary qualities of an "appropriate physical system" are unknown and would in any case need a much more elaborate and thorough defense than what can be provided here. Still, it would be able to bypass the criticism that any physical system can implement a Strong AI.

Discussion over Strong AI forced us to juggle between the words "conscious AI" and "conscious machine". The two concepts they represent – Strong AI and machine consciousness – clearly have much in common, but if embodied cognition is to be believed, some form of embodiment may in fact be necessary for consciousness. This suggests that we may want to abandon Strong AI in favor of MC. This, however, depends on what kinds of need embodiment actually places on consciousness. As we saw, a loose interpretation is that even simu-

lated embodiment might be sufficient, which would still make Strong AI as a program possible if it is provided suitable sensory information. Still, the empirical foundations of this position do seem a little short on concrete evidence.

Searle was adamant that an appropriate physical body is indeed needed for consciousness, but it would have to be a *biological* body – biological naturalism. While some interpretations of embodied cognition move along similar lines, Searle does not provide a convincing account to defend his position. Searle's defense relied on analogies made from lactation and photosynthesis, but as Russel and Norvig pointed out, these analogies alone do not prove that consciousness cannot be duplicated by replicating formal cognitive processes – playing chess with a computer is duplicating, not simulating, the game under a different implementation.

Pentti Haikonen represented a practical approach to machine consciousness. Haikonen's model of consciousness allowed artificial consciousness to exist, and the cognitive architecture he proposed for conscious robots was similar enough to human cognition that he could claim it permits consciousness to exist in robotic bodies. The architecture was a sub-symbolic, connectionist neural network coupled with a sensorimotor system, and this way evades Searle's criticism over symbol manipulation and lack of embodiment.

Key factors for Haikonen's approach are whether his model of consciousness is close enough to the truth to work, and whether the cognitive architecture he proposed is truly sufficient for machine consciousness. Some philosophical shortcomings were also detectable in Haikonen's writing. His model narrowed the explanatory gap between physical brain activity and consciousness, but did not appear to completely close it. Haikonen's suggestion on how consciousness is observed in robots also suffered from behaviorist undertones, even though he rejected behaviorism as a valid position in determining consciousness. To be fair, Haikonen also acknowledged the difficulty of the problem of other minds, and granted that it may be impossible to prove beyond any reasonable doubt that any machine is in fact conscious.

The central difference between the various views on Strong AI seem to boil down to on how tightly consciousness is tied to embodied systems, biological or otherwise. There is plenty of room for disagreement, since the only demonstrably conscious entities we know of are biological in nature. We know from experience that a biological body is a logically sufficient condition for consciousness to exist, but we do not know whether it is a logically *necessary* condition. Be that as it may, the shadow of the Chinese room – while inconclusive – still

looms over the proponents of Strong AI as a disembodied, symbol-manipulating artificial intelligence.

In the introduction, I very briefly explained why it is important to know whether Strong AI or MC can exist at all. The reasons were mostly ethical: the societal impact of conscious AIs or machines would most likely be severe. AIs and machines of the present day are people's property, and if a Strong AI/MC were ever to be created, by any modern law it would legally be the property of its owner. This is not a problem today, but if we end up owning conscious and quite possibly very human-like robots, we would have to seriously re-think our moral and legal commitments towards them. It seems apparent that if Strong AI/MC is possible and if it is someday created, there is a dire need for a serious discussion over the moral and legal rights of AIs or machines. It is conceivable that we do not even need to decisively prove that the AI is a *Strong* AI. All that is necessary is a public consensus – even if uninformed – that an AI is humane enough for ethical questions to rise to the surface, as Turing predicted with his "polite convention".

What could be the starting points of this kind of ethical debate?

The first questions would most likely concern the AI's agency, general and moral. What kinds of cognitive capabilities does the AI have? What kinds of needs does it have, intrinsic or otherwise? What is the extent of its moral agency – is it capable of moral consideration, should we expect it from it, and how do these relate to the amount of moral and legal rights we decide to bestow to it? These questions have not seen easy answers even when applied to humans, but they are probably central in a philosophical discussion over the rights of machines.

An interesting "precedent" of sorts for a moral debate also comes from animal rights. While discussion over the rights of animals can be traced all the way back to antiquity, the traditional Cartesian way of thinking about animals' inner life has held for a long time, and has only changed radically over the last few decades. Descartes saw animals as machines with no feelings, emotions, or thoughts, and this draws significant parallels to machines. We now see animals as thinking, feeling beings, and although this has not translated into an end of use of animals for production, it is worth asking if there are similarities that may warrant a comparison to AIs or machines. While animals were – and still are – seen as tools for production, this perspective is at odds with animals as thinking beings. A similar path with advanced AIs or machines does not sound completely implausible.

Another meaningful analogy – perhaps even more appropriate – is one of slavery. Since AIs and machines are people's property, the status of a Strong AI or MC would very much resemble a slave, even more so than an animal. The consequences of accepting this as a *de facto* situation would be tremendous. A return to slavery? Our moral consensus has moved past that institution. Even though slavery is still widely practiced today in the form of human trafficking, all legislatures in the world officially prohibit it. This is in relation to *humans*, however. Could slavery be a valid form of legislature for machines – a compromise between moral or legal rights and practicality? Would slavery be as reprehensible if we created a Strong AI that *wants* to be enslaved? A "moral compromise" may certainly raise many justified red flags, but our current legislature does seem to be heading that way if no changes are made when the issue first rises in practice.

All of these considerations are dependent on the assumption that Strong AI or MC is possible, not only in some possible world, but in reality. It may be wise to plan ahead for low-chance, high-impact contingencies, but as we have seen, our current understanding of consciousness does not make Strong AI an eventuality. Yet the possibility cannot be entirely dismissed, either. The jury is still out, so perhaps it is wise to close with a quote from Professor Aaron Sloman from 1996:

> I am embarrassed to be writing about consciousness because my impression is that nearly everything written about it, even by distinguished scientists and philosophers, is mostly rubbish and will generally be seen to be rubbish at some time in the future […] This means my own work is probably also rubbish.

Words to live by.

5. REFERENCES

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., J. Hall, J. S., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S. C., Sowa, J. (2012). Mapping the Landscape of Human-Level Artificial General Intelligence. *AI Magazine, 33*(1), 25-42.

Avramides, A. (2019, May 2). *Other Minds*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/other-minds/

Blackburn, S. (1996). Other Minds. In *The Oxford Dictionary of Philosophy.* Oxford: Oxford University Press. Retrieved from: https://epdf.pub/the-oxford-dictionary-of-philosophy-oxford-paperback-reference.html

Block, N. (1981). Psychologism and Behaviorism. *The Philosophical Review, 90*(1), 5–43.

Boden, M. (1987). *Escaping from the Chinese Room*. Retrieved September 3, 2019, from http://shelf2.library.cmu.edu/Tech/19297071.pdf

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Bringsjord, S., & Govindarajulu, N. S. (2018, July 12). *Artificial Intelligence*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/artificial-intelligence/

Buckner, C., & Garson, J. (2019, August 16). *Connectionism*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/connectionism/

Chrisley, R. (2008). Philosophical Foundations of Artificial Consciousness. *Artificial Intelligence in Medicine, 44*, 119–137.

Churchland, P. M., & Churchland, P. S. (1990). Could a Machine Think? *Scientific American, 262*(1), 32-39.

Cole, D. (2014, April 9). *The Chinese Room Argument*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/chinese-room/

Dennett, D. (1992). *Consciousness Explained.* New York: Back Bay Books.

French, R. (1990). Subcognition and the Limits of the Turing Test. *Mind, 99*(393), 53–65. Retrieved April 5, 2019, from https://pdfs.semanticscholar.org/db2e/35bb875bce89474e60e814881dff30fb3228.pdf

Gamez, D. (2008). Progress in Machine Consciousness. *Consciousness and Cognition, 17*(3), 887–910.

Haikonen, P. (2015, November 2). *BICA2015 XCR-1 robot demo*. Retrieved April 23, 2019, from YouTube: https://www.youtube.com/watch?v=q0pIlc-MnaY

Haikonen, P. (2017). *Tietoisuus, tekoäly ja robotit.* Helsinki: Art House.

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

Heylighen, F. (2012). Brain in a Vat Cannot Break Out. *Journal of Consciousness Studies, 19*(1–2), 126–142.

Hickey, L. P. (n.d.). *The Brain in a Vat Argument*. Retrieved April 16, 2019, from The Internet Encyclopedia of Philosophy: https://www.iep.utm.edu/brainvat/

Hyslop, A. (1995). *Other Minds.* Dordrecht: Kluwer Academic Publishers.

Lappi, O., Rusanen, A.-M., & Pekkanen, J. (2018). Tekoäly ja ihmiskognitio. *Tieteessä tapahtuu, 36*(1), 42–46.

Manzotti, R., & Tagliasco, V. (2008). Artificial Consciousness: A Discipline Between Technological and Theoretical Obstacles. *Artificial Intelligence in Medicine, 44*, 105–117.

McDermott, D. (2014). On the Claim that a Table-Lookup Program Could Pass the Turing Test. *Minds & Machines, 24*(2), 143–188.

Mill, J. S. (1865). *An Examination of Sir William Hamilton's Philosophy* (2nd ed.). London: Longman, Green, Reader and Dyer.

Oppy, G., & Dowe, D. (2016, April 9). *The Turing Test*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/turing-test/

Pargetter, R. (1984). The Scientific Inference to Other Minds. *Australasian Journal of Philosophy, 62*(2), 158–163.

Pinker, S. (1997). *How The Mind Works.* London: Penguin Books Ltd.

Raatikainen, P. (2015, January 20). *Gödel's Incompleteness Theorems*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/goedel-incompleteness/

Rey, G. (1986). What's Really Going on in Searle's "Chinese Room". *Philosophical Studies, 50*(2), 169–185.

Russel, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach* (3rd ed.). Harlow: Pearson Education Limited.

Searle, J. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences, 4*, 417–457.

Searle, J. (1990). Is the Brain's Mind a Computer Program? *Scientific American, 262*(1), 26–31.

Searle, J. (2001). The Failures of Computationalism. *Psycoloquy, 12*(60). Retrieved June 17, 2019, from http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?12.060

Searle, J. (2002). *The Rediscovery of the Mind.* Cambridge, MA: The MIT Press.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutsgever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, Thore., Hassabis, D. (2017). Mastering the Game of Go Without Human Knowledge. *Nature, 550*(7676), 354–359.

Sloman, A. (1996). A Systems Approach to Consciousness. *RSA Journal, 144*(5470), 40–46.

Thompson, E., & Cosmelli, D. (2011). Brain in a Vat or a Body in a World? Brainbound versus Enactive Views of Experience. *Philosophical Topics, 39*(1), 163–180.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind, 59*(236), 433–460.

Wilson, R. A., & Foglia, L. (2015, December 8). *Embodied Cognition*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/embodied-cognition/

Ziemke, T. (2016). The Body of Knowledge: On the Role of the Living Body in Grounding Embodied Cognition. *BioSystems, 148*, 4–11.