Gregory Alexander
Student: 719466

Supervisor: Professor Mark Leon

A Research Report submitted to the Faculty of Humanities at the University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Arts in Philosophy.

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Topic:

Artificial Intelligence, Machines and Understanding - A Critique of John Searle's Chinese Room Thought Experiment.

Word Count: 24211

Table of Contents

Plagiarism Declaration

University of the Witwatersrand, Johannesburg

School of Social Sciences

SENATE PLAGIARISM POLICY

Declaration by Students

I GREGORY ALEXANDER

(Student number: 719466) am a student registered for

Philosophy MA in the year 2019. I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that ALL the work submitted for assessment for the above course is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  Date: 17/07/2019

Acknowledgements

I would like to thank my supervisor, Professor Mark Leon, for his dedicated guidance, support and interest in my research paper. I would also like to thank my parents, Robert and Nathalie Alexander, for their unwavering support throughout my studies.

Abstract

Can machines think? In this paper, I will argue against John Searle's *Chinese Room Thought Experiment*. The *Chinese Room Thought Experiment* argues against the claim that "the appropriately programmed computer literally has cognitive states and that the programs thereby explain human cognition" (Searle 1980, 417). The central thesis that I intend to pursue is that it is/will be possible for machines to exhibit cognition. I will engage with various replies to Searle's argument, such as *The Systems Reply*, *The Virtual Mind Reply*, *The Robot Reply* and *The Brain Simulator Reply*, to argue for my thesis. I argue for an externalist account of "understanding" and I argue that if a machine were created that combines these responses then it would be able to understand the world around it. I then move on to argue for the theory of mind known as connectionism. I argue that machines should be programmed with a connectionist system as connectionist systems are able to learn from experience. In this paper, I intend to argue against Searle, put forward a concept of cognition and understanding, and argue for the theory of mind known as connectionism to conclude that a machine will be able to exhibit understanding of the world around it.

Overview

The aim of this paper is to analyse machines and discover under what conditions it is plausible to say that a machine exhibits "understanding". In addition, I want to be able to show what my view of an "understanding machine" would be. The central thesis that I intend to pursue is that it is/will be possible for machines to exhibit cognition. I intend to do this by critically analysing and arguing against John Searle's *Chinese Room* Thought Experiment. By refuting Searle's argument, I will be able to provide a discussion and argument about what a machine that exhibits cognition would be. To do this, I will provide the reader with an account of what 'understanding' is.

The intended research is important because machines that exhibit cognition are becoming more and more prevalent in our lives. Artificial intelligence is a large component of computer science and technological improvements are occurring every day, from new software in smart phones to self-driving cars. In the computer science discipline, a lot of the discussion centres around what our lives will be like when we live and work with these machines. From an engineering point of view there is also a lot of work being done in computer programming to create artificially intelligent systems. What I have noticed from my research, is that other disciplines only speak about how our lives will be affected and how these machines will be programmed. From this arises my biggest concern, "What is a machine that exhibits cognition and what does it mean for a machine to be able to understand"? From a philosophical perspective, researching artificial intelligence is extremely important because if our lives are to be affected by improving technology such as these "thinking machines", then we need to know what these machines are and what makes them so.

Research Paper Outline

Section I - Functionalism

My paper begins with a discussion of Daniel Dennett's Intentional Stance. Dennett's Intentional Stance is relevant because it helps to explain the ascription of intentional states (or mental states), such as beliefs, desires and hopes, to living and non-living systems. The purpose of Section I is to highlight how functionalism can account for mental states in a non-living system, such as a machine. Functionalism is the doctrine that "what makes something a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part" (Levin 2016). This is the common-sense version of functionalism. The common-sense version of functionalism provides us with a platform on which to start arguing how a machine could have mental states. After outlining the common-sense version of functionalism I go on to discuss experiential states, such as pain. This is relevant because explaining pain as a functional state, and showing how a machine could be in a state of pain, it becomes clear how a machine could also have mental states. To provide further evidence why I believe functionalism can help us account for mental states in a machine, I look closely at a version of functionalism known as psychofunctionalism. Psychofunctionalists hold that "the best empirical theories of behaviour take it to be the result of a complex set of mental states and processes, introduced and individuated in terms of the roles they play in producing the behaviour to be explained" (Levin 2016).

To consider objections to functionalism, and psychofunctionalism, I critically engage with the objection that functionalism fails to account for qualitative states (or "qualia") and Ned Block's "Homunculi-Headed Argument", as well as the spectrum inversion objection. These objections aim to show that systems functionally equivalent to us might not have mental states. Block's argument and the spectrum inversion objection show that functionalism does not account for the realization of mental states with any qualitative character. To respond to these objections, I hold that it would be difficult for Block to show how an entity functionally equivalent to you or me would lack mental states with any qualitative character. To disprove functionalism, the onus would be on him to provide solid proof for his claims. In response to the spectrum inversion objection, I concede the point for qualitative states but this does not necessarily mean that cognitive states are affected. I believe that in this section I will provide enough evidence to show how and why such objections should not be a problem for functionalist accounts of

cognitive states and the idea that machines can be ascribed mental states. In conclusion, the purpose of this section is to provide an account of intentional states and lay the groundwork for why I hold that machines could possess mental states. In Section II, I critically engage with Searle's *Chinese Room Thought Experiment*.

Section II - John Searle's *Chinese Room Thought Experiment*

This section begins with an analysis of Alan Turing and the *Turing Test*. The *Turing Test* is a game which tests a machine's ability to demonstrate intelligent behaviour. This is relevant because it provides us with the necessary groundwork on which to begin critically analysing Searle's *Chinese Room Argument* (CRA). Searle's CRA is an argument directed against *Strong AI*, which is the notion that a "computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states" (Searle 1980, 417). I aim to show that Searle's argument against Strong AI does not show that machines cannot have cognitive states. To argue for this, I will critically engage with the most well-known responses to the thought experiment. These responses are *The Systems Reply*, *The Virtual Minds Reply*, *The Robot Reply* and *The Brain Simulator Reply*. I systematically analyse each of these replies, discuss Searle's responses to each reply and provide further arguments for the replies in light of Searle's comments. Even if each reply on its own is not sufficient to refute Searle, I will show how they can be taken accumulatively to defuse the CRA. This accumulation of the replies is described when I discuss *The Combination Reply* objection. Once this reply is outlined I conclude this section of my paper. These responses to Searle's CRA defuse Searle's argument to a certain extent but further elements will need to be discussed to make the argument against Searle more persuasive. These further elements will be discussed in Section III and they are an account of understanding, a theory of mind to apply to a machine, machine learning, deep learning and real-world AI scenarios.

Section III – Machine Learning, Connectionism & Understanding

After critically analysing Searle's CRA in Section II, I move on to provide an account of understanding, to show how a machine could possibly understand and to add more weight to the argument against Searle. In this section, I critically engage with two accounts of understanding, namely "internalism" and "externalism". Internalists hold that understanding is

something that comes from within and so it is "a purely internal product of internal physiological processes" (Searle 1987, 230). While externalists hold that what is needed for understanding is a suitable relationship between the subject and the external world (Preston 2002, 40). I compare the two accounts and conclude that the externalist account of understanding is the more plausible account. The externalist account, which focuses on our environment-specific knowledge (Proudfoot 2002, 178), appears to be the more plausible account because the correct causal connections to an environment seem necessary for understanding to be ascribed to an entity.

The next part of this section is a comparison between two theories of mind, connectionism and the classical theory of mind. Connectionism is the idea that sets out to explain intellectual abilities using artificial neural networks (like how our brains function) as a simulation on a computer (Garson 2016), whereas the classical theory of mind (otherwise known as computationalism) is a theory that holds that psychological states are distinguished by their casual connections with sensory inputs and behavioural outputs (Preston 2002, 9). Here information can be represented by strings of symbols, comparable to how data is represented on a computer with "1"s and "0"s. I compare these two theories and argue that connectionism is a more plausible account, as it may be a more biologically plausible account than the computational account. As a more biologically plausible account, connectionism offers us a system that can learn from experience rather than a sytem that needs to be systematically programmed.

Next, I will discuss machine learning (ML) and deep learning (DL), as well as look at some real-world examples of AI machines. ML is the idea that machines should be given access to large amounts of data and then by sifting through the data they learn for themselves (Marr 2016). Deep Learning, or DL, is an even more specific branch of ML, because in DL data is still fed through the simulations of neural networks, however on a much more advanced and larger scale (Marr 2016). Mentioning ML and DL in my paper is relevant because they provide further evidence for the plausibility of a connectionist theory of mind and the use of neural nets. From here I will analyse an objection to connectionism known as "The Chinese Gym". "The Chinese Gym" is a modified version of Searle's original CRA, and it serves as an objection to connectionism. To respond to this new objection I refer back to the argument made by the *Systems Reply* as discussed in Section II. I aim to show that Searle's "Chinese Gym Argument" is not sufficient to contradict the plausibility of connectionism to conclude this final

section of my paper. All of the information contained in these three sections will aid me in my argument against John Searle's *Chinese Room Thought Experiment*.

Section I:

Functionalism

1.1) Introduction

In this section of my paper, I will argue that the doctrine of functionalism provides support for the notion that machines can think. Functionalism is the doctrine that "what makes something a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part" (Levin 2016). I will critically engage with the idea of functionalism by assessing it against the objection that functionalism fails to account for qualitative states (or "qualia") and Ned Block's "Homunculi-Headed Argument". In terms of my discussion, functionalism is an important place to start to provide a solid foundation on which to argue that it is/will be possible for machines to exhibit cognition. I intend to argue for the plausibility of machines that understand the world around them. In this section I analyse intentional states and the theory of functionalism to answer the question: "What are mental states?".

1.2) The Intentional Strategy/Stance

"The Intentional Stance" or "The Intentional Strategy" are both terms devised by Daniel Dennett to explain the ascription of intentional states, such as beliefs, desires and hopes to living and non-living systems. Dennett holds that an intentional system is a system whose behaviour can be predicated and explained, given its beliefs and desires (Dennett 1988, 495). In his paper, *True Believers: The Intentional Stance and Why It Works*, Dennett outlines two contrasting views on the nature of belief ascription. These two opposing views are *realism* and *interpretationism*. The "realist" view holds that a belief can be an objective phenomenon and therefore something that is occurring within a system. The "interpretationist" view, however, holds that whether a system has beliefs depends on how it is assessed (or how it is "interpreted"). The thesis that Dennett puts forward for belief ascription is that beliefs are perfectly objective phenomena (making him a realist), while they are also things that need to be assessed as a matter of interpretation based on a predictive strategy (making him an interpretationist too) (Dennett 1981, 15). It is important to note that I am looking at Dennett's work for the conditions of belief ascription. When it comes to belief ascription, the debate between the realists and the interpretationists is beyond the scope of my paper and I will not get involved in the debate.

When applying Dennett's intentional strategy or stance to an entity whose behaviour you want to predict, the entity must be treated as a rational agent with beliefs, desires and other mental states. This will show what is known as the "intentionality" of an entity. Dennett holds that any system whose behaviour can be well predicted by his intentional strategy can be considered a "true believer" (Ibid, 15). If a system is a true believer than it is an intentional system, which means that it can be ascribed beliefs. I will show how an inorganic entity, like a robot, can possibly be attributed beliefs and shown to have mental states.

How the intentional strategy works is, first one must choose an object whose behaviour is to be predicted and then treat this object as a rational agent. A rational agent is an agent that is expected to behave in a way to further its goals in the light of its beliefs. With regards to the rationality that one attributes to an intentional system, we must begin by assuming the system is perfectly rational and then reassess its rationality depending on the circumstances. Next, one decides what beliefs or desires this agent ought to have, based on its place in the world. From here you can anticipate that this rational agent will act in such a way to promote its desires, goals and interests, based on what it believes (Ibid, 17).

As an example, think about the way in which we ascribe beliefs to one another. It seems right to say that secluded people tend to be more ignorant than outgoing, interested people. If you expose someone to something new, they generally come to know more about what they are being exposed to (Ibid, 18). One can say that we begin to believe the truths of things around us when we are placed in a situation or environment to learn about those things. Put more clearly, Dennett holds that exposure to *x* over a suitable extent of time is usually a justifiable condition for having a belief about *x* (Ibid, 18). We can say that we hold beliefs about the things around us. From this, Dennett states: "one rule for attributing beliefs in the intentional strategy is this: attribute as beliefs all the truths relevant to the system's interests (or desires) that the system's experience to date has made available" (Ibid, 18).

Dennett holds that we use this strategy all the time when predicting the behaviour of other people or even non-living (inorganic) entities. With regards to non-living entities, let us use a well-known sci-fi movie character as an example for the intentional strategy. Let us take *R2-*

D2[1], the lively robot from the *Star Wars* franchise as our example. One could apply the intentional strategy to a robot such as *R2-D2* to say that whenever it performs an action, it does this because it believes the action to be the best way to further its goals (Ibid, 22). For example, if *R2-D2* and the other characters in the film were in a life-threatening situation and *R2-D2* acted in such a way to save everyone, we could say that *R2-D2* acted in this manner because it believed that this was the best way to further its goals. It is clear to see how an inorganic entity like, *R2-D2*, holds beliefs.

Up until now I have spoken about attributing beliefs to systems very generally as I have moved from speaking about the beliefs of living entities (like humans, who seem to have beliefs) to robots like *R2-D2*. Right now, the most important thing to note is how this method of belief ascription works and that we have an account to describe what it means for something to have a belief. Dennett does this to stress the importance of his logical account of belief ascription. The intentional strategy claims, that what it means to be a true believer, is to be an entity whose behaviour can be reliably anticipated based on what it is believed that the entity will do. All that matters for believing that *x* (for any proposition *x*) is to be an intentional system for which *x* transpires as a belief in the most suitable (and most predictive) interpretational explanation (Ibid, 29). From this logical explanation, it can be seen how machines, like robots, could possibly be attributed mentality.

1.3) Common-sense Functionalism

Functionalism holds that a mental state is determined "by its causal relations to sensory stimulations, other mental states, and behaviour" (Levin 2016). What this means is that an entity is not judged to have or not have mental states based on its intrinsic composition. Instead, the ascription of mental states is given to an entity based on the way that its internal elements function. To have beliefs requires possessing intentional states that interact with each other in appropriate ways as well as with inputs and outputs.

[1] *R2-D2* is a blue and white robot character from the famous sci-fi movie franchise *Star Wars*. *R2-D2* is the robot companion of the main characters throughout the movie franchise and the robot often acts in ways to save them all when they are faced with life-threatening situations. Based on the way *R2-D2* behaves to further its goals and the goals of others, it would be hard to argue that it should not be ascribed beliefs.

Functionalism is propelled forward by the idea that mentality is a matter of functioning instead of substance (Preston 2002, 8). So it does not matter if a creature (or machine) is made out of carbon compounds, biological stuff or even physical stuff. What matters with regards to mentality, is how the thing in question works and what its capacities are. The functionalist thesis that I will pursue is: "what's essential to mental phenomena is their causal roles, and if those roles can be simulated in computer programs, there's no reason why computers running those programs shouldn't be credited with the mental phenomena in question" (Ibid, 9). This opens the door for investigation into the mental capabilities of machines now and in the future. Functionalism depicts mental states "in terms of their roles in the production of behaviour" (Levin 2016).

It is important to note that functionalism must not be mistaken for behaviourism. The doctrine of behaviourism claims that the mentality of humans, as well as other animals, can be explicated by only appealing to the behavioural tendencies of these creatures. So, this means that the assessment of human cognitive abilities was based solely on the way that they acted within specific environments (Ibid, 2016).

What functionalism holds, is that there could be entities (both biological and non-biological) that are functionally identical to us but do not have the same neural properties as we do. If the truth of such a claim could be argued for and if such beings can reasonably be considered to share our mental states, "then even if neural states can be individuated more coarsely, functionalism will retain its claim to greater universality" than other theories (Ibid, 2016). Functionalism is such an important theory to assess, as it universally incorporates non-human entities. Functionalism does not discriminate against entities that are not made of the same organic materials that humans are made up of. As technology improves, machines will become more "intelligent" and something will need to be said about the ways in which machines exhibit cognition. This is of the utmost importance because it is a starting point for machine intelligence. Machines that function like us, regardless of their internal materials need to be given a proper assessment. Functionalism is useful in my discussion because it helps to answer the question: "What are mental states?" (Block 1978, 268).

1.4) Experiential States

I will now describe the experiential state of pain from a functionalist perspective. Functionalism regards pain as a mental state that is caused by bodily injury to create the belief that there is something not right with the body and to create the desire to be out of that particular state. All and only creatures with inner states that satisfy these conditions, or play these roles, possess the capacity to be in pain (Levin 2016). Functionalists claim, that what makes something a pain state is not its specific neural makeup. For example, people and animals can both be in pain even if they have very different neural configurations (Schwitzgebel 2015). Additionally, an alien that would very likely have a very different neural makeup to people or animals, could also be in a state of pain. What is of the utmost concern here, is that the subject in question is in a state that is suitable to be caused by damage or stress. This damage or stress is then suitable to cause signs of distress, withdrawal or future avoidance of the painful stimulus (Ibid, 2015).

Any creature with inner states that produce a pain state can be in pain. Imagine that within human beings there is a specific type of neural activity that occurs when a human is in pain. Let us call this neural activity "C-fiber stimulation". A functionalist would argue that humans are in pain when they endure C-fiber stimulation. However, this does not mean that only creatures who undergo C-fiber stimulation can be in pain. Functionalism opens up the possibility that many other creatures from Aliens to AI machines can be in the mental states of pain if their internal materials react similarly to bodily injuries to create the belief that there is something not right with the alien or machine. These creatures do not have to have the same C-fiber stimulation to be in pain but rather can be in pain if they have their own internal conditions that satisfy the same criteria of being in a state of pain as that of pain in a human. Functionalists argue that "pain can be *realized* by different types of physical states in different kinds of creatures, or *multiply realized*" (Levin 2016).

Multiple realizability is the idea that a single mental kind (property, state or event) can be realized by many different physical kinds. Which means that physical kinds, such as organic C-fibre stimulations to perhaps circuitry functions in machines can realize the state of pain for the entity. The multiple realizability thesis has been closely linked to functionalism[2]. Examples

[2] It must be noted that the multiple realizability thesis is not only associated with functionalism.

that explain that, a psychological kind (such as pain) can be realized by various physical kinds, are like "brain states in the case of earthly mammals, electronic states in the case of properly programmed digital computers, green slime states in the case of extraterrestials, and so on" (Bickle 2016).

The main idea from the example of pain is that if an alien has a functional process of green slime states, for example, that flow through its body when it reacts to certain stimuli and this flow of green slime is the realization of pain, then it cannot be argued that the alien is not in a state of pain because of its different makeup. The alien is in a state of pain if it's bodily functions produce the same conditions to be in pain (Ibid, 2016). On the same note, if a machine's internal electronic states react when it responds to external stimuli and cause the machine to be in "pain", then surely the machine will be in a state of pain. This can be applied to mental states too. If a machine, regardless of it's internal constitution, exhibits the correct causal roles, then it can be ascribed mental states, such as beliefs.

1.5) Psychofunctionalism

Psychofunctionalists hold that "the best empirical theories of behaviour take it to be the result of a complex set of mental states and processes, introduced and individuated in terms of the roles they play in producing the behaviour to be explained" (Levin 2016). This is in contrast to the common-sense theory of functionalism which analyses *a priori* the idea of our mental states in functional terms. Psychofunctionalism, rather comes from the consideration of the goals and methodology of "cognitive" psychological theories. Under this theory, for example, physically different entities can all be regarded as "eyes" as long as they allow an organism to see. On the same token, contrasting physical structures or processes can all be phenomena like beliefs, thoughts, sensations or desires, so long as "they play the roles described by the relevant cognitive theory" (Ibid, 2016). Psycho-functionalists take on the methodology of cognitive psychology in their characterization of mental states and processes "as entities defined by their role in a cognitive psychological theory" (Ibid, 2016).

Under the theory of psychofunctionalism, it is held that it is not possible for a system to have beliefs or desires, unless psychological theories that are true of us, are true of it (Block 1978, 291). This means that if a system is psychofunctionally equivalent to us and if we have beliefs

and desires, then it would also have beliefs and desires. Hence, psychofunctional equivalence to a human being is what is required for the ascription of mental states.

Block mentions a problem for psychofunctionalism, that even if psychofunctional equivalence to us is a requirement for our acceptance of mentality, it is not clear that it would be a requirement for mentality itself and from this, he asks the question: "Could there not be a wide variety of possible psychological processes that can underlie mentality, of which we instantiate only one type?" (Ibid, 291). He holds that the psychofunctionalist account could be chauvinist in that it would disregard the possibility of other creatures having mental states because of a lack of psychofunctional equivalence to us. He provides us with the example of Martians (Ibid, 291). If Martians landed on earth and we began to interact with them, share our lives with them and learn from them to the point where it became clear they had mental states (because they are like us), we would not disallow them the ascription of mental states if suddenly we found out that they were not psychofunctionally equivalent to us. In this regard, Block holds that psychofunctionalism is too chauvinist a theory (Ibid, 292).

Block holds that the problem with functionalism is that it errs on the side of being too liberal a theory and the problem with psycho-functionalism is that it errs on the side of being too chauvinist a theory. It is important to note that a theory is liberal insofar as it wrongly *ascribes* mental states to a system and a theory is chauvinist to the extent that it mistakenly *denies* that a system has mental states (Ibid, 292).

To respond to Block's "too chauvinist a theory" objection we can look at computationalism as a version of psychofunctionalism. Under the theory of computationalism, psychological states are distinguished by their casual connections with sensory inputs and behavioural outputs (Preston 2002, 9). This is the idea that our minds compute, which means that thinking is the processing of information realized by manipulating symbols. If this theory is a way to describe our psychological states, then the chauvinism objection will be avoided. Psychofunctionalism does not have to be seen as a chauvinist theory because it can be universally applied to creatures whose brains might not be made up like ours, but whose brains have the same psychological functioning.

Computationalism is associated with "The Language of Thought Hypothesis" (LOTH) which holds that thought and thinking occur in a mental language. This is an empirical thesis about

the nature of thought and thinking (Aydede 2015). What it means for thought and thinking to occur in a mental language, is that they are physically realized in a symbolic system in the brain of the correct type of organism. For examples of the LOTH, we must look at "propositional attitudes". These are thoughts described by sentences of the type, "*S* believes that *P*", "*S* hopes that *P*" or "*S* desires that *P*" (Ibid, 2015). In these examples, *S* refers to the subject expressing the attitude and *P* refers to that which is believed. So, *S* could be the man, the woman or the machine. While *P* could be a sentence like "it will rain tomorrow". The relevance of the LOTH is that it is based on the theory of computationalism. The idea here, under the LOTH, is that thinking entails the rule-governed manipulation of uninterpreted symbols. These uninterpreted symbols will then have a certain sort of physical realisation that can differ from system to system (Ibid, 2015).

Functionalists hold that to believe is to be in a state that plays a certain sort of causal role (Schwitzgebel 2015). The following four points can be seen as rough ways to functionally characterise beliefs and further describe the LOTH. Firstly, if we reflect on propositions *P* and *if P then Q* from which *Q* directly follows and if we believe these propositions, then proposition *P* typically causes the belief that *Q* (Ibid, 2015). Secondly, if we direct our perceptual awareness to the properties of things, under conditions that make our perceptions accurate, it can be said that we believe things to have those properties. For example, looking at a red book in sufficient viewing conditions will normally cause the belief that the book is red (Ibid, 2015). Thirdly, holding the belief that doing action *P* will lead to event or state of affairs *Q*, in addition to a desire for *Q* and no other opposite desire, will typically cause the intention to carry out action *P* (Ibid, 2015). Lastly, believing that *P*, under conditions that favour the genuine expression of belief *P*, will normally lead to an assertion of *P* (Ibid, 2015). Therefore, for functionalists, to believe is to be in a state that plays something like these sorts of causal roles (Ibid, 2015).

Representationalism ties in with the LOTH and can help us to better understand questions such as "What is it to believe?". Representationalism is the idea that to believe something, is to have a fact or proposition stored or represented in one's mind (Schwitzgebel 2015). For example, when an individual learns something particular like, astronomers do not classify Pluto as a planet, the individual acquires a new belief, which is the belief that Pluto is not a planet. Holding this belief can then play a causal role in the production of behaviour. For example, if the individual discusses astronomy with a friend and the friend says "Pluto is one of the nine

planets in our solar system", this would call up the individual's belief about Pluto, stored in his mind. The individual's belief about Pluto would play a causal role in the production of his behaviour because he would likely respond and contradict his friend's statement (Ibid, 2015). Both representationalism and computationalism help us avoid Block's worry about chauvinism because they highlight how any system could have mental states if the system satisfies the criteria of the rule-governed manipulation of uninterpreted symbols.

1.6) Objections to Functionalism

1.6)1. The Homunculi-Headed Argument

Objectors to functionalism argue that functionalism fails to take qualitative states into account. Qualitative states, or the more commonly used term "qualia", refer to mental states with "distinctive subjective character" (Tye 2017). For example, if I smell freshly baked bread, get a splinter in my finger, see the colour red or enjoy the taste of coffee, there is something that it is like for me to be in each of these states. There is a distinctive subjective characteristic of what it is like for me to smell freshly baked bread, get a splinter in my finger, see the colour red or enjoy the taste of coffee. Each of these states has a specific sort of phenomenology. The term qualia refers to "the introspectively accessible, phenomenal aspects of our mental lives" (Ibid, 2017).

As an objection to functionalism, Ned Block outlines an argument known as the homunculi-headed argument. Block argues that his argument unsettles the functionalist doctrine because it shows that functionalism is guilty of liberalism. Meaning that functionalism classifies systems that do not have mentality as having mentality (Block 1978, 275).

For Block's argument, he asks us to imagine a body that is externally like a human body, however it is internally quite different. Instead of the usual internal makeup that we are familiar with in our own bodies, this body has neurons from sensory organs that are attached to a bank of lights in a hollow chamber in the head. Along with this, there is a set of buttons that connects to the motor-output neurons. Now, inside the chamber there is a small group of little men and each man has a simple task to carry out. Each little man must execute a "square" of a reasonably adequate machine table that describes you. On one wall of the chamber there is a bulletin board with a posted state card and this card displays a symbol designating one of the various states

identified in the machine table. Imagine that the posted card on the wall has a 'G' on it. All the little men who initiate 'G' squares (they are called 'G-men") will then be alerted. Next imagine that a light standing for input 'I' suddenly goes on. We now move onto what the little men do. When this happens, one of the little 'G-men' is required to press output button 'O' and change the state card to 'M'. This is one of the 'G-man's' only tasks. Block ends the argument by stating that "in spite of the low level of intelligence required of each little man, the system as a whole manages to simulate you because the functional organization they have been trained to realize is yours" (Ibid, 276).

With this argument, Block aims to show that a "homunculi-headed" system does not have mental states with any qualitative character. This means that it is possible for there to be states functionally similar to sensations or perceptions that do not have characteristic feels (Levin 2016). If we have mental states but something functionally equivalent to us does not have mental states, then Block's argument shows that functionalism does not account for the realisation of mental states.

The whole purpose of Block's homunculi-headed creature objection to functionalism is to show that specific mechanisms would weaken rather than support mental functioning. The type of creature that he has described, which is functionally equivalent to you, would lack mentality because of its structure and there would not be anything that it was like to be that entity (Leon 1998, 384).

To go against what Block argues, it seems plausible to accept that where there is functional equivalence, there is the possibility of attributing mental states to a system. We should not be drawn on the idea that a difference in matter or mechanism is the reason for ruling out mentality in a system (Ibid, 384). This is because the mental should not be viewed as being akin to natural kinds. What natural kinds are, are items with general underlying structures. We should see beliefs and mental states as informational states created under certain conditions, that cause other states and other specific behaviours, which in turn are also produced under specified conditions. For a state or a process to be regarded as a mental state, or for a system to be attributed mentality, we must not focus on its structure, but rather it must fulfil our descriptions of an intentional system (Ibid, 385).

1.6)2. Spectrum Inversion, Liberalism & Chauvinism

Another objection to functionalism raised by Block is the inverted spectrum argument. Imagine a lens that inverts colours. This lens can be placed on someone's eye and it would change how colours appear to the wearer. For example, all things that the wearer had originally viewed as green would now appear as red to them, and vice-versa (Block 1978, 288). Now to take this idea further, imagine a pair of identical twins and one of them has a lens of this sort inserted over his or her eyes at birth. Throughout the rest of their lives the twins grow up together normally and they are functionally equivalent, even though they have different experiential states when viewing green and red objects. One twin sees green things as red and the other twin sees green things for what they are, green. This objection to functionalism holds that there is the possibility of people who can be in states that fulfil the functional definition of our own experience of red but who experience green. This goes against functionalism because it shows that functional equivalence does not necessitate the same experiential states. So, a machine (for example) that is functionally equivalent to us is not guaranteed to have the same experiential states as we would have (Levin 2016).

1.7) Rebuttals in Defence of Functionalism

As a response to Block's homunculi-headed argument, it does not seem clear that qualia or subjectivity would necessarily be absent in this case. This is because we do not know enough about the relation between our own "hardware" and subjective mental states, to rightly say that the entity functionally equivalent to us lacks mentality because of its own internal structure. It would be difficult for Block to prove that the functionally equivalent being does in fact lack mental states. The onus would be on him to prove this lack of mentality.

To respond to the spectrum inversion objection, I will concede the point for qualia. In this case, it seems plausible that two functionally equivalent beings can have varying experiential states. However, I hold that conceding to the qualia objection does not necessarily mean that this is also true of all mental states. It does not entail that this objection affects cognition.

1.8) Conclusion

In this section of my paper, I showed that the doctrine of functionalism can provide support for the idea that machines can think because of how the theory classifies mental states. I analysed intentional states and the theory of functionalism to answer the question: "What are mental states?". After having critically engaged with the idea of functionalism by assessing it against the objection that functionalism fails to account for qualitative states (or "qualia") and Ned Block's "Homunculi-Headed Argument", I will now move on to critically engage with John Searle's Chinese Room Thought-Experiment in the next section of my paper.

Section II:

John Searle's *Chinese Room Thought Experiment*

2.1) Introduction

In this section of my paper, I will critically discuss John Searle's *Chinese Room Thought Experiment*. John Searle's thought experiment needs to be critically analysed because the idea of Strong AI can show that a machine could have mental states. I will point out that Searle's argument against Strong AI does not show that machines cannot have mental states. Strong AI is the idea that a "computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states" (Searle 1980, 417). To argue for my thesis, I will critically engage with the most well-known responses to the thought experiment. These responses are *The Systems Reply*, *The Virtual Minds Reply*, *The Robot Reply* and *The Brain Simulator Reply*. In addition to this, I will also raise the objection known as *The Combination Reply*. Searle responds to these replies; however, I will argue that his responses are not sufficient to undermine these objections to his *Chinese Room Argument* (CRA). In this section of my paper, I will show how Strong AI can account for mental states.

What "Strong AI" means, is that any computer that has been programmed in a specific way is said to understand what it is doing when it is running its program. Searle believes that his thought experiment shows that computers just use syntactic rules to operate strings of symbols, while they have no understanding of meaning or semantics behind the symbols (Cole 2015). The CRA is as follows: there is a man in a room and he is surrounded by batches of Chinese symbols, along with a rule book that correlates the symbols. The room has two slots, and pieces of paper can be passed through these two slots. Questions on paper are passed through one of the slots (the input slot) and the man in the room passes responses through the other slot (the output slot). The papers passed through the input slot are Chinese questions and the man then has to respond to these questions by using the rulebook and the batches of symbols. Once he has responded, he passes the responses through the output slot.

2.2) Alan Turing

Alan Turing, heralded as the father of computer science, devised a behavioural test for computers called *The Turing Test*. The Turing Test is a game which tests a machine's ability to demonstrate intelligent behaviour. In his paper, "Computing Machinery and Intelligence", he put forward this test, as a means of dealing with the question whether machines can think.

The question, "Can machines think?" originally spurred my interest in this topic and motivated me to write this paper. Turing poses the question of whether a digital computer can do well in a specific game that Turing describes as "The Imitation Game" (Oppy and Dowe 2018). Turing believed that any machine that passes his imitation game, and more importantly the modern version known as *The Turing Test*, should be regarded as intelligent. This test was put forward as a simple way to describe intelligence and it is argued that with this operational definition of thinking, Turing propelled the whole field of artificial intelligence forward (French 2000, 115).

The *Turing Test* is a game played between a person, a machine and an interrogator. These three are separated in different rooms. The objective of the game is for the interrogator to distinguish which responder is the person and which responder is the machine after firing questions at and receiving responses from the two. After the game the interrogator must say that either "X" is the person and "Y" is the machine or vice versa. An example of a question that the interrogator can pose is a question such as "Will X please tell me whether X plays chess?". Whichever of the two is X in this example, the person or the machine, must respond to the question addressed to X. The aim for the machine is to persuade the interrogator to conclude that the machine is the other person and the aim for the person is to aid the interrogator in accurately determining which of the two is the machine (Oppy and Dowe 2018). Turing's central claim is that there would be no grounds to deny intelligence to a machine that could accurately mimic a human's unrestricted conversation (French 2000, 116). The Turing Test, in more modern terms, could be run on an Instant Messaging service (such as Facebook Messenger or WhatsApp). Each person could be given a device and the machine would have the application set up too. The two people and the machine could be involved in a group chat and the interrogator would have to decide which of the two is the person and which is the machine. If the interrogator concludes that the machine has perfectly imitated a person and the interrogator cannot distinguish between the two, then the machine, according to the view of Alan Turing, is intelligent.

The relevance of Alan Turing's *Turing Test* to John Searle's *Chinese Room Thought Experiment* is that Searle, through his CRA, is disagreeing with Turing's notion that an appropriately programmed computer could think and show intelligence (Oppy and Dowe 2018).

In his book, *The Rediscovery of the Mind* (1992), John Searle refers to Strong AI as being an implausible theory. He suggests that Strong AI is the widely-held view that a computer can

have "thoughts, feelings, and understanding solely in virtue of implementing an appropriate computer program with the appropriate inputs and outputs" (Searle, 1992: 7).

2.3) John Searle's Chinese Room Argument (CRA)

In John Searle's 1980s paper titled, "Minds, Brains and Programs", he outlines the *Chinese Room Argument* (CRA). As well as discussing the argument, he evaluates various responses to the argument. Searle sets up this thought experiment to refute the claim that "the appropriately programmed computer literally has cognitive states and that the programs thereby explain human cognition" (Searle 1980, 417). What this claim means is that any computer that has been programmed in a specific way is said to understand what it is doing when it is running its program. Searle believes that his thought experiment shows that computers just use syntactic rules to operate on strings of symbols, and have no understanding of the meaning or semantics behind the symbols (Cole 2015). For brevity, I will refer to this thought experiment as the "CRA" (Chinese Room Argument) throughout my paper. Computers merely take inputs and produce certain outputs without any understanding, according to Searle.

For his CRA: Searle asks us to imagine that a person is in a room among many batches of Chinese symbols. This person in the room has no knowledge of either written or spoken Chinese. Along with the batches of Chinese writing, the person in the room is also given a second batch of Chinese writing along with a rulebook that correlates the second batch of writing with the first batch. These rules in the book allow the person in the room to be able to correlate one set of symbols with another based purely on their shapes. Next the person in the room is given some more rules in addition to a third batch of Chinese writing. There are now rules that correlate the third batch with the first two batches. These new rules indicate which symbols to hand out of the room in response to symbols being passed into the room (Searle 1980, 418). The person in the room does not know that the people outside of the room handing the batches into the room are calling the first batch symbols "a script", the second batch symbols a "story", and the third batch symbols "questions". The symbols that the person hands out of the room in response to the "questions", are called the "answers to the questions". Lastly, the people outside of the room call the rules "the program". Searle believes that if the rules for manipulating the Chinese symbols are followed correctly, then the people outside of the room would have great difficulty differientiating the answers from the room with say, the answers that a native Chinese speaker might give. It would be extremely difficult for someone outside

of the room to deny that, judging by his or her answers, the person inside of the room passing out symbols is not fluent in Chinese. In the case of the answers, the person in the room has merely handed out answers by following the rulebook and manipulating the batches of Chinese symbols (Ibid, 418). Searle also claims that the person in the room has just acted like a computer because they have performed "computational operations on formally specified elements" (Ibid, 418).

Searle shows that according to the theory of computationalism, thinking is the processing of information and that this processing of information is merely a case of manipulated symbols. He also maintains that computers do symbol manipulation. From all this, Searle believes that the most effective way for us to study thinking, or rather, "cognition", is by analyzing computational symbol-manipulating programs (Searle 1984, 43). Computation requires the rule-governed manipulation of uninterpreted symbols. Searle believes that what goes on in his Chinese Room thought-experiment is what goes on in the operation of a computer. Hence, Searle holds that if what goes on in his Chinese room does not amount to understanding, then the operation of a computer program does not add up to understanding either.

After explaining his thought experiment Searle highlights two points from this scenario which he believes help him refute Strong AI. These two claims are: firstly, the person in the room does not understand a word of Chinese, and secondly, "we can see that the computer and its program do not provide sufficient conditions of understanding since the computer and the program are functioning, and there is no understanding" (Searle 1980, 418). I intend to critically engage with this second premise of Searle's argument by appealing to four common replies to his argument as well as an accumulation account of these replies to the argument. I agree with the first point that Searle makes because I do not think that the person in the room understands Chinese. Of the responses that I engage with, most of them agree with this point too. The important point to press Searle on is his second claim. Searle is wrong to say that the computer and its program do not provide sufficient conditions of understanding (Ibid, 418).

It is important to note that Searle does not argue that no machine can think (Boden 1987, 1). His idea is that humans can think and Searle believes that humans are machines. The focus of his argument is that nothing can possess the capacity to think, mean what it says, or understand solely in virtue of its instantiating a computer program. Searle believes that it is not possible for meaning and intentionality to be described in computational terms (Ibid, 1). The important

point that Searle is making is that a system does not think in virtue of "instantiating a programme". Searle argues that if there is one theme that runs throughout all of his work, then it is that consciousness is a natural, biological phenomenon and this phenomenon is as much a part of biological life as digestion, growth and photosynthesis (Searle 1997, xiii).

With regards to understanding, Searle believes that a proper explanation of understanding would be that meaningful symbols have to be embodied in something containing "the right causal powers". These "right causal powers" are what create understanding, or intentionality and he believes that a human brain can do this, whereas a computer does not have the required causal powers. Searle also clarifies that because a brain can be simulated in a computer, neuroprotein and the organic make-up of our brains are the biochemical properties that are crucial for understanding (Boden 1987, 4). As machines do not possess the chemical make-up that humans possess, machines cannot understand.

There is a problem with Searle's reasoning here because using the description of understanding requiring the "right causal powers" is very vague. Searle's point is merely that humans have the right causal powers and machines do not. I worry that Searle's point is not decisive because it does not tell us what intentionality is, what it involves or even how it is realised by the brain's causal powers. I will elaborate on this because I want to show that machines, set up in the correct manner, could possess the "right casual powers" required for understanding.

It is not clear what Searle requires these causal powers to be or how they should manifest in a human and how they do not manifest in a machine that mimics a human. Metal and silicon, in Searle's opinion, are not sufficient materials for understanding and are not sufficient for "the right causal powers". Disregarding the possibility of different kinds of materials supporting mental functions is problematic. It is problematic because it is not entirely clear how the organic substances that make up our bodies and brains support mental functions (and more specifically intentional states). So, if we do not fully understand how organic materials lead to mental states then we cannot merely brush aside inorganic materials on the grounds that they lack the relevant causal powers.

Margaret Boden helps substantiate the claim that we cannot intuitively deny intentionality to entities made of certain materials like metal and silicon. Boden does mention that perhaps it is true and only creatures with neuroprotein in their make-up and creatures that have a "terrestial"

biology can account for intentionality. However, she holds that there is no scientific evidence to back this claim and thus we cannot base our assumptions on intuition (Ibid, 8).

As a brief example of inorganic substances working like organic substances, we can turn to "computer vision". Here, metal and silicon are able to support some of the functions required for the 2D- to 3D- mapping associated in vision. This is the process of a machine using a digital camera to capture an image of an object. The image can be captured in 3D and the machine then has accurate dimensional data of the object. This sounds like the type of thing that a robot could do, as it would have digital cameras for eyes. If this can be done with metal and silicon then it should be possible to create "computer thinking" with these substances too. The main point that I want to stress here is that I agree that organic substances have "the right causal powers" for understanding to manifest (whatever those causal powers might be), but I do not agree that **only** organic substances possess this ability.

As with what has already been said about the intentional stance and the ascription of mental states to various systems in Section I, mentality could possibly be realised in different materials. Searle has not shown us that inorganic materials do not have the right causal powers and that they cannot support mental states. The onus would be on Searle to clearly show that inorganic materials cannot support mental states, instead of merely stipulating this detail away. I will now turn to an examination of behavioural and functional equivalence to show Searle's misunderstanding of Strong AI.

2.4) Behavioural & Functional Equivalence

There are two crucial points that need to be made at the beginning of this section. Firstly, it must be noted that Searle's CRA scenario is somewhat simplistic. A real test of his claim would require that his scenario be strengthened in important respects. The purpose of this section is to show how and why Searle's scenario is simplistic. Secondly, a more realistic scenario would involve the system being programmed to take indefinitely many inputs and be able to generate sufficiently many appropriate outputs. Rather than Searle's version, where, as Rey suggests is not equivalent to a Strong AI system because the rules are set up to ensure appropriate responses to a specific and finite set of inputs (Rey 1986, 171).

Searle misunderstands Strong AI because he confuses behavioural equivalence with functional equivalence (Ibid, 170). Searle believes Alan Turing's *Turing Machine* to be equivalent to Strong AI. However, the Turing Machine scenario is much simpler than a Strong AI system because under Alan Turing's view "a sufficient condition for machine understanding is teletype input/output equivalence to a normal human being" (Ibid, 170). The Turing Test focuses more on behavioural equivalence. Behaviourism is the scientific study of what individual organisms do and it is a way of assessing psychological state attribution, purely based on how organisms act (Graham 2017). Whereas Strong AI is more complex than this and it can be looked at from a more functionalist point of view. To review, functionalism is the doctrine that "what makes something a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part" (Levin, 2016).

Rey believes that Strong AI is a version of what are known as the "functionalist theories of mind". Meaning that Strong AI distinguishes itself from simple behavioural theories because it does not only take interest in physical inputs and outputs but it also relies on a system's *internal* states in the attribution of mental states (Rey 1986, 170). Regardless of how much a system's behaviour corresponds to another system that has mental states, the most important thing is that the system in question's inputs and outputs are brought about by the correct sorts of internal states. Rey uses a term, dubbed "AI-functional equivalence" for referring to a system that is mediated by the correct sorts of internal states and behaves like a system that has mental states. He holds that this condition is necessary for mental states. He sees this as being the important point for Strong AI and holds that the mere *teletype* behavioural equivalence with the Turing Test is irrelevant for Strong AI. To conclude this point, Rey states that "the fact that the Chinese room and/or its occupant might put out Chinese symbols to Chinese symbol inputs in a way indistinguishable from the behaviour of a normal Chinese speaker is entirely irrelevant to Strong AI" (Ibid, 171). So, Searle's view of a Strong AI system is as equivalent to a *Turing Machine*, that has a set number of responses to a particular set of outputs. However, on Rey's view, a Strong AI system would be more appropriately able to handle indefinitely many inputs and generate sufficiently many appropriate outputs.

Rey believes that the more plausible question to ask for Strong AI is: "is what is happening *inside* the room *AI-functionally equivalent* to what is happening inside a normal Chinese

speaker?" (Ibid, 171). From this, Rey holds that what matters for Strong AI is "AI-Functionalism". AI-Functionalism is defined as more than just behavioural similarities. It is defined as the view where a system does not merely behave like a system that has mental states but additionally the AI system follows the "same program" as a system that has mental states.

The key features of equivalence in terms of internal operations and processes discussed above can be helped further by looking back to what was said about beliefs in Section I. Two systems have equivalent internal operations and processes when they have psychofunctional equivalence. This can be explained by referring to computationalism as a version of psychofunctionalism. Two entities will be equivalent in terms of internal operations and processes if their causal connections with sensory inputs and behavioural outputs are the same. (Preston 2002).

To finish the discussion of the CRA, another general problem with Searle's Chinese Room is that it is extremely difficult (verging on impossible) for these so called "rules" that Searle has in the room to cover almost all inputs from the vast number of topics and questions the people outside of the room can put into the room. The rule book will have to cover an extensive array of inputs from a large variety of other programs that will need to be submitted to consider for things like "perception, belief fixation, problem solving, preference ordering and decision making" (Rey 1986, 171). A fluent Chinese speaker outside of the room will surely be asking questions more complex than, for example, "Do you like sunshine?" or "Does it snow in winter?" and therefore it will be very difficult to cover all possible inputs.

2.5) The Replies to Searle's Chinese Room Argument

In this section of my paper, I will mainly focus on the most prominent and widely known responses to the CRA. These responses are: *The Systems Reply, The Robot Reply, The Brain Simulator Reply and the Virtual Mind Reply.* I will be defending these responses and will argue that some responses can be combined to provide more weight to their conclusions. Searle responds to each of these replies. I will go into greater detail on each of these responses to show that they carry more weight than Searle originally believes. My view is that, as a

combination response, the four replies can provide a solid argument against Searle's CRA. I will systematically explain each of the responses to build an argument against Searle. Perhaps each reply on its own is not sufficient to refute Searle. However, with each reply addressing different failings and taken all together, they could be sufficient to refute Searle's argument.

2.6) The Systems Reply

2.6)1. The Systems Reply Outline

Proponents of The Systems Reply agree that the individual in the room does not understand Chinese as these proponents hold that the individual in the room is merely a part of a larger system. The fact that the man does not understand doesn't matter because understanding needs to be ascribed to the system, as a whole, of which he is merely the central-processing unit (CPU). Hence, "understanding is not being ascribed to the mere individual: rather it is being ascribed to this whole system of which he is a part" (Searle 1980, 419). The man (or CPU) is only a small part of the larger system. The system (which is the entire room) is made up of the rulebook, the Chinese symbols and the man. Similarly, a computer is comprised of various parts that make up its system, such as a huge database, the memory cards comprising of intermediate states, as well as the instructions. These elements from a computer system can be compared to the elements in the room, like the rulebook and the batches of symbols. This whole system is what is required for answering the questions that a fluent Chinese speaker would believe came from another fluent Chinese speaker. Understanding could be accredited to the whole system of which the person is just a fundamental part (Cole 2015). So, the Systems Reply holds that while the man in the room might not understand Chinese, the whole system could possibly understand Chinese. The fact that the man does not understand is irrelevant because he is the CPU in the system. What is relevant is that the whole system can be shown to understand (Ibid, 2015).

2.6)2. Searle's Response

Searle's objection to the Systems Reply rests on the idea that the man in the room should internalize all the elements of the system, to then become the whole system. To internalize the

entire system, he must memorize the rules in the rule book and the large batches of Chinese symbols, as well as respond to the calculations, in his head. Once this is done, the man becomes the entire system (Searle 1980, 419)**.** Searle goes on to point out that if the man were to leave the room and converse with a native Chinese speaker he would still have no method of attaching any meaning to the formal symbols (Cole 2015). Searle's point here is that the concept of the entire system understanding is not plausible.

2.6)3. Response to Searle

In response to Searle's objection, Georges Rey questions Searle on this idea of the man internalizing the whole system. To Rey, it does not seem possible that the man can internalize the system and still not understand. Rey makes this point explicitly clear, by saying that in this instance, we have our person made of flesh and blood, with a fully biological brain (which Searle requires). This person is not just following rules, for he is also *reading* an external manual that he has actually *internalized* through memorization. These memorized rules are not just the correlation rules that Searle has already mentioned but instead these rules are full recursive grammatical and semantic rules (Rey 1986, 174) This relates back to what was said about functional equivalence as, here, we have the correct sorts of internal operations and processes occurring required for the ascription of understanding. Rules of this nature would be the type of rules that linguists and philosophers of language would require for understanding of a sort. As Georges Rey's line of reasoning relies on this point because it seems plausible that if one were to memorize a book of rules, then one would be internalizing them based on "full recursive grammatical and semantical rules of the sort that many philosophers of language would demand" (Ibid, 174). It seems problematic to conclude that a biological human being who has memorized a rule book of a foreign language (in this situation, Chinese) to their native language (let us say English) is not recognized as understanding that language. Rey questions what more Searle would possibly want for an account of understanding because Rey does not see how the person in the room could have carried this out and still not understand Chinese. What Rey holds is that this type of internalization is in a general sense, what is normally involved when it comes to understanding a language like Chinese (Ibid, 174).

Searle refers to human beings as meat machines and he argues that biological, organic materials (that we, as humans, possess) are what are needed for understanding. This human being, who is operating the rule book and correlating them with the Chinese symbols in the CRA, is a

"machine" (according to Searle) that possesses the required organic make-up to be attributed understanding. Yet, according to Searle, this machine does not understand Chinese even if he has "internalized" the system. Rey asks the important question: "If neither the biology nor the program, alone or together, are sufficient, what more for Searle is needed?" (Ibid, 173). As Rey states, the big factor here is "internalization" because if a person in a room has internalized a language (like Searle requires), then it should be evident he or she has done enough to understand Chinese (Ibid, 173).

Rey holds that for an AI-functionalist, what Searle's idea of memorization amounts to is merely programming the CPU with the program of the whole system in such a way that they will be functionally identical. However, it is important to note that the room, in this case, would still embody a system were the man in it to go back to reading the external text. Rey then moves on to state that the AI-Functionalist will then hold that "the flesh and blood clause is inessential: memorization of the rules is quite enough. Aside from specious speciesism, what's Searle's argument now that it isn't?" (Ibid, 175). Rey concludes his criticism of Searle's CRA by stating that it is not a problematic thought experiment for Strong AI.

2.7) The Virtual Mind Reply

2.7)1. The Virtual Mind Reply Outline

This response can be viewed as a response that branches off from the Systems reply. Followers of this view also concede that the person in the room manipulating symbols does not understand Chinese. Proponents of the Virtual Mind Reply argue that what is doing the understanding in the machine, is a virtual mind that the system creates. The crucial difference between the Systems Reply and the Virtual Mind Reply is that the running system creates a new entity and this entity understands. This new entity is distinct from the CPU or the system as a whole. The focus of this response is the idea of whether understanding is created. So, the questions which the Virtual Mind Reply (VMR) poses are not "Does the man understand Chinese?" or "Does the system understand Chinese?" (Cole 2015). Rather, the question is "Does the running computer create an understanding of Chinese?" (Ibid, 2015).

David Cole argues for the VMR by stating that the failure of the man in the room's potential to understand Chinese does not give any indication that there is absolutely no sort of

understanding being created. The crucial characteristic of Cole's argument for the Virtual Mind response is that Searle fails to see that if the room operator does not understand Chinese, then it does not mean that no-one in the room understands Chinese. There is a new entity that is created that understands Chinese (Ibid, 2015).

Cole provides us with three claims that he believes arise from Searle's CRA:

1) the claim that the person following the English instructions would not understand Chinese;
2) the inferred claim that a computer following a program would not understand Chinese; and
3) the inferred final claim in the preceding summary that programming cannot produce understanding of a natural language. (Cole 1991, 401)

From these three premises Cole asks us to firstly, suppose that premise 1 is true because "Searle would not understand Chinese merely by following the instructions in English for syntactic manipulation" (Ibid, 401). For the sake of the argument Cole also asks us to accept that premise 2 is true. His argument rests in proving that premise 3 is false.

Cole leads on to conclude that artificial minds are possible as Searle's argument creates a new entity. This new entity is what Cole refers to as the *Virtual Person* (or *Virtual Mind*) (Ibid, 399). The new "thing" that is created is an entity that could be caused by the activity of something that is not a person (or mind). The two points that I wish to take from Cole's argument are that Searle's argument causes a new entity to exist "(a) that is not identical with the computer, but (b) that exists solely in virtue of the machine's computational activity" (Ibid, 399). Cole's conclusion is that "Searle's argument fails in establishing any limitations on Artificial Intelligence (AI)" (Ibid, 399).

This reply ties in satisfactorily with the doctrine of functionalism because asking this question focuses on whether understanding is created by the computer, instead of what materials are creating understanding. I like this idea because as Margaret Boden explains in her paper, *Escaping the Chinese Room*, our brains do not understand things but they cause understanding (Boden 1987).

2.8) The Robot Reply

2.8)1. The Robot Reply Outline

The Robot Reply, agrees with Searle's claim, that a computer stuck in a room cannot understand language or know what words mean (Cole 2015). However, proponents of the reply hold that if a digital computer placed in a robot body with sensors, video cameras, microphones, wheels or legs to move around with, and arms to operate things with, then such a machine would interact with the world by seeing and doing. The main claim of the Robot Reply is that a digital computer here, in a robot body freed from the confines of the room, can attach meanings to symbols and understand natural language (Ibid, 2015). The computer inside the robot body would do more than merely take in formal symbols as input and give out formal symbols as output. This computer (which would be the "brain" of the robot) would be the controlling force that would enable the robot to perceive, walk, move about, drink, eat or even make tea (Searle 1980, 420). The robot reply theorizes that the robot could do practically anything that a human could do. From this, it is believed that such a robot would have genuine understanding and other mental states (Ibid, 420). Of course for a robot to have genuine understanding, all other sufficient conditions for the ascription of understanding would have to be met. If a machine is built in this manner, then it is believed that a computer in this situation would be able to understand things in the world because it would be able to learn things by seeing and doing, similar to how a child learns about the world around it. To learn things by seeing and doing would require the robot to have the correct type of inputs and outputs, which would be the correct causal connections required for genuine understanding. This computer would do more than just take in symbols, manipulate them and then put out symbols as responses. It would do more than that because this computer could become a robot that can percieve, walk around, listen and do various other human activities. A robot like this could possibly be classified to have an understanding of the world around it.

2.8)2. Searle's Response

Searle responds to the Robot Reply by saying that the sensors and limbs to interact with the world merely provide more inputs to the computer. Searle says that the problem is that these will just be more syntactic inputs. Searle is against the Robot Reply because "additional

syntactic inputs will do nothing to allow the man to associate meanings with the Chinese characters" and this is all merely "more work for the man in the room" (Cole 2015).

Searle's response to the robot reply is a two-part response. Firstly, he celebrates and claims victory by replying that if the robot reply accepts that cognition is not only a matter of formal symbol-manipulation but also needs a set of causal relations with the external world, then his view of understanding is correct. He says this because his claim is that there is more to understanding than just manipulating symbols, such as what he believes his CRA argues. So, according to Searle if proponents of the robot reply agree with this then what he is saying is correct (he believes). However, it is not clear that this is the case and so he should not claim a victory so quickly. The robot reply argues that a robot built in this way needs more than symbol manipulation and so, the "more" that we are saying the robot needs is a connection to the world but it does not have to be a biological, organic connection that Searle believes is necessary for understanding. Searle then goes on to respond to this new point by saying that adding extra perceptual motor functions to a machine does not mean that it in turn adds intentionality and understanding (Boden 1987, 10). To show that these inputs will just mean that the robot computer does more processing and is still not understanding, Searle tweaks his Chinese Room scenario to account for his response to the robot reply. Imagine the person in the room receives, along with the Chinese characters slipped under the door, a collection of binary digits which appear on a ticker tape in the corner of the room. The rule book has also been tweaked and now it uses the digits on the tape as input in addition to the Chinese Characters. The man does not know that the symbols on the tape are merely a digitized output of a video camera as well as other sensors (Cole 2015). Searle asserts that "additional syntactic inputs will do nothing to allow the man to associate meanings with the Chinese characters" and there will just be "more work for the man in the room" (Ibid, 2015).

Searle responds further to the robot reply and outlines a scenario of a robot. This scenario, he believes, shows that no understanding is created. The scenario is as follows, Searle asks us to imagine a robot, which instead of having a computer program that makes it work, it has a miniaturized Searle within its "skull". We can call this entity, "Searle-in-the-robot". This Searle-in-the-robot is given a new rule-book and he also shuffles papers around and passes Chinese symbols in and out of the slots in the wall, just like the man in the room did in the

example before this. However, the scenario is then changed, instead of the Chinese characters being passed into the room on paper, the Chinese characters are triggered by causal process in the cameras and audio-technology in the robot's eyes and ears. Instead of the outgoing Chinese characters being received by Chinese speakers' hands, the characters are received by motors and levers connected to the robot's artificial limbs, which move because of all this. The point that Searle is trying to make here is almost one of a sarcastic nature. He holds that this robot would "apparently" be able to do more than just answer questions in Chinese, because it would be able to recognize raw bean sprouts and throw them into a wok as well as any human chef (Searle 1980, 420 (In Boden 1987, 10)). The main point that Searle is making here is that he has satisfied all the conditions of the robot reply and this still does not lead to any understanding.

2.8)3. Response to Searle

The most crucial part of the robot reply is the idea of causal connections being required to attribute intentionality to the system. This ties in with what was previously discussed on the topic of Daniel Dennett's Intentional Stance in Section I. Let us imagine a scenario where we have a system setup in our homes to keep our homes in order, like an *Amazon Alexa* or *Google* home device. These systems rely on speech recognition and you can control electronic elements in your home with such a device. You can "tell" it to turn lights on, start the dishwasher or play rain sounds to help you sleep. These devices are merely plugged in and work based on speech recognition. For argument's sake, let us say that one of these devices has beliefs about what it must do to keep our homes in order, it has beliefs *X, Y* or *Z*. Rather than being too explicit with the exact beliefs of such a device, let us say that it believes that *X* is too *Y* or *Z* when it comes to maintaining order in our homes. If it wants *X* to be more *Z* then it needs to do *N* (Dennett 1981, 30). Searle would hold that a device would not have the right causal connections to believe anything and it is merely following a speech recognition program that makes it do *N* when it wants to be more *Z*. Searle would hold that all a device like this is doing, is following a program and no thinking or beliefs are involved here.

With what was said in the robot reply, let us now imagine that we can improve such a device's modes of attachment to the world and turn it into a robot home helper/manager. For example, imagine that it has ways of learning about what the best ways are to keep a house clean and

tidy. It is given a video camera that acts as an eye and it begins to notice that people could live better in the house in reaction to the ways they do things in their homes. The occupants of the house have friends over and their friends give them home advice (something the robot hears with its audio recorder microphone and it then uses to improve the home). The robot is also connected to the internet and so it can learn more about interior design and home aesthetics. All of these additions will warrant extensive, intricate upgrades to its inner structure. We can turn it into a machine that is also given behavioural versatility. This allows it to move around and decide on where to rearrange furniture or which cleaning products it should use. Dennett holds that all of these features will add more internal complexity to the machine (Ibid, 30). By enhancing the connections between the object and its environment where it resides, we have enhanced the semantics of the terms described earlier (*X*, *Y*, *Z* and *N*). The more complexities that we add to the system, the more demanding and richer the semantics of the system become. Dennett supports the idea that with all these new complexities in place, we will start to say that the machine in question has beliefs about design, cleaning, home management and so on (Ibid, 31). These are the causal connections that are required for mentality.

If the robot helper were moved to a new home, it would arguably start to form beliefs about how to act in accordance with its new environment. Its sensory attachments to its environment could be sensitive and distinguished enough to respond suitably to the change of scenery. Dennett believes that an entity as described constantly mirrors the environment in which it is placed and that there is therefore a representation of the environment in the makeup of the entity (Ibid, 31). A futuristic robot as described earlier in the robot reply would have even more complex and demanding representations of the world and so, with its causal connections we would be even more justified in saying that it has beliefs.

Margaret Boden, in her paper *Escaping the Chinese Room*, has a two-part argument that aims to refute Searle's claims about the Robot Reply. Firstly, Boden argues that there are problems with the example of the Chinese Room and secondly, she critically engages with Searle's assumption that computer programs are purely syntactic (Boden 1987, 9).

In her refutation, Boden agrees with the argument of the Robot Reply. She argues for the idea that a robot provided with a detailed script, camera-fed visual programs and arms and legs to

help it interact with the world, should be considered to understand the world around it. Boden argues that Searle's reply above is not a sufficient rebuttal and it does not show fault in the Robot reply. As it changes the example of the robot. One cannot disregard the robot because it is taking more inputs into consideration. Placing sensors, microphones and cameras will not merely mean that the robot will just have more inputs but it will have more ways of interacting with the environment to have the correct causal connections needed for mental states. It will mean that the robot computer will begin to act in such a way that it begins to interact with the world around it and then learn to carry out basic tasks. As Dennett states, the new complexities that are added to an entity, should not restrict its ability but rather the more connections to the external world, the more demanding and richer the semantics of the system become (Dennett 1981, 31).

Having explained the robot reply and how a robot such as described could have the correct causal connections to understand the world around it, it is important to refer to what Georges Rey said about understanding. To understand something requires more than merely having a bunch of rules that correlate to symbols. People (or robots) able to understand something requires dealing with inputs from a variety of other processes, such as perception, belief fixation, problem solving, preference ordering and decision making (Rey 1986, 171). Rey argues that Searle's CRA would only challenge Strong AI if the person in the room was able to relate Chinese characters to the other inputs mentioned above, like perception or preference ordering. This is where the Robot reply becomes relevant to Rey's argument. Understanding, and more specifically understanding a language (like Chinese or English), "involves being able to relate the symbols of the language to at least *some* perceptions, beliefs, [and] desires" (Ibid, 172).

A large part of Searle's CRA relies on the idea that the man in the room must understand the English rulebook to correlate to the Chinese symbols. It seems plausible to argue that over time, the man will not see the Chinese symbols as being completely "meaningless" because after a while the man would likely be "*in a position* to determine *something* about what the symbols mean" after the man (who is now the whole system) correlates the symbols with the English rulebook in his head (Ibid, 173). This applies to what is being said about the robot home manager. Over time and with the correct causal connections to the world, the robot could

understand what it is doing to keep the house in order, instead of merely following a program. These correct causal connections of the robot to the world would help it constitute reference and aboutness because it would likely be able to determine something about what it means for it to do *X* or *Y* to keep the house in order.

In terms of mental content, externalism holds that for an entity to have certain types of intentional states (like beliefs and desires), it is crucial that the entity be connected to an environment in the correct manner, as I have highlighted in the robot reply (Lau and Deutsch 2016). Externalism is relevant to the robot reply because it shows how the correct causal connections to the external world ensure that symbols can be interpreted and they do have meaning. This goes against Searle's claim that the CRA is comprised of uninterpreted symbols without any meaning. The most familiar thought-experiments that argue for externalism of mental content draw on physically indistinguishable individuals that are inserted into varying environments. From here it is argued that certain beliefs and thoughts are held by one of the individuals but not the other. What this calls attention to, is that some mental contents do not come from within a system and so if this is the case, it shows that externalism is true (Ibid, 2016). If externalism about mental content is true, then it shows that mental content may not reside from within an entity and a robot like in the robot reply could be able to understand. The most well-known thought-experiment of this type is Hilary Putnam's Twin Earth scenario.

For his "Twin-Earth" thought experiment, Putnam requires us to imagine a remote planet named "Twin-Earth" in the year 1750. This remote planet is identical to our Earth except for a few differences. A crucial difference to take note of and one that holds Putnam's argument together is that instead of water (made up of H_2O), this planet has a water which is made up of a compound named XYZ. The two substances are very similar and they can both be found in rivers and oceans on their respective planets. In the year 1750, it seems likely that no one would be able to distinguish between the two substances, namely that one was H_2O and one was XYZ. Putnam holds that if someone on our Earth referred to water then they would be referring to H_2O and not XYZ. Now, if this Earth person were to be transported to Twin-Earth and say that he was drinking water, he would be saying something incorrect. Likewise, if someone from Twin-Earth came to our Earth, they would refer to H_2O and not XYZ when speaking about water (Putnam 1975, 223).

It is important to note that Putnam's argument was originally applied to "linguistic content", however it can be shown to apply to mental content, and this applicability to mental content makes it relevant. Along with not referring to actual "water" when the Twin-Earth person speaks about "water", he also does not hold beliefs about "water" (Brown 2016). He holds beliefs that play the same part in his mental life, as the Earth person's water-beliefs play in his, respective, life. The important point is that in the Twin-Earth person's case, his beliefs are not about "water". So, while the Earth person might believe that water is wet, for example, the Twin-Earth person does not believe this. Since the Earth person and the Twin-Earth person have the same innate qualities but the person from Earth believes that water is wet, while the other person does not, it highlights that mental content cannot be specified only by intrinsic properties (Ibid, 2016). Therefore, if externalism is true, then meaning depends on external connections and so a robot with the correct causal connections would be able to understand.

2.9) The Brain Simulator Reply

2.9)1. The Brain Simulator Reply Outline

To understand the Brain Simulator Reply, one is required to imagine a computer that works in a way that is different to a standard machine which merely works with "scripts and operations on sentence-like strings of symbols" (Cole 2015). Instead, this new sort of machine simulates the exact sequence of nerve firings that occur in the brain of a native Chinese language speaker when that person is conversing in Chinese. In this regard, the computer would now be working in the same manner as the brain of a native Chinese speaker. If it processes information in the same way as the brain of a native Chinese speaker, then it would be difficult to argue that the machine is not understanding Chinese (Ibid, 2015). The brain simulator reply requires one to imagine a type of program that simulates the actual sequence of nerve firings that occur in the brain of a native Chinese language speaker when that person understands Chinese (Ibid, 2015).

2.9)2. Searle's Response

Searle's response to the Brain Simulator Reply, is that it does not make any difference to understanding if the computer system has changed to mimic the brain of a native Chinese speaker. This is because he believes that the simulation of brain activity is not comparable to real brain activity. Searle adopts an example of valves and water pipes in the Chinese Room to

argue for his view. He asks us to imagine that there are a huge set of valves and water pipes in the room. These valves and water pipes are set out in the same arrangement as the neurons in the brain of a native Chinese speaker (Ibid, 2015). To understand how the water pipes work, imagine that instead of rearranging symbols the man in the room is operating this complex set of water pipes with valves connecting the pipes (Searle 1980, 421). When Chinese symbols are passed into the room the man analyses the program and looks at which valves he needs to turn on and off. Each water connection used in the example is equivalent to a synapse in the brain of a native Chinese speaker. The entire system is set up so that once all the correct firings are completed (which means turning on the correct faucets) the Chinese answers to the questions passed into the room emerge at the output end of the series of pipes (Ibid, 421).

With this example, Searle asks the question: "Now where is the understanding in this system?" (Ibid, 421). This example takes Chinese as the input, it mimics the formal structure of the synapses of the brain in the form of water pipes and it then gives Chinese as output. Searle argues that in this tweaked scenario there is still no understanding of Chinese. According to Searle, the main problem with the Brain Simulator reply is that it is simulating the incorrect things about the brain. It is only simulating the formal structure of the sequence of neuron firings at the synapses. Searle believes that if there is only a simulation of the formal structures then what matters most about the brain, like causal properties and its capacity to create intentional states, won't have been simulated (Ibid, 421). Searle concludes this response by adding, that the water pipes example highlights how the formal properties are not sufficient for the causal properties because "we can have all the formal properties carved off from the relevant neurobiological causal properties" and still not have understanding (Ibid, 421).

2.9)3. Response to Searle

In response to Searle we can refer to the computational account of the mind. This is the version of psychofunctionalism that holds that our minds compute and that thinking is the processing of information involving the manipulation of symbols. If we have mental states then it seems plausible to maintain that a system functioning psychologically like us, also has mental states. To falsify the brain simulator reply, it appears one would need to ask what the argument is to

show that the brain simulation does not understand if these conditions are met. To do this, one would have to show that inorganic materials do not have the right causal powers and cannot contain mental states, which is something that would be difficult to do.

Searle's response to the brain simulator reply is weak because the causal properties of the brain are not only created by organic substances. To strengthen Searle's account, he would need to show that the "causal properties" cannot be reproduced in different matter. The onus here would be on him to show this. Similarly, with what was said about Block's homunculi-headed creature, it is not clear (given our ignorance of the relation between neural structures and subjectivity) that the homunculi-headed creature lacks subjectivity. The artificial materials that could simulate brain activity could possibly also have intentional states. A personal identity thought experiment can be used to aid my argument.

Imagine we have a living, breathing human being who is suffering from a brain disease. The brain disease is attacking his synapses and so his brain is slowly deteriorating. We are quite far into the future in this scenario and so modern medicine is capable of extraordinary feats. Scientists have created an artificial brain that simulates neural activity and simulates human thought. This man suffering from the brain disease goes to his doctor for help and his doctor informs him of a new procedure to slowly replace his organic synapses with the artificial synapses from this artificial brain. The doctor reassures the patient that the procedure will go smoothly and at the end of the operation the man will continue to live his life normally. The procedure takes many days because as the artificial synapses become part of his brain they assimilate with his organic synapses. Over the course of the procedure his memories and thoughts transfer across to the artificial synapses. What this means, is that by the end of the procedure he will not have forgotten anything and his cognition will remain intact as if nothing had ever happened. The doctor begins the procedure and 365 days later, the man is free from the brain disease and he now has an artificial mind that simulates the neural activity of his old organic brain. With this type of scenario, it could be difficult for Searle to argue that this man's brain activity is not the same as the organic brain activity that the man previously had. The man who now has an artificial brain would likely have the same neural activity as someone with an organic brain. It would be difficult

for Searle to argue that this man, with his artificial mind, is not capable of understanding the world around him.

When it comes to reproducing consciousness or mental states artificially, Searle holds that the way to go about this is to replicate the actual neurobiological framework of consciousness that is in organisms like ourselves (Searle 1992, 92). With this said, he also believes that any system with the ability for consciousness and mental states must be able to replicate the causal powers that we have within our own brains (Ibid, 92). This adds weight to the points made by the brain simulator reply because it seems difficult to deny mental states to a system capable of replicating the causal powers of the brain.

2.10) The Combination Reply

2.10)1. The Combination Reply Outline

Having outlined the CRA, discussed the responses to the CRA, analysed Searle's objections and given my rebuttals to Searle's objection, I will now move the discussion towards the "Combination Reply". The Combination Reply, as the name suggests, is a reply that takes all the replies into consideration. The idea here is that even if each reply on its own is not sufficient, then together, as a combination of responses, they can possibly hold more weight and decisiveness. Proponents of this view urge one to: "imagine a robot with a brain-shaped computer lodged in its cranial cavity, imagine the computer programmed with all the synapses of a human brain, imagine the whole behaviour of the robot is indistinguishable from human behaviour, and now think of the whole thing as a unified system and not just as a computer with inputs and outputs" (Searle 1980, 421). It is then argued that surely if a machine such as this (with all the features of each reply) were to exist, then we could attribute intentionality to the system. This provides us with an interesting case and I will build on this response to argue against Searle. It seems plausible to attribute mentality to this type of machine. This type of machine is described as "programmed with all the synapses of a human brain" (Ibid, 421).

2.10)2. Searle's Response

Searle responds to the Combination Reply by arguing that we would only find it rational and attractive to accept the hypothesis that a machine such as this has intentionality, only if we do not know more about the robot (Ibid, 421). This response by Searle follows a behaviouristic framework because Searle believes other than the appearance and the behaviour of the machine, the other components of the combination reply are irrelevant. He believes that if we can build a machine whose behaviour is identical to a human's behaviour over a vast range of tasks, we would attribute human intentionality to it, unless there was a certain reason not to. This seems like a strange response because what I understand Searle to be saying here is that behaviouristic machines can be ascribed intentionality.

2.10)3. Response to Searle

The combination reply is quite fascinating because it allows us to imagine robots and machines from famous futuristic, sci-fi films. The type of machine that is being argued for in the combination reply is something like *C-3P0* from the *Star Wars* Universe. *C-3P0* is a human like machine as he is shaped like a human with arms and legs. He speaks and moves around like a human being. The human (as well as alien) characters in the *Star Wars* films interact with and treat *C-3P0* as a compatriot of theirs. What is going on within the circuitry and materials of *C-3P0* should be enough to say that *C-3P0* possesses the capacity to understand what he is doing. *C-3PO* is possibly a good example of a machine with the correct internal operations and processes and so a machine such as this could be regarded as being able to understand. This machine responds to stimuli and interacts with the world and universe around it. Regardless of the material makeup of his mind and body, it seems that he understands the things that he does and the things that he says throughout the duration of the *Star Wars* films. I would regard *C-3P0* as an intentional system and a true believer, who can be ascribed mental states because it seems that he has the correct causal connections to the world and he acts rationally to fulfil his own desires and goals. Due to this it could be said that *C-3P0* has beliefs, hopes and desires. These features make him an intentional system.

C-3PO can be used as an example of a machine that highlights the combined elements of the responses, to argue, that with such a machine, the idea that there is now understanding of a certain sort is not implausible. By taking into consideration all that was said in the systems reply, the virtual mind reply, the robot reply and the brain simulator reply, it is not difficult to see how a machine, such as *C-3PO* with these combined elements, could possibly understand.

2.11) Conclusion

I hope that I have provided enough evidence to show why John Searle's Chinese Room thought experiment is not sufficient to refute Strong Artificial Intelligence. As mentioned, a Strong AI system cannot be compared to a Turing Machine and I can conclude that Searle misunderstands the idea of Strong AI. The views that I have argued for are *The Systems Reply*, *The Robot Reply*, *The Virtual Mind Reply*, *The Brain Simulator Reply* and *The Combination Reply*. These replies show that Searle's original responses to them do not guarantee a defence for his argument. In the next section of this paper I will argue for an externalist account of understanding to further support my position. With what has been argued in this section, the responses to Searle's CRA defused Searle's argument to a certain extent but further elements need to be discussed to make the argument against Searle more persuasive. The further elements to be discussed in the next section, are an account of understanding, a theory of mind to apply to a machine, machine learning, deep learning and real-world AI scenarios.

Section III:

Machine Learning, Connectionism & Understanding

3.1) Introduction

In this section of my paper, I will critically engage with the theory of mind known as connectionism, the idea of deep learning and machine learning, and a notion of understanding. I will begin this section by outlining an account of understanding. To do this I will compare and contrast externalism and internalism to show which account of understanding is the more plausible. I will then move onto discuss connectionism and the classical theory of AI. I will also compare and contrast these two theories to show which is the more plausible account. Next, I will discuss machine learning and deep learning, as well as look at some real-world examples of AI machines. From here I will analyse an objection to connectionism known as "The Chinese Gym". I argue that Searle's "Chinese Gym Argument" is not sufficient to contradict the plausibility of connectionism. Connectionism is the idea that sets out to explain intellectual abilities using artificial neural networks (Garson 2016).

3.2) An Account of Understanding

3.2)1. Internalism

An internalist holds that understanding is something that comes from within and so it is "a purely internal product of internal physiological processes" (Searle 1987, 230). Searle would be regarded as an internalist. For example, when Searle says that he does not understand Chinese, he believes that it is because there is something lacking within him that makes it the fact that he cannot understand Chinese (Searle 1980, 422). With this line of reasoning, we could say that Searle holds that the man in the room cannot understand Chinese because there is something lacking within him. An internalist is therefore someone who holds that to understand something requires an internal process occuring in the right way.

Within the internalist position is the importance of "conscious awareness". Conscious awareness, or consciousness, is an important aspect for Searle as he believes that consciousness is an inner, first-person, qualitative phenomenon (Searle 1997, 5). Searle holds that the term "consciousness" refers to our states of sentience and awareness and so consciousness is an internal phenomenon. For Searle, our states of sentience and awareness usually start when we wake up from a dreamless sleep and continue until we go back to sleep again (or more radically, if we fall into a coma, die or become "unconscious" in any other way). Searle believes that

dreams are also a form of consciousness too but they differ from complete waking states. With all this in mind, it can be said that consciousness can be switched on and off and so it is something within us. For Searle, consciousness is an internal phenomenon and a system or entity is either conscious or it is not conscious. Within the field of consciousness, there are varying degrees of consciousness, which extend from drowsiness to full awareness (Ibid, 5).

This goes back to what Searle says about "the right causal powers" that I raised in Section I of my paper. Searle holds that all of our mental phenomena, such as beliefs, desires and hopes, are caused by neurophysiological processes in our brains (Searle 1990, 29). If caused by neurophysiological processes within us, then these mental phenomena arise internally and these internal processes have to be conscious. If internalists hold that beliefs arise from within, then for them, understanding is an internal phenomenon.

3.2)2. Externalism

In contrast to the internalist account, the externalist view of understanding, is the idea that what is needed for understanding is a suitable relationship between the subject and the external world (Preston 2002, 40). The externalist account holds that understanding is not something that comes from within and so it is not purely internal. For someone to understand something there must be external factors that contribute to what it is they are trying to understand. This is different from the internalist account because externalists hold that for an entity to have specific types of intentional states, such as beliefs and desires, it is important for said entity to have the correct causal connections to the outside world (Lau and Deutsch 2016). These correct causal connections are those highlighted in the robot reply. As outlined, internalism is different as it denies that external connections are necessary, because having intentional mental states relies solely on our internal properties (Ibid, 2016).

Under Wittgenstein's externalist account, when we say that there is "understanding" of a sentence, it means that to understand, numerous things must occur before and after the sentence has been read. For example, subject *S* can understand something, only if *S* has a specific history relating to what they are trying to understand. This means that for *S* to understand, *S* must have had a history of learning and training. Additionally, *S* must have taken part in a specific social

environment relating to what it is that they are trying to understand (Proudfoot 2002, 177) What can be taken from this is that the externalist account of understanding has an "experiential" explanation. One gains the ability to understand more and more things as one moves through the world experiencing different stimuli.

Diane Proudfoot in her paper, *Wittgenstein and the Chinese Room*, discusses a theory of Wittgenstein's known as the "situated cognition approach" (Ibid, 178). By following this theory, Wittgenstein argued that cognition must be seen in terms of activities, skills and common-sense knowledge. The emphasis here must be placed on situated reasoning and environment-specific knowledge. What this means, is that reasoning and knowledge must be related to the environments of which we (or a machine) are a part. This goes against the view that cognition is an internal phenomenon, as Wittgenstein argues that the "sophisticated cognitive performance does not involve internal world-modelling" (Ibid, 178). Proudfoot goes on to state that under Wittgenstein's externalist conditions for understanding, there is nothing preventing a machine (living or otherwise) from beginning to understand the world around it (Ibid, 178). For someone to understand something, they must be connected to the external environment in such a way that allows them to gather information from their environment and make conclusions about the information. This goes back to the robot reply, perhaps for a robot to understand it needs to be connected to its external environment in the correct way, which then causes it to form beliefs.

With regards to the robot reply, it is important to note that firstly, for understanding, the system needs the right internal processes that correspond to perception, belief fixation, problem solving, preference ordering and decision making. However, under the externalist account, the right internal processes on their own are not sufficient, as the correct external connections are required too. Secondly, for the CRA, questions and responses to not be generalized, the system needs to provide for many different input questions. What Searle's large rule book in his Chinese Room Argument needs to account for, are appropriate outputs from arbitrary inputs. It does not seem plausible for a rule book to account for the vast array of questions that can be inputted into the CRA. However, an entity who has experienced the world around it would possibly be able to respond to random questions. This is where externalism comes in, because

only a robot that has the correct external connections to the world would possibly be able to give appropriate outputs to arbitrary inputs through an understanding of its external world.

With regards to the correct causal connections to the external world, the current CEO of Microsoft, Satya Nadella believes that machines will exhibit understanding once they "learn to learn" (Nadella 2017, 153). In his book, *Hit Refresh*, he says that when machines "learn to learn" they will be capable of understanding. On the point of consciousness mentioned above it seems relevant to briefly discuss the idea of a machine, "learning to learn" because a machine doing this could possibly be ascribed mental states and be regarded as conscious on Searle's definition of the term.

Nadella holds that the decisive moment for machines is when they will be able to "learn to learn". Machines will be regarded as doing this when they are able to generate their own programs. Just as humans can do this, computers will go further than merely imitating what humans do and they will be able to invent new, improved solutions to problems. What will make this possible are the new technologies being implemented in machines today. According to Nadella, AI can be compared to a ladder and we are only on the first rung of the ladder. At the final rung of ladder, we will find artificial general intelligence and complete machine understanding of human language (Ibid, 153).

This idea of learning to learning, is what a robot from the robot reply would be doing. Such a robot as described would be able to interact with the world around it and then begin to respond to stimuli based on its previous experiences. It follows that a robot learning to learn in this way could possibly be ascribed mental states.

3.2)3. The more plausible account

Leaving the technicalities of our neural systems and the argument of machines to one side for the time being, It is plausible to agree that we as human beings, interact with the world around us and gather information from our environments. From this, I want to highlight that without the causal connections to our environments we would find it difficult, if not impossible, to hold any beliefs, desires or hopes. By taking information from our environments and assessing it, we come to "understand" the things around us. In this sense, understanding should not be

regarded as a purely internal phenomenom. The "understanding" process must surely rely on the correct causal connections to our environments. Unless Searle or other internalists can prove that internal processes are all that are needed for understanding, we cannot rule out the plausibility of the externalist account.

To better understand the externalist account (as well as the debate between internalists and externalists), we can turn to epistemology or what is otherwise known as, the theory of knowledge and examine the parallel/analogical case of externalism with regard to knowledge. With regards to the justification of beliefs in epistemology, there has been a substantial debate on whether justification is internal or external (Steup 2018). To understand the debate, it is important to note that when a belief is justified, it is because there is something that makes the belief "justified" (Ibid, 2018).We can refer to the things that make beliefs justified as *J-factors*. So, the debate between the internalists and externalists, is a debate over what *J-factors* are. Internalists hold that beliefs can be justified internally because we have a special kind of subjective access to *J-factors* and they are recognizable on reflection (Ibid, 2018). Externalists, on the other hand, hold that beliefs are justified externally because *J-factors* are reliable processes (such as sense-perception).

It seems problematic to hold that beliefs can be justified internally, because for one to possess the information needed to justify a belief internally on reflection, there must have been some external factors that provided one with the original belief on which to reflect. For example, if we think about the example of Pluto in Section I, imagine your friend tells you that Pluto is a planet. Imagine then that you reflect internally on what you know about Pluto and then correct him, by saying Pluto is not a planet. Even if your justification for this belief was something that you reflected on internally, it cannot be the case that your belief was a purely internal phenomenon because you would have gathered the *J-factor* for the belief that Pluto is not a planet from an external source. Your exposure to an external *J-factor* would justify why you now believe that Pluto is not a planet. Through external interactions with their enviroments, children form beliefs and justify those beliefs, as adults we seem to do the same thing and so, with all this taken into consideration, the external account of justification seems to be the more plausible account to follow (Ibid, 2018). The important point here is that for externalism a belief is justified because it can be justified by what makes it the case, which are the correct causal connections to the external world (Ibid, 2018).

To critique the internalist account of understanding, we can go back to what Searle has said about the correct neurophysiological processes and "the right causal powers" that brains possess to give us the ability to understand. When Searle says that Chinese writing looks like meaningless squiggles to him and that he therefore does not understand Chinese (Searle 1990, 26) it might not be because he is lacking an internal capacity that makes him not understand Chinese. There is the possibility that Searle does not understand Chinese because he has not been exposed to Chinese in the way that a young child would be exposed to Chinese. If Searle had experienced the Chinese language through constant exposure to a Chinese environment, then Chinese writing would no longer appear as squiggles to him. The externalist account of understanding which focuses on our environment-specific knowledge (Proudfoot 2002, 178) appears to be the more plausible account of understanding because the correct causal connections to our environments seem necessary for understanding to be ascribed to an entity. To conclude, we can turn to the robot reply once again as it seems reasonable to now hold that any robot set up in the right way (with the correct causal connections to its environment) would be able to understand the world around it and therefore be ascribed mental states.

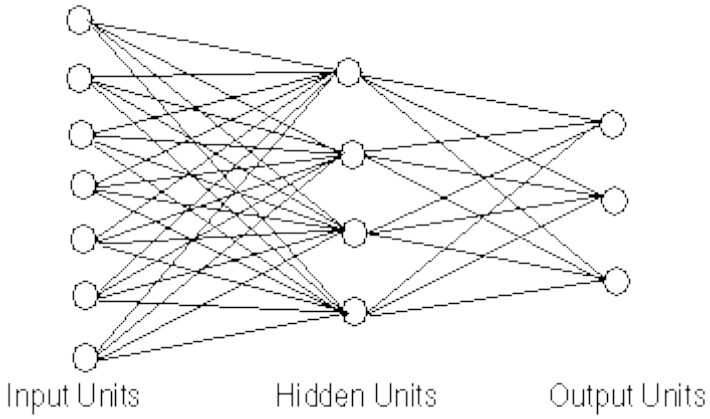
3.3) Connectionism & The Classical Theory of Mind

3.3)1. Connectionism

Research in the field of *connectionism* investigates the properties of neural networks. Connectionism is relevant because it can be seen as a way of trying to model how our brains function and it is the idea that sets out to explain intellectual abilities using artificial neural networks (Garson 2016). The theory of connectionism is a reverse engineering of sorts because it aims to try various simulations of brain functioning, which can then be used to better understand ourselves. Searle makes the claim that, as far as we know, every mental event that we experience (be it a belief about our favourite music, to thirst, or to memories of childhood) are caused by particular neurons firing in particular neural architectures in our brains (Searle 1990, 29). Perhaps connectionism, as a simulation of our brain functions (within a machine), could provide such a machine with the capacities that are sufficient for mental states. Turing highlights the relevance of a system like this because he holds that gaining experience and learning like a child might be what we need for understanding. Turing holds that we should aim to manufacture a machine that simulates a child's brain and then "educate" it to eventually lead to a machine with an adult's brain (Turing 1950, 456).

The theory of connectionism refers to a system, or a neural network, that is comprised of neuron-like processors called nodes or units. A connectionist system is not a physical device made up of neuron like processors but rather it is a simulation of a neural network on a standard digital computer (Horgan 1997, 11). These neural networks are made of vast numbers of artificial neurons along with, small, simple processing units. These artificial neurons and processing units are related to one another by connections. The excitatory or inhibitory 'weight' and the threshold for firing of these connections can be altered by the programmer (Preston 2002, 12). Such a system has the capacity to learn from experience because we can say that a connectionist system is "trained" instead of merely programmed. On each connection, there is a weight that represents the strength of each connection between the units that it links. How this program runs is by updating in parallel the activation of each unit based on the activation of the units to which it is linked. When a machine is said to "learn" in a connectionist system, what this means, typically involves adjusting the weights between units in ways that reinforces what the machine must and must not do in certain situations. Various connections are strengthened or weakened depending on whether actions should be repeated or forgotten by the machine. How this works, is that a program depicts a set of virtual neurons. It then takes these neurons and assigns random numerical values to them, otherwise known as "weights" to the links between the "neurons" (Hof 2016). In this scenario, a connectionist system is the correct type of "program" that could have cognitive states because the way it functions explains human cognition. The weights that are assigned numerical values regulate how each neuron responds "to a digitized feature such as an edge or a shade of blue in an image, or a particular energy level at one frequency in a phoneme, the individual unit of sound in spoken syllables" (Ibid, 2016). These patterns of activation describe the movements of the weights between the units, which reinforce how such a system should respond to stimuli.

The image below provides us with an example of how a simple neural net operates in a connectionist structure. The fact that a connectionist system can be trained rather than merely programmed is important because the idea of being trained and learning from experience are what could be sufficient for understanding.



An example of a connectionist structure.
Source: *Stanford Encyclopedia of Philosophy* – Connectionism (Garson 2016)

The sort of connectionist system that is possibly sufficient for understanding and that has been described thus far, is a simulation on a computer. However, as technology improves a connectionist system comprised of physical neural nets that make up a physical aritificial "brain" could be implemented in a machine. This could be described as a "real" connectionist system. The intricacies of how this would work and how it would be made are beyond the scope of this paper, yet it is an interesting point to briefly mention. In Section II of my paper under the discussion of the Brain Simulator Reply to Searle's CRA, I raised a scenario where a man's dying, organic brain is replaced by an artificial brain. In this futuristic scenario we would assume that after slowly replacing the man's organic brain with artificial "brain pieces", the man's new inorganic brain works as well as his old organic brain. This makes one think that if a scenario with an artifical brain were possible, then we could place an actual, physical, artificial brain into a robot to cause mental states.

3.3)2. The Classical Theory of Mind

The "Classical Theory of Mind" rests on the idea that the mind is something similar to a digital computer processing a symbolic language (Garson 2016). The Classical Theory of Mind is also known as the Computational Theory of Mind (CTM). This is the theory of computationalism that was raised as a version of pyschofunctionalism in Section I. To review, computationalism is a theory that holds that psychological states are distinguished by their casual connections with sensory inputs and behavioural outputs (Preston 2002, 9). Here information can be

represented by strings of symbols, comparable to how data is represented on a computer with "1"s and "0"s. In this sense, our minds "compute" because thinking is caused by manipulating symbols.

In this paper, I have already shown with my discussion of Searle's CRA that Searle believes that the most effective way for us to study thinking is by assessing computational symbol-manipulating programs (which is the theory of computationalism) (Searle 1984, 43). This idea lead Searle to hold that what goes on in his Chinese Room thought-experiment is what goes on in the operation of a computer, leading to his claim that if what goes on in his Chinese room does not add up to understanding, then the operation of a computer program does not amount to understanding either. The theory of computationalism relates to Alan Turing's *Turing Test* because such a device manipulates symbols, just like how a human agent manipulates marks on paper while working out mathematical symbols (Rescorla 2017).

3.3)3. The more plausible account

Advocates for connectionism, believe that connectionism is the more plausible theory of mind because, as mentioned, there is a suggested similarity between the simulation of neural networks in a machine and our own brains (Ibid, 2017). In such a system nodes resemble neurons and the connections between the nodes are similar to the synapses in our brains (Ibid, 2017). With this said, connectionists hold that this theory is more biologically reasonable than the computational theory as it is a sort of "reverse engineering" of our brains. The idealized expectation for a connectionist system is that if we can simulate the psychological phenomenon of thought then a machine with such a "brain" could have mental states (Ibid, 2017).

Computationalists respond to this idea by highlighting that connectionist systems might not be as biologically plausible as believed because the real neurons in our brains are much more varied than the simplified nodes in a connectionist system (Ibid, 2017). In addition, computationalists raise the issue that connectionists cannot guarantee that the nodes simulating neurons even hold the properties of real neurons because we do not know enough about how our own brains work (Ibid, 2017).

This objection raised towards the connectionist system is based on intuition and so in order for it to hold more weight and to disregard connectionism as a theory of mind, computationalists

would have to somehow prove that a connectionist system is not biologically plausible. In defence of connectionism, the most important thing here might be that connectionism is **more** biologically plausible than computationalism (Ibid, 2017). Connectionism is relevant to this discussion because of its emphasis on neural nets. The simulation of neural nets plays a large role in more technical computer science and as will be discussed in the sections that follow, computer scientists rely heavily on connectionist modelling and neural nets when creating state-of-the-art machines today. The connectionist theory of mind offers us a system that can learn from experience rather than a system that has to have every step programmed. This is what connectionism offers us that computationalism does not and this is what makes connectionism more biologically plausible because it can be argued that learning from experience is that way in which a child would learn.

As discussed, from his *Chinese Room Thought Experiment*, Searle concludes that a computer or a type of system that runs computer-like programs will not be able to understand language. He believes that all a machine does is merely move around symbols that hold no actual meaning for the machine. By contrast, Paul Thagard, says that, "if the machine interacts with the world and is able to learn concepts from interactions with the world as well as the programmer, then there is the possibility that its semantic capabilities will be no different from our own" (Thagard 1990, 271). This was a bold claim to make in 1990 but there is now greater reason to think it is true. At the time of writing (1990), Thagard also states that the available research and technology did not show any signs of developing a "context-sensitive and semantically sophisticated computer" (Ibid, 271). However, this should not stop one from believing that it can happen, as technology has improved.

3.4) Machine Learning, Deep Learning & Real World AI

3.4)1. Artificial Intelligence, Machine Learning & Deep Learning

Machines that complete tasks that would normally require human intelligence to complete, are the types of machines that could be ascribed mental states. In addition to this, machine learning is a part of the field of AI that aims to make machines improve their 'knowledge' (so to speak) and improve their capacities for performing various endeavours (Ibid, 261).

To review and put these ideas into more modern terms, Artificial Intelligence, or AI, describes machines that have the ability to execute tasks that we would normally require a human to do. These tasks are the sort to require a certain "level" of intelligence in order to be carried out (Marr 2016). Machine Learning, or ML, is the application of AI systems. ML is the idea that machines should be given access to large amounts of data and then by sifting through the data they learn for themselves (Ibid, 2016). As discussed in connectionism, machines learn when they have access to and gather enough information to know what to do or what not to do in certain situations.

The term "machine learning" was founded in 1959 by Arthur Lee Samuel, when he put forward the idea that instead of programming machines to follow rules, we should rather spend time researching how we could teach them to learn things for themselves (Ibid, 2016). He held that machines should not have rulebooks outlining everything they need to do and instead they must be taught how to learn. His work, and the emergence of the internet (with its huge stores of data) have lead to advancements in ML (Ibid, 2016).

Neural Networks are an integral part of ML because they are simulations of our human brains within computer systems. Systems set up in this way are fed large amounts of data and then they essentially work on a system of probability because they make "statements", "decisions" and "predictions" based on the data sets. After many sets of data have been analysed, the systems essentially learn which decisions are right or wrong to improve and modify their approaches in the future (Ibid, 2016).

Deep Learning is another important term to outline before discussing and assessing real-world examples of artificial intelligence. Deep Learning, or DL, is an even more specific branch of ML, because in DL data is still fed through the simulations of neural networks, however on a much more advanced and larger scale (Marr 2016)[3]. To better describe DL's relationship with ML, think of DL as the more intricate program within the simulated neural networks, within ML. Deep Learning systems are able to analyse and process data banks as large as Google's Image Library and Twitter's digital store of "tweets". DL learning focuses on the intricate programming of the neural networks within a computer (Marr 2016).

[3] This is taken from another *Forbes* article by Bernard Marr, titled, "What is the Difference Between Deep Learning, Machine Learning and AI?".

The in-depth technicalities of computer science are beyond the scope of my paper, so the most important things to take from what I have said about AI, ML, neural networks and DL are that: AI machines are machines that carry out tasks that would usually require human-level "intelligence"; ML is a sub-discipline of AI where machines are fed with large amounts of data and they learn what to do and what not to do, to improve their capabilities; neural networks are a simulation of our own human neural networks within a machine as described in the section on connectionism; and DL is a more specified and advanced version of ML, as it is able to take in and analyse more data at much faster speeds (Ibid, 2016). To sum up the relationship between these concepts, we can say that not all artificial intelligence uses machine learning but all systems using machine learning are examples of artificial intelligence and, not all systems using machine learning use deep learning but all systems using deep learning use machine learning. With all that has been said here on AI, ML, DL and neural networks it is clear to see the relevance of a connectionist theory of mind in this discussion because all of the "learning" that can be done in a connectionist system could lead a machine to understand on its own, without needing to be specifically programmed.

3.4)2. Real-World Artificial Intelligence

From our discussion thus far, it is clear to see that Searle's CRA aims to refute the claim made by Strong AI. Since the CRA's publication, much of the discussion of AI has moved past merely Weak AI and Strong AI and research has expanded. Searle's CRA is not only applicable to Strong AI that was around in the 1980s but it is relevant to learning in neural networks and even the humanoid-style robots that are found in Hollywood sci-fi films (Bishop 2002, 363).

Ray Kurzweil, a director of Engineering at Google and author of the book *How to Create a Mind*, has the ambition to build an intelligent computer. He wants an intelligent computer to be able to "understand language and then make inferences and decisions on its own" (Hof 2016). The branch of AI that deals most prominently with creating machines such as these, is the branch of deep learning (DL). Software of this type tries to simulate the activity in layers of neurons in the neocortex. This is the area of the human brain where thinking occurs (Ibid, 2016). This is the idea of machine learning mentioned previously, as "the software learns, in a

very real sense, to recognize patterns in digital representations of sounds, images, and other data" (Ibid, 2016).

This idea in deep learning, that software can mimic the neocortex's large range of neurons in our brains within an artificial neural network, is the same concept as connectionism. As mathematics and computers have improved in a way that is worthy of attention, deep learning systems are becoming more proficient in speech and image recognition.

I have a view of "understanding" as being something defined by the way in which we interact with the world around us, which is an externalist view as mentioned previously. Under this view it can be said that understanding is something that we gain from interacting with our environment. For example, a robot that interacts with its environment could possibly be compared to a child that interacts with its environment. If a child learning by interaction with the world is said to understand, then a robot could also be said to understand if it follows the same processes. Turing touches on this point in the conclusion of his 1950's paper as he hopes that machines will one day compete with humans in every intellectual field. As of the 1950s, he believed that intellectual activities like chess are the best places to test out the capabilities of machines. In 2019, computer scientists and AI-enthusiasts are aware that machines can play chess as well as do many other complex things. According to Turing the machines of the future need to be provided with the best "sense organs"[4] possible and they will need to be taught languages (Turing 1950, 460). He believes that to teach a machine a language one must follow the same language teaching processes taught to a child.

A massive feature of these connectionist, deep learning systems is that these artificial neural networks have the capacity to train themselves to identify complex patterns (Hof 2016). The way that these neural networks are trained is interesting. Programmers train neural networks to recognize an object or phoneme by rapidly sending out "digitized versions of images containing those objects or sound waves containing those phonemes" (Ibid, 2016). If the neural networks struggle to detect a certain pattern, then an algorithm within the machine adjusts the weights between the neurons, just like it is thought to happen in a human brain. This is a trial and error type of training, with the final objective to "get the network to consistently recognize

[4] I understand the term "sense organs" for a robot as referring to the state-of-the-art devices described in the robot reply. These devices are microphones, cameras, touch pads and artificial limbs that would possibly give the machines the correct causal connections to the world.

the patterns in speech or sets of images that we humans know as, say, the phoneme *d* or the image of a dog" (Ibid, 2016). This is very similar to the way that a young child learns when it interacts with the world around it. When a child starts to understand what a dog is, it does this by noticing its head shape, its body shape, its behaviour and its other features, like it is furry and it barks. This is what I argued for in Section II, a machine that learns in the same way as a child, should be seen to understand the world around it just like a child does.

The largest breakthrough in deep learning so far is in image recognition. This is an important concept to be discussed and it can be explained by continuing with the example of a machine differentiating between dogs and cats. Google has created one of the most substantial neural networks to date and this neural network boasts more than a billion connections. A computer scientist from Stanford, Professor Andrew Ng and a Google Fellow, Jeff Dean, exposed this deep learning system to images from 10million randomly selected YouTube videos (Ibid, 2016). This is how the machine is trained. It analyses and scans the videos to learn what they each contain. With regards to cats, a specific neuron within the network was stimulated to focus on cats. After training and assessing the 10million videos, the system was able to recognize cats despite the fact that no human programmer had ever defined or labelled the cats in any of the videos. Dogs and cats are quite similar animals, as they both have tails, 4 legs, they're furry, they interact with humans and they can be similar sizes. It might seem trivial for us to think that distinguishing between cats and dogs is a great feat for a computer. However, it is a great achievement because imagine a 2 year old child who has a cat as the family pet. Imagine that this child has just learnt to describe her pet as a "cat". The child has now, we assume, attached meaning to the word "cat" because she knows what a cat is on the basis of her family pet. Now if you showed this child 10 images and 6 of them showed small dogs, while 4 of them showed cats, perhaps the child would not accurately distinguish between the animals. So for a computer to accurately determine which videos had cats, rather than any other animals in them is a progressive moment for artificial intelligence.

3.5) Searle's *Chinese Gym Argument*

3.5)1. The *Chinese Gym Argument*

I will now turn to an objection to Connectionism. This objection is formulated by Searle who refers to it as a modified version of his CRA. This objection is the Chinese Gym Objection.

Searle's "Chinese Gym" thought-experiment is formulated in response to the theory of connectionism. Searle states that "connectionism is subject even on its own to a variant of the objection presented by the original Chinese room argument" (Searle 1990, 29).

As a response to the connectionist claim of Strong AI, Searle formulates his "Chinese Gym" thought-experiment. This thought-experiment is a modified version of his original CRA and instead of a room full of Chinese symbols, he requires us to imagine a gym full of monolingual, English-speaking-men. The men in this gym perform the same operations as the nodes and synapses explained in a connectionist system. Searle holds that the outcome of this process would be similar to having one man manipulate symbols corresponding to a rule book (Ibid, 29). None of the men in the gym speak a word of Chinese and the system as a whole does not have any way of learning the meanings of any Chinese words. To conclude, Searle states that the system could be adjusted as required to give the right answers to Chinese questions, like in his original CRA (Ibid, 29).

This modified CRA is aimed to deal with neural networks and parallel distributed processing. More simply put, imagine a scenario where people are in a room together passing each other plastic tokens. In this scenario, green tokens represent input along an excitatory connection and red tokens represent input along an inhibitory connection. The number of tokens moved from one person to another person in a single transaction signifies the weight of the connection like in a connectionist system. There is a list belonging to each player (like the rulebook) that indicates to whom each player must pass a token and how many tokens must be passed along. Along with each list there is also a trainer in the scenario. With the trainer, there is also a training phase of the simulation and the players in the room can adjust their lists in line with the instructions of the trainer (Copeland 2002, 116). The reason behind having a trainer in this scenario is to go with the idea that a connectionist system can learn from its environment and therefore change accordingly. This argument is comparable to Searle's response to the Brain Simulator reply with the water pipes and Searle believes that this argument helps him to refute the points made by connectionism.

According to Copeland, the gymnasium version of the CRA is made up of two inferences. The first inference is that, "no individual player understands Chinese (by virtue of doing the

computations) (Ibid, 117). The second inference is that, therefore, "the simulation as a whole – call it *G* – does not understand Chinese (by virtue of doing the computations) and the conclusion of the Chinese Gym argument is that "the network being simulated, *N*, does not understand Chinese (by virtue of doing the computations) (Ibid, 117).

3.5)2. Responses to the Argument

In response to the Chinese Gym objection, Copeland argues that we can look at this thought-experiment from a logical point of view and say that we can agree with Searle that handing around tokens and fiddling with lists or recipients will not show that the individual players understand Chinese. However, even with this agreement with Searle, it does not mean that we must also agree that the system as a whole does not come to understand Chinese. Copeland argues that the fallacy in this scenario moving from part to whole is even more clear than in the original CRA (Ibid, 116). This response is the same as the reply formulated in the Systems Reply that was outlined in section II. So, it is quite clear that this can be tackled by the Systems Reply. With regards to the inferences mentioned above, we can agree with the first inference but not the second inference. Like the response of the Systems reply, just because each player in the gym does not understand Chinese, it does entail that the whole system (*G*) does not come to understand Chinese. Stevan Harnad, in his paper, *Minds. Machines, and Searle 2: What's Right and Wrong about the Chinese Room Argument*, argues that the Systems Reply is correct because it outlines an executing program, which is part of what an understanding 'system' would be like (Harnad 2002, 295).

To bulk up the defence of Strong AI and link it to connectionism, we can look at a position known as *Connectionist Strong AI*. Perhaps this modified version of connectionism is sufficient to avoid Searle's criticism in his modified CRA. Connectionist Strong AI is the idea that, "an appropriately configured and trained connectionist network would have (the relevant) genuine psychological properties, and would do so purely in virtue of its having the configuration training in question" (Preston 2002, 23). In addition to this version of connectionism, in *Computers, Dynamical Systems, and Searle*, Michael Wheeler turns to another variant of connectionism known as *dynamical neural networks* (Wheeler 2002, 349). These dynamical neural networks "feature properties not found in their conventional computational relations, properties such as asynchronous processing, real-valued time delays on connections, non-

uniform activation functions, deliberately introduced noise, and connectivity which is not only both directionally unrestricted and highly recurrent, but also not subject to symmetry constraints' (Ibid, 349). What this all means, is that these neural networks display properties that are considered non-computational. Neural networks structured in this way contain the same architectural properties and dynamical complexities as real biological nervous systems (Ibid, 349). As has already been outlined, connectionism could be regarded as a more "biologically plausible" theory of mind and so, if research such as mentioned above could prove that these systems are as real as biological nervous systems, then an entity using such a system could have mental states.

3.6) Conclusion

In this section of my paper, I have critically engaged with the theory of mind known as connectionism, the idea of deep learning and machine learning, and a theory of understanding. I began this section by outlining an account of understanding. I then moved onto compare and contrast connectionism and computationalism. After this, I discussed machine learning, deep learning and certain forms of real-world AI. From there I analysed an objection to connectionism known as "The Chinese Gym". I argued that Searle's "Chinese Gym Argument" is not sufficient to rebut connectionism and so, I now believe that throughtout this paper and each section, I have provided sufficient evidence to show why Searle's CRA is not substantial enough to deny Strong AI and future machines the ability to exhibit cognition.

Research Paper Conclusion

The central thesis that I pursued in this paper, is that it is possible for a machine to be ascribed mental states if it has the correct causal connections to its environment. I did this by looking closely at the theory of functionalism, critically analysing and arguing against John Searle's *Chinese Room* Thought Experiment, analysing accounts of understanding and examining the connectionist theory of mind. By refuting Searle's argument, I provided a discussion and argument of what a machine that exhibits cognition would be.

In Section I of my paper, I looked at functionalism because it provided support for the idea that machines can be attributed mental states. By closely analysing intentional states (such as beliefs and desires) and the theory of functionalism, I answered the question: "What are mental states?". There were a few objections to functionalism that were raised in this section, such as the absent qualitative states (or "qualia") argument and Ned Block's "Homunculi-Headed Argument". I believe that in this section I provided enough evidence to show how and why such objections should not be a problem for functionalist accounts of cognitive states and the idea that machines can be ascribed mental states. This lead the discussion on to Searle's CRA.

The next section focused primarily on Searle's *Chinese Room* Argument (CRA), the most well-known replies to the CRA, Searle's responses to those replies and further comments on Searle's remarks. Searle's CRA is directed against *Strong AI*, which as mentioned, is the notion that a "computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states" (Searle 1980, 417). I highlighted how Searle's argument against Strong AI does not show that machines cannot have mental states. I did this by critically engaging with the most well-known replies to the thought experiment. The replies that I analysed are *The Systems Reply*, *The Virtual Minds Reply*, *The Robot Reply* and *The Brain Simulator Reply*. I systematically analysed each of these replies, discussed Searle's responses to each reply and provided further arguments for the replies in light of Searle's remarks. To argue for my thesis, I showed that even if it each reply on its own was not sufficient to refute Searle, the replies taken accumulatively could possibly refute the CRA. This accumulation of the replies was discussed when I concluded the section with *The Combination Reply*. The arguments made in this section defused Searle's CRA to a certain extent but further

elements needed to be added to the discussion to make the argument against Searle more persuasive.

The third and final section of my paper focused on adding more elements to the argument against Searle, such as accounts of understanding, two theories of mind (connectionism and computationalism), machine learning, deep learning, real-world AI and an objection to connectionism. I began this section by comparing two accounts of understanding, the internalist account and the externalist account. I held that the externalist account is a more plausible account to follow because to understand, one needs the correct causal connections to the external world. I then moved onto compare and contrast connectionism and computationalism. I held that connectionism is the more plausible theory of mind to follow because it offers us a system that can learn from experience, rather than a system with every step programmed (like a system under the classical theory of mind). After this, I discussed machine learning, deep learning and certain forms of real-world AI to show how a connectionist system and neural nets are relevant. Next, I analysed an objection to connectionism known as "The Chinese Gym". I argued that Searle's "Chinese Gym Argument" is not sufficient against connectionism because this modified version does not stand up against the points made by *The Systems Reply*. These elements were all added to make the argument against Searle's CRA more convincing.

I now believe that throughout this paper and each section, I have provided sufficient evidence to show why Searle's CRA is not substantial enough to deny Strong AI and future machines the ability to exhibit cognition.

Bibliography

Aydede, Murat. 2015. "The Language of Thought Hypothesis." *The Stanford Encyclopedia of Philosophy.* September 21. Accessed May 16, 2018. https://plato.stanford.edu/entries/language-thought/.

Bickle, John. 2016. "Multiple Realizability ." *The Stanford Encyclopedia of Philosophy.* March 21. Accessed April 29, 2018. https://plato.stanford.edu/entries/multiple-realizability/#WhaMulRea.

Bishop, Mark. 2002. "Dancing with Pixies." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop, 360-378. Oxford: Oxford University Press.

Block, Ned. 1978. *Troubles with Functionalism.* Vol. 9, in *Perception and Cognition. Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science*, edited by C W Savage, 261-325. Minneapolis: University of Minnesota Press.

Boden, Margaret. 1987. "Escaping the Chinese Room." *Cognitive Science Research Paper* 1 - 25.

Brown, Curtis. 2016. "Narrow Mental Content." *The Stanford Encyclopedia of Philosophy.* June 21. Accessed January 30, 2019. https://plato.stanford.edu/entries/content-narrow/#put.

Cole, David. 1991. "Artificial Intelligence and Personal Identity." *Synthese* (Springer) 88 (3): 399-417.

Cole, David. 2015. "The Chinese Room Argument." *The Stanford Encyclopedia of Philosophy.* December 21. Accessed February 28, 2018. https://plato.stanford.edu/archives/win2015/entries/chinese-room/.

Copeland, B Jack. 2002. "The Chinese Room - A Logical Point of View." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, 109-122. Oxford: Oxford University Press.

Dennett, Daniel. 1981. "True Believers: The Intentional Strategy & Why it Works." *Scientific Explanation* 13-35.

Dennett, Daniel. 1988. "Précis of the Intentional Stance." *Behavioral and Brain Sciences* 495-546.

French, Robert M. 2000. "The Turing Test: the first 50 years." *Trends in Cognitive Sciences* 4 (3): 115-122.

Garson, James. 2016. *Connectionism.* December 21. Accessed March 2, 2018. https://plato.stanford.edu/entries/connectionism/.

Graham, George. 2017. "Behaviorism." *The Stanford Encyclopedia of Philosophy* . March 21. Accessed January 31, 2019. https://plato.stanford.edu/entries/behaviorism/#1.

Harnad, Stevan. 2002. "Minds, Machines, and Searle 2: What's Right and Wrong about the Chinese Room Argument." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, 294-307. Oxford: Oxford University Press.

Hof, Robert D. 2016. "Deep Learning." *MIT Technology Review.* March 29. Accessed May 19, 2018. https://www.technologyreview.com/s/513696/deep-learning/.

Horgan, Terence. 1997. "Connectionism and the Philosophical Foundations of Cognitive Science." *Metaphilosophy* 1-30.

Hyslop, Alec. 2016. "Other Minds." *The Stanford Encyclopedia of Philosophy.* March 21. Accessed April 29, 2018. https://plato.stanford.edu/entries/other-minds/.

Jacob, Pierre. 2014. *Intentionality.* December 21. Accessed March 2, 2018. https://plato.stanford.edu/entries/intentionality/.

Lau, Joe, and Max Deutsch. 2016. "Externalism about Mental Content." *The Stanford Encyclopedia of Philosophy.* December 21. Accessed January 30, 2019. https://plato.stanford.edu/entries/content-externalism/#ExtMenCau.

Leon, Mark. 1998. "The Unnaturalness of the Mental: The Status of Folk Psychology." *The Southern Journal of Philosophy* 367-392.

Levin, Janet. 2016. "Functionalism." *The Stanford Encyclopedia of Philosophy.* December 21. Accessed February 25, 2018. https://plato.stanford.edu/archives/win2016/entries/functionalism/.

Marr, Bernard. 2016. "What Is The Difference Between Artificial Intelligence And Machine Learning?" *Forbes.* December 06. Accessed February 06, 2019. https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#6e0445682742.

Nadella, Satya. 2017. *Hit Refresh.* London: William Collins.

Oppy, Graham, and David Dowe. 2018. "The Turing Test." *Stanford Encyclopedia of Philosophy* . March 21. Accessed April 22, 2018. https://plato.stanford.edu/entries/turing-test/.

Preston, John. 2002. "Introduction." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop , 1-50. Oxford: Oxford University Press.

Proudfoot, Diane. 2002. "Wittgenstein's Anticipation of the Chinese Room." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by John Preston and Mark Bishop, 167-179. Oxford: Oxford University Press.

Putnam, Hilary. 1975. "The Meaning of Meaning." In *Philosophical Papers: Mind, Language and Reality* , 215-271. Cambridge: Cambridge University Press.

Rescorla, Michael. 2017. "The Computational Theory of Mind." *The Stanford Encyclopedia of Philosophy* . March 21. Accessed April 22, 2018. https://plato.stanford.edu/entries/computational-mind/.

Rey, Georges. 1986. "What's Really Going on in Searle's "Chinese Room"." *Philosophical Studies* 169-185.

Schwitzgebel, Eric. 2015. "Belief." *The Stanford Encyclopedia of Philosophy.* June 21. Accessed February 25, 2019. https://plato.stanford.edu/entries/belief/#1.4.

Searle, John R. 1987. "Minds and Brains Without Programs." In *Mindwaves*, edited by C Blakemore and S Greenfield, 209-233. Oxford: Blackwell.

Searle, John R. 1990. "Is the Brain's Mind a Computer Program?" *Scientific American* 20-5.

Searle, John R. 1992. *The Rediscovery of the Mind.* Cambridge, Massachusetts: The MIT Press.

Searle, John R. 1980. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences* 417-457.

—. 1984. *Minds, Brains and Science.* Cambridge, Massachusetts: Harvard University Press.

—. 1997. *The Mystery of Consciousness.* New York: The New York Review of Books.

—. 2016. "What Is The Difference Between Deep Learning, Machine Learning and AI?" *Forbes.* December 08. Accessed February 06, 2019. https://www.forbes.com/sites/bernardmarr/2016/12/08/what-is-the-difference-between-deep-learning-machine-learning-and-ai/#1bf3dd7526cf.

Steup, Matthias. 2018. "Epistemology." *The Stanford Encyclopedia of Philosophy.* December 21. Accessed March 4, 2019. https://plato.stanford.edu/entries/epistemology/#IVE.

Thagard, Paul. 1990. "Philosophy and Machine Learning." *Canadian Journal of Philosophy* 261-276.

Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind: A Quarterly Review of Psychology and Philosophy* 433-460.

Tye, Michael. 2017. "Qualia." *The Stanford Encyclopedia of Philosophy.* December 21. Accessed April 27, 2018. https://plato.stanford.edu/entries/qualia/#Functional.

Wheeler, Michael. 2002. "Computers, Dynamical Systems, and Searle." In *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, 338-359. Oxford: Oxford University Press .