MASTERS BY RESEARCH

On the equivalence of brains and machines
consistency and determinism in first-order logic machines

Prosser, Thomas

Award date:
2019

Awarding institution:
Bangor University

ON THE EQUIVALENCE OF BRAINS AND MACHINES: CONSISTENCY AND DETERMINISM IN FIRST-ORDER LOGIC MACHINES

Thomas V. Prosser

A thesis submitted for the degree of Master of Arts by Research

University of Bangor
2019

Contents

Abstract

The mechanist thesis and the brain as a Turing machine is shown to be a poor model of mind due to its idealisation concerning memory length and precision algorithms. While the Gödelian argument demonstrates the the brain not to be of first-order arithmetic in the language of PA, this is not all which may be concluded, meaning the Gödelian argument stands, yet requires extension. We shall demonstrate that the Chinese room cannot be intelligence since it lacks the required properties concerning sub-symbolic networking and representation. Both the Gödelian argument and the Chinese room operate on the intention of proving the limitations of formal systems, they architecture as deterministic systems raises a problem of free-will. The alternative to the mechanist thesis and the computational theory of mind are addressed here. Due to the far reaching scope of mechanism, our purpose is to remain focused on the grounding works of mechanism from formal systems of logic and the equivalence of minds and these formal systems. It is determined that equivalent systems which are described in terms of a formal system lead one to determinism. This is explored under a direct, yet expansive set of philosophic and mathematical works.

Notation on symbols

The symbolic notation throughout is as consistent as possible, but in dealing with multiple authors there are interruptions which may lead to confusion, particularly surrounding lettered functions.

We use basic symbols of logic/mathematical notation, and have choose to remain as in keeping with referenced notion for the sake of readability and consistency. Any negation shall be given as $\sim$, and $\subset, \supset$ where $\subset$ shall always denote subset. Where $\subset$ is given as 'negation' by any author it will be altered to $\sim$. Where certain notations which introduce new, or prior used symbols, it hall be clearly stated what they are - "ϕ, where ϕ is a formula of F", etc, - so as to minimise confusion. he ϕ notation is commonly used throughout the literature and we have used it to refer to some knowledge which could be known by our system. The use of ϕ is common towards the latter sections of this paper, appearing multiple times in the works of Reinhardt (Reinhardt, 1980), Carlson (Carlson, 1999), Alexander (Alexander, 2006) and Shapiro (Shapiro, 2003), in addition to those on Penrose's works (Penrose, 1994).

Concerning the equivalence of brains and machines - *B, M* - the notion of Turing equivalence is followed where it is meant that B is the same a M if and only if both have the same properties such that what can be expressed in B can also be expressed in M such that they are identical. This notation is commonly given as $P \leftrightarrow Q$ but we alter this notation to $B \leftrightarrow M$, since we refer to a specific type of system.

There is consistency with these however; Alexander (Alexander, 2003) is based on the work of Carlson and Reinhardt, and where we refer to specific authors notation is given. We do not deviate from the standard by much and every attempt to remain as readable as possible has been given.

The symbols of *S, G,* and F afford confusion with F in particular denoting a number of functions, each individual to the author. Much of our symbolism will be taken from automata theory, often directly from Hopcroft, Motwani, Ullman (Hopcroft, Motwani, Ullman, 2007). Beyond this, general symbolism is used such as where S is defined as denoting a *step* of the machines function; $S, S_1, S_2, ...$ and F as being a function, with the *accepting state* being some

function. The one benefit of F notations is that it commonly denotes any given function, or where it denotes a formal system, it is easy to see which it refers to due to the informal description or its placement in the logic with relation to the logics pre-defined respective symbols.

Due to the use of 'G' for the Gödel sentence, and for grammar, there is a clear overlap. Spavrek (Spavrek, 2007) uses G to be a statement of Searle (Searle, 1980), which shall be discussed in §4. Since the Gödelian sentence shall not be discussed in chapter concerning Searle, in addition to Spavrek's comments being something I highlight, it is not necessary to alter this notation as it will always be clear what is being referred to, however Spavrek's notation will be written as (G).

Grammar will not come into play with the mathematics since mathematical grammar will be given as the rules of formation, and all sentences of the system are given as theorems and denoted by Φ, for example. In dealing with automata we will not speak of the Gödel sentence and G will refer to grammar. In all instances, we shall specify 'where G is grammar', or 'where G is the Gödel sentence'. Grammar, G, is utilised in §5.2 where there is no mention of the Gödel sentence.

For language, there is no overlap concerning L with it being almost universal across the literature. Chomsky[1] used F to denote function, which is also used to denote a formal system. Formal systems we denote as being 'F'. This standard font notation is used and the use between this and the italicised 'F' is more of an author preference than a general notation. All instances of 'F' for formal systems will be given as 'F' Chomsky (1959)used F to denote 'function', which we alter to `F` since F is so frequently used to denote a formal system. All quotes from Chomsky (Chomsky, 1959) will use the same altered `F`.

[1] Across all literatures.

Chapter 1

Introduction

What are the fundamental capabilities and limitations of computers?

- Michael Sipser

1.1 Outline

It is well-established that the mechanist thesis of brain as a Turing machine is problematic. The two primary reactionary arguments which are addressed are J. R. Searle's *Chinese Room argument* (Searle, 1980) and his claim that syntax is not sufficient for semantics, and what has come to be known as the Lucas-Penrose argument (henceforth the L-P result), formed independently by J. R Lucas (Lucas, 1961) and R. Penrose (Penrose, 1989). These two arguments differ in their approach and respective fields, yet they stand as two of the most concise - and highly referenced - arguments against artificial intelligence. Critically, both concern some program and a brain/machine equivalence; $B \leftrightarrow M$.

This research concerns a number of extensions upon Gödel's *incompleteness theorem* (Gödel, 1931) and the brains/machines equivalence as noted by Bringsjord (Bringsjord, 2012) and explored further in A. Freidman (Freidman, 2002) and J. Weng (Weng, 2015). Broadly, it is on the mechanist versus anti-mechanist claims toward the mind as a machine: what is commonly referred to as the *problem of mechanism*. The foundation of philosophic - and indeed much of - artificial intelligence, is the claim that human behaviour can be defined with such precision that a machine can simulated.

Within anthropic mechanism, in relation to artificial intelligence, the debate between

mechanist and anti-mechanist is extensive, yet it is often centralised around two point: the mind and consciousness, and free-will. What I discuss here shall not concern consciousness, beyond where relevant, instead choosing to focus the free-will argument since in concerns itself to a greater degree with logic and formal systems.

Mechanism highlights a deeper problem within any theory of the brain as a machine/formal system which in turn causes faults to arise in the anti-mechanist claims. Despite the level of research which has been given, the mechanist thesis is simple and the most fundamental criticisms were given at the start of mechanisms life, so to speak. J. R. Lucas (Lucas, 1961) gave response to Rogers (Rogers, 1957) who introduced early issues of consistency in the Gödelian arguments, and to Putnam (Putnam, 1960) who proposed the initial mechanist/anti-mechanist argument built on the early computational theory of mind. Much of Lucas's response to Putnam is from private conversation as Lucas states *Minds, Brains, and Gödel.*

The Chinese room concerns a program running a code for natural language conversation - an adaptation on the Turing test - where a man manipulates symbols of a language he does not understand in accordance with a rule. The purpose is to illustrate how semantics cannot come from the bare syntax and instead requires a brain. Both Lucas and Penrose concern brain not being a Turing machine of some sort since the brain can recognise the Gödel sentence - a true formula which is not provable in a consistent system - as being true. There are notable differences between Lucas and Penrose, with Penrose provided an expanded version and an alternative theory of consciousness. The Chinese room and the Gödelian argument are not too dissimilar since both utilise some form of the Church-Turing thesis and a formal system.

Despite the large difference between the two arguments of Searle's pure philosophy versus the mathematics of the Lucas/Penrose argument, they both concern the same thing; the mind as a Turing machine or digital computer of some equivalence. Into his formal versions, Searle takes use of the Church-Turing thesis, bringing his result more in line with the L-P result (Searle, 1990b). This is thereby an attack on computationalism, and by extension,

functionalism. Searle's argument is a direct attack on both - something which he does so willingly - by both terms are viewed as being synonymous by many with both being used. 'Computationalism' is used more frequently since it is the prevailing term and argument. It also bears strong semantic relation to the study.

The difference between Searle and the L-P result, often given as the Gödelian argument, concern a Turing machine. This is what shall be referred to as the Turing model (TM model) where the brain is seen as a Turing machine. There clearly is a large distinction between language-based systems versus mathematics-based systems. The intention is certainly not to present a unifying theory between the two, but only to explore the limits of certain AI systems. Critically, there is one flaw in Searle's argument which has gone strongly unchallenged, and that is his use of the Church-Turing thesis (Searle, 1990b). Lucas makes no reference to the Church-Turing thesis (Lucas, 1961), while Penrose gives extended note to it (Penrose, 1989: Penrose, 1994).

It would not be possible to refer to all AI systems and regardless of the logic used there will always be a system cable of doing, or not doing, something since all that is required is that one alters the specifics or places focus on one specific task. This technique is widely employed in much of standard AI research.

I shall not wholly disagree with the L-P result concerning Gödel's theorem, only say that it is problematic as many have noted in the decades following their respective publications.

At the center of both results is a theory of mind. Lucas never produced a theory of mind, nor is his Gödelian argument as extensive as Penrose's, or Searle's. Additionally, Lucas and Penrose never collaborated on a paper. Both Searle and Penrose have a theory of mind which alters how their respective arguments can be viewed. As such this adds a level of complexity as it often distances us from the central argument against AI.

What I propose here is a shift on the conventional method of analysis of the brains as Turing machines thesis. Both Shapiro (Shapiro, 2003) and Lindström (Lindström, 2002) felt that that question of mechanism is perhaps not adequately given specificity. What shall be proposed is the brain as *form* of some finite automata based on connectionist networks.

This is nothing new; McCulloch and Pitts (McCulloch & Pitts, 1943) explored this in its earliest form and a vast level of connectionist research has demonstrated the potential for mapping the brain and its functions. This is however where the more traditional philosophic response enters the situation with a focus towards consciousness and a definition between capacity. This is where it makes sense to begin problem of meaning and challenge J.R. Searle's Chinese room thought-experiment (Searle, 1980) by demonstrating how systems are able to learn and build relations between words and concepts and physical inputs. Searle's argument includes a great deal of thought on the nature of consciousness and speaks of a distinction between the abilities of brains and machines, including those which simulate brains and their activities.

What shall be discussed is that any formal system of an with an output, Σ^*, from a finite alphabet, Σ, the totality of output symbols and initial states, q is finite thereby making all possible states, q_n, thereby making the systems output states knowable from some algorithm. Lucas (Lucas, 1961) highlights this as a problem of free-will versus determinism in his paper[1].

I am by no-means the first to highlight this problem and indeed all critiques of mechanism have highlighted this problem in some manner. This critique is not exclusive to the anti-mechanism (see Putnam, 1960; Lucas, 1961; Searle, 1980; Penrose, 1989; Searle, 1990; Penrose, 1994.) frequently being contained in those who engage with the possibility of a machines/brains equivalence or the brain as a machine of some sort. I do wish to briefly note that the scope of mechanisms and its problems encompass just about every aspect of mechanism and as such it is impossible to give a concise critique of mechanism. Our 'problems' will be restricted to those concerning the equivalence of brains and machines.

With Chapter 2, Searle's thought experiment of the *Chinese room* shall be introduced and the broad issues shall be outlined. The purpose here to deal with and present Searle's main argument concerning the replication of intelligent action and if there is a link between imitating an intelligent action, and actually performing the intelligent action. The use of pure rule-following is problematic and despite Searle's frequent mentions of the rule-following

[1]This presents a significant opposition to formal systems: See §6.5.1.

operations of the room being inefficient, the dismissal of it because the argument is only an analogy is difficult to rectify due to the limitation that are placed on the Chinese room system. The informal version encompasses a great deal of what was contemporary philosophic literature and Searle provides a series of defences towards the most frequent objects placed against the Chinese room.

Chapter 3 introduces Searle's formal version of the argument which allows for a more precise critique. Much that was discussed in chapter 2 remains here, and direct response is given to each of Searle's axioms and subsymbolic computation is introduced. Subsymbolic computation provides a direct counter towards Searle's foundational argument and demonstrates how the Chinese room deals with symbol processing at too high of a level. By grounding axioms and concepts into a single entities as opposed to an axioms of sub-axioms, a network of relations is bypassed which leads to a poor representation of information.

Chapter 4 continues with symbolic meaning in a formal manner focusing on formal systems and mathematical works, particularly Chomsky's work concerning grammar. From this, a formal variant of the Chinese room system is developed (§5.2) where its alphabet and function are discusses in a formal manner in line with Chomsky automata and grammar before the problems of the Chinese room are fully explored in chapter 5, with specific reference to Searle's misuse of the Church-Turing thesis and symbol manipulation as thinking. Here we are able to demonstrate the link between Searle's system and the L-P result.

Our final chapter deals purely with the brain as a Turing machine and the L-P result from Gödel's incompleteness theorems and discusses the equivalence of brains and machines.

My intent within this thesis is not merely to provide an updated reply to the L-P result, and Searle's Chinese room with respect to all recent developments in the field of computer science, but to address the claim of brains as Turing machines and further explore the problem of free-will; a response which has not often been fully given towards Lucas and Penrose. The free-will problem seems detached from intelligent systems, yeah informally the power to choose between alternatives in pre-determined systems leads one to determinism. Paradoxically, it is difficult to escape from determinism and in-determinism when dealing

with finite states.

It is careless to proceed boldly down this line of reasoning as Lucas, Penrose, and to a lesser extent, Searle have done since the Gödel and Turing results can easy result in paradoxical conclusions which applied to rigorously to brains, particularly when one utilises a significant deal of idealisation on the part of the brain. The significant failure of the anti-mechanism arguments is that they attack a faulty argument; mechanism is poorly defined.

Ultimately, the L-P result is not all which may be concluded from the Gödelian theorems since it is problematic to assert the consistency of brains, which is only part of the problem. A Turing machine is a poor model upon which to base a brain network since brains are finite, and the poorly defined nature of the mechanist model requires a specific form of the computational theory of mind. What is concluded here concerns three propositions of the brain as (i) a Turing machine, (ii) not directly a Turing machine, and (iii) a Turing machine being able to directly simulate a brain. The discussion prior to these shall provide a clear exploration of each of these such that each propositions response is direct.

1.2 The Dartmouth Proposal

Established in 1955, the proposal states:

> "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves." (McCarthy, Minsky, Rochester, Shannon. 1955: p. 1)

A concise definition of the outputs of both brains and machines is severely lacking in much of the literature from both sides. The Dartmouth Proposal discussed automatic computers, computers using language, neuron nets, the theory of the size of a calculation, among others. Each is highly significant yet none individually form the mechanist argument. Instead mechanism encapsulates all which the Proposal discusses with primary focus that intelligence can

be “so precisely described that a machine can be made to simulate it" (McCarthy, Minsky, Rochester, Shannon. 1955: p. 1) The proposal was left open to give the field as much room to grow as possible, but subsequent research and publications have each dealt with broad definitions and idealisation of the machine with much built on the Church-Turing thesis. This has ultimately become the mechanist thesis. The collective research of each individually, published as independent research is of great significance towards logic-based AI, but I don't wish for the proposal to be seen as a significant to this thesis.

The Dartmouth Proposal invariably posses the question concerning the limits of computer science. Searle (Searle, 1980) challenged this with a thought experiment that acts to demonstrate that computers cannot display consciousness since they lack the correct mental properties, and J. R. Lucas (Lucas, 1961) and R. Penrose (Penrose, 1989; Penrose, 1994) take an alternate route by exposing the limits of formal systems by holding brains to be of a higher power; each exposes a limitation which machines inherently possess,

The Gödelian argument and the Chinese room contain a very clear similarity which - while I'm sure has been noticed - has not be presented an any manner which concisely links the two. Penrose (Penrose, 1994) wrote on the Chinese room, yet perhaps because the reasoning is that Searle makes no comment towards it. What this concerns is that Searle fails to make a single very precise point: there must be a limit to the Chinese room. The point of the Gödelian argument is not what the system can do, but what it cannot and *finding* that limits; Searle does not seek to find that limit. I shall not elaborate further here, but will detail this fully in §6, once both arguments have been covered in adequate detail.

It is clear that mechanism is problematic on both sides, and the broader point that the brain is a Turing machine is of equal dispute due to the idealisation of a Turing machine and of *effective computation* (Siegelmann, 1995). The reason for this is difficult to explain collectively since there are certain aspects of the brain as a Turing machine argument which function perfectly well (Weng, 2015) where as others avoid clear discussion on the function of the machine and deal only with a machine which follows an algorithm (Searle, 1980; Searle, 1990a; Searle, 1990b.)

If we consider both Gödel's incompleteness theorem and Tarski's theory of formal truth then we have an argument against the limitations of a formal system. That shall not be disputed, but the L-P result has failed to convince a vast number of academics, and it is these failings and limitations of both Lucas and Penrose which I wish to highlight. MeGill (MeGill, 2004) posits that we may be paraconsistent and Putnam (Putnam, 1960) challenged the anti-mechanist claims by writing that the argument ignores the issue of consistency in brains. LaForte et al (LaForte, Hayes, Ford, 1998) issued this against Penrose, writing that this belief ad the idealisation of human mathematics constitutes an academic hubris.

1.3 The Theory of Computation

The theory of computation began life in the 1930's from the works of Alonzo Church, Kurt Gödel, Stephen Kleene, and Alan Turing. In nature, it is purely mathematical with little consideration for philosophy. Despite this, Aaronson (Aaronson, 2011) claims that the purpose of their works was the clarification of philosophical issues. There is some argument to be made here that this is partly true due to the relation to metamathematics which the works hold, despite both Church and Kleene being concerned little with philosophy.

While having its origins in mathematical logic, and based on some formal system within first-order logic, computability finds itself seeping into all aspects of AI philosophy; from Bayesian mechanisms, probability reasoning in complex systems, right up to human co-habitation, and co-evolution. These systems have since extended beyond mathematical logic and into linguistics and formal grammar, most notably with Chomsky's work in formal grammar (Chomsky, 1956; Chomsky, 1957) Chomsky's two early works of formal grammar mark a notable shift from the more traditional philosophy of linguistics/language into a more structured and logical manner. There remains a notable difference between Chomsky's work and the works of Frege and Tarski since they remained sceptical towards the formalisation of natural language and remained in the development of formal language for formalised parts of language[2].

[2]See Tarski, 1936; *The Concept of Truth in Formalized Languages.*

Turing's work related more to computer science and his hobby of cryptography, while Church's and Gödel's was mathematical. For nearly three decades it remained that way, and only once computers became more widely used in the 1960's computability began to expand. All problems we wish to solve are generally computable in the 'Turing sense'. That is that there is some algorithm by which we can use. This all may cause one to ask where mechanism fits into all this; it is the collection of the theory of computation.

1.4 Hilbert's Problems of Mathematics

In 1900, David Hilbert presented his 23 questions. The second of these was the proof that the axioms of arithmetic are consistent. Emil Post explored the halting problem for tag systems further [3]. It's unsolvability was shown by Minsky (Minsky, 1967). In 1931, a year following the completion of his PhD, Gödel published *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*; his incompleteness theorems and in 1936 Alfred Tarski published *The Concept of Truth in Formalised Languages*. Church developed λ-calculus and showed Hilbert's *entscheidungsproblem* to be unsolvable [4]. This lead to something of a stalemate between Church and Gödel with the later rejecting Church's thesis [5]. This was ultimately ended in 1936 with Turing's publication of *On computable numbers, with an application to the entscheidungsproblem.*

1.5 Alan Turing: *Tests, Machines, and Symbol Manipulation*

Turing's 1950 paper *Computing Machinery and Intelligence* first established the now famous Turing Test. The design of which was such to demonstrate if a machine could exhibit intelligent behaviour when under human examination. Turing's paper is not a theory, but an exploration of the potential of an AI system and a small piece of what would become the Dartmouth proposal. The test does not serve as a method to show machines can think,

[3] See E. Post. *Absolutely unsolvable problems and relatively undecidable propositions-account of an anticipation. Pp. 375-441.*

[4] *A Note on the Entscheidungsproblem* in Davis

[5] See Kleene [1936] and Gödel [1934]

or even present a framework upon which one may develop such a machine. As we have prior discussed, providing an accurate definition of both 'thinking' and 'intelligent' is doable, albeit with some degree of difficulty, but it becomes more complex when we apply it to machines. Turing's note of these problems is clearly given in paper (Turing, 1950).

D. Hofstadter has long abandoned the AI community due to problems which have arisen from this debate along with others; mainly the overall direction of the study. His long-held belief is that systems ought to teach us something about ourselves and human intelligence, yet modern AI systems are built purely on data, following a strain similar to Turing's proposed system for a Turing test. Systems such as *Siri*, *Alexa*, and IBM's *Watson* and *DeepBlue* are not 'intelligent' since they operate in a simple manner; they are built upon data. *Siri*, *Google*, and *Alexa* are all marketed as being advanced AI's but they again work on data; voice-recognition software allows the device to perform a search of sorted or accessible data, and then text speak informs the person of the result. Naturally, these speak with a pleasant and human-like tone to make consumers respond better towards the systems.

DeepBlue, while capable of the completion of what is regarded as a highly intelligent action in humans, was not better than Kasparov, or even 'good' at chess; it won by brute force. Where *DeepBlue* differs from *AlphaGo* and *AlphaZero* is in the brute force method. Both *AlphaGo* and *Zero* made use of this method, they had the additional component of a learning algorithm which played a lot more tactically, with *Go* making a number of poor moves during Game 4, which was viewed as someone against the program by its developers, and Zero learning by playing the game of Go thousands of time to essentially learn without being programmed how to play Go. *DeepBlue* is and its method are a form of imitation since the system could not actually 'play' chess; it analyses every possible move of both black and white and when with the option with the highest probability of taking pieces and winning.

The Turing Test is similar; it's not concerned with intelligence, but with a seamless replication of a particular human behaviour; natural language conversations. Turing was aware of this, even referring to his test as an 'imitation game'. Much of the criticism and arguments against were anticipated by Turing or not the subject of his paper .

The Turing test introduced the academic community to the concept of a 'thinking' machine and his own predictions gave some 50-years before a machine could pass the Turing test, with 'thinking' - of some clearly defined term - being much beyond this. Of the question Turing posed; *can machines think?*, Turing wrote: "The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous". (Turing, 1950, p. 40) 'Think' is a term he referred to as being an ambiguous one, and sought to replace it with a more categorical term; hence *imitation*.

It does therefore seem odd that so many have placed the Turing test so heavily under the spotlight. As with Hofstadter's overriding argument that AI ought to tell us something about ourselves and human intelligence the test is not about intelligence but imitation. This disillusionment with the state of current AI philosophy and research has caused him to abandon the field.

By 'imitation' we mean something different from 'replication'. This is certainly difficult to define in a constructive manner, since there would be no reason to see them as something different, yet we shall view them separately as it will be required in §2. As such the technique is based primarily on achieving the task by any means even if it means cheating. With replication, it is generally meant that it is a more scientific production of the results based on developing the system around the brains neural network, or intuition from prior knowledge. Where this contains relevance is that the systems proposed by Searle, Lucas, and Penrose follow a similar patter; they take use of data and follow a rule. This is less precise in the Gödelian versions since the focus is shifted away from human interaction towards the limitations of a system,.

1.5.1 Passing the test

Passing the test is not easy and still has yet to have been done so as to appease a significant portion the academic community. In 2014 - the 60th anniversary of Turing's death - a Chatterbot by the name of Eugene Goostman was deemed to have passed by the event organiser Dr Kevin Warwick. The program was designed to be a 13-year old boy from Ukraine. This was intentionally done so as to allow for gaps in knowledge, grammatical

errors and the nature of conversation. Often he would appear enthusiastic about speaking, ask questions about the other person, and frequently used emoji's, and/or text-style speak. This further served the purpose of misdirection.

Of interesting note, the system appeared to have been trying hard to mask itself. The enthusiastic nature is rather distracting and to those over the age of eighteen, it may cause them to being doubting everything said and cause them to believe the program to be a program. It is difficult to know where the line is between youthful neglect of grammar and care meets going too far, and where the happy middle-ground exists between this and being too clever and precise.

Programs of this nature are not the most sophisticated but when ran on dedicated servers such as in the case of Goostman where the server was only dealing with a select group of judges versus more general online Chatbots which have to contend with thousands of users at any given second.

Goostman's low pass rate is not convincing to suggest it has passed the Turing test, and many of these bots often breakdown on select questions:

> Scott: Which is bigger, a shoebox or Mount Everest?
>
> Eugene: I can't make a choice right now. I should think it out later. And I forgot to ask you where you are from. . .
>
> Scott: How many legs does a camel have?
>
> Eugene: Something between 2 and 4. Maybe, three? :-))) By the way, I still donk͡now your speciality - or, possibly, I've missed it?
>
> Scott: How many legs does a millipede have?
>
> Eugene: Just two, but Chernobyl mutants may have up two five. I know you are supposed to trick me.[6]

[6]From Aaronson, S. *My Conversation with "Eugene Goostman," the Chatbot that's All Over the News for Allegedly Passing the Turing Test.* https://www.scottaaronson.com/blog/?p=1858

It would seem obvious that a system would be able to provide answers to these sorts of questions since they are easily verifiable, yet at the same time Goostman could not provide an exact answer to the height of Everest since it would not be believable that a 13-year old would know that.

This natural language conversation is what Turing envisaged as the purpose of the test; as we shall see shortly, Searle has misunderstood the point of this test greatly. Turing had no intention of producing a framework for artificial general intelligence[7].

We have mentioned consistency and how it has been proposed that humans are inconsistency, as counter to the L-P model, and Goostman is inconsistent; it must be since 'he' is a 13-year old boy. If the Gödel sentence and its argument is applied here, obviously Goostman would not recognise it. Surely Goostman's outputs are not satisfactory as that are not in perfect grammatical form. Clearly this isn't of any overriding philosophic significance, but the test does raise the opposition that perfect and consistent systems are not always a correct model of the brain.

1.5.2 Objections to the Turing Test

Bostrom's work (Bostrom, 2012) posits a number of existential threats faced by humanity in an increasingly AI dominated world, but systems of this nature are built solely around data.

Turing gave a number of possible objections to the test in his original paper which make for an interesting read. This was an objection Turing saw possible, which he humously titled the *head in the sand objection* that "The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do this." (Turing, 1950: p. 50)

The one response which I find most intriguing is the *theological objection* which almost collectively encompasses all arguments which has been given against AI. However, AI is not alone here. Humanity has something of a superiority complex that we humans are immensely more evolved an intelligent that any other being and nothing can come close to humanity.

This objection holds the brain to be a superior system, and regardless of whether or not one is religious, the theme is prominent across a variety of replies against AI in subtle ways,

[7]Sometimes given as 'artificial superintelligence'. See Bostrom, 2012

including in the works of Searle and Penrose's *Orch-Or* proposal of an alternative model of consciousness.

1.5.3 The Problem of the Computer Metaphor

From this, it ought be clear that the test is limited. Turing gave only limited scope to the test and its limitations must be clearly acknowledged. There is a clear danger in following the computer metaphor and much of Turing's work though when likening the brain to a computer.

What I hope to have done here is in providing a brief history of computation and the Turing test is present a context towards AI philosophy and why it is so complex. This subject touches on so much that to concisely provide a clear narrative is as difficult as providing one of any philosophy subject. It is likely that AI philosophy will always contain problems which will be immensely discussed, especially where consciousness is involved.

There is a sense of waiting on science to catch-up. AI is still a new field and it is only in the last 30-years that we have seen a rise into every person owning multiple forms of a personal computer, and as this continues, as we increase our understanding what what is currently available and as computers get more powerful we shall be able to deal with more of AI.

Neuroscience has much to contribute here, and as memories and brain activity is better mapped, AI models shall become more precise. This is the aim of connectionism, and this - while a computational model - allows for the problem of the computer metaphor to become clear. A set 'receive/store/execute' model of a A, B, C-style process is going to fail as it does not capture the operational nature of the brain.

This is what needs to be rectified; there may not be a straightforward algorithm as is currently believed under certain models. It may be more fluid.

1.6 The Gödelian Argument

Gödel's theorem is one of the first 'incompleteness' theorem's of logic and mathematics during the 1930's - Tarski's theorem (Tarski, 1936), Church-Turing thesis, the Kleene-Church theorem - all of which have had a profound influence on mathematics which.

As a mathematical variation on the Liar paradox, consider the sentence 'This statement is untrue': if the statement is truth, then the statement is false; and if it is false then it is true. While a fun little language paradox, they don't cause much us much hassle since self-referential statements are well known to lead to paradoxes which are easily avoided by removing self-reference. Russell provided a simple escape of asking 'What statement?'; something which does not enter into Gödel's theorem for clear reason.

Rather that ask if the statement is truth, Gödel's asks if the statement is provable. We now know from Gödel (Gödel, 1931) and Tarski's (Tarski, 1936) proof that truth does not equate with provable, nor does it translate into formal systems. Through replacing 'true' with provable, we find Gödel's escape.

The theorem is almost the birth of modern mechanism in relation to the computer. Since its publication, it has been of paramount importance to abstract computing. With Gödel's first theorem, any consistent first-order theory which formalises the basic arithmetic of the natural number - particularity any formal system of Peano arithmetic - is proven to be hold certain formalisations which are true, yet cannot be proven in the system.

Exactly what the Gödel theorems do for philosophy and mechanism is rather open; nothing stately conclusively either way, but it does prevent a number of concrete claims from being established[8].

1.6.1 Gödel's Conjecture

Gödel's 1951 Gibbs lecture (Gödel, 1951) was one of the first attempts at using the theorem to reason about human intelligence. From the theorem, he drew the following conclusions that the human brain is not a consistent finite machine, and there exists Diophantine equations for

[8] See§6

which it cannot determine whether solutions exist. Gödel found the second point concerning Diophantine equations to be implausible since it could not be disproved.

A Diophantine equation is a polynomial equation of the form $P = Q$, where P and Q are polynomial with coefficients of some field, often of a rational number, $\mathbb{R}$.

The Diophantine equation was Hilbert's tenth problem and closely related to the *Entscheidungsproblem* since it concerns decidability. Hilbert's tenth ask for a general algorithm which can decide whether or not the equation has a solution where all unknowns take an integer value.

Given that there are many applications of the incompleteness theorem - many are subject to a misunderstanding - and while only a conjecture, it stands to be highly convincing that Gödel himself would hold this position. The theorem in its complete form is an immensely complex piece of (meta)mathematics. The attempt to fully detail the theorem here would be futile for this very reason. This conjecture only raises further questions. Gödel's writings on AI - while brief - are thought-provoking.

This is the problem with mechanism. Shapiro (Shapiro, 2003) stated the exact content is left to unspecified. This lack of clarity concerning the type of machine and its outputs places a vast amount of empty space surrounding the claims. Of the outputs, commonly these are given to be those rendered in first-order logic or first-order Peano arithmetic. This is the case with Lucas and Penrose and much who have followed on from their work remain committed to outputs of Peano arithmetic since this is what Gödel's theorem dealt with. Penrose restricts these outputs to those of Π_1-sentences, while Lucas makes no direct comment on the nature of these outputs; their language, propositions, what they express, or even what they cannot express. Of any first-order or propositional logic there are limitation which are imposed in the sentence as we have seen from both Gödel and Tarski's theorem.

While there is idealisation on all sides, limitations inevitably remain yet one thing that isn't frequently mentioned is that is seems obvious that there *would* be limitations. If we disregard the question surrounding how outputs are represented and assume that they are given in a concise manner, it still seems that there would be a limit to what the system can

expressed. Can everything be expressed in first-order logic? Can everything be expressed as a proposition? The totality of all which can be expressed which can be asked - i.e. concerning everything - is immeasurable. Similar applies for mathematics in that there is a finite number of propositions which may be expressed, with fewer being expressed in first-order logic with the addition of those which cannot be expressed in consistent systems.

1.6.2 Gödel Numbering

Gödel was able to block the ban on self-reference through the use of his own scheme of Gödel numbering. The scheme is very simple to understand but its operation is difficult. Gödel's method is to assign a symbol to each symbol used of the formula so that statements within a system can be represented by natural numbers. The system used was based on prime factorisation where each natural number can be factorised as its primes: $72 = 2^3 \times 3^2$.

Lucas writes the use of Gödel numbering to allow us to "refer [to], difficult to prove that it can represent, or translate the meta-logical features needed for Gödel's argument" (Lucas, 2002; p. 203).

1.7 Effective Computation and Turing Machine Machines

From the 1930's onwards until the 1950's, before the world began to experience the early days of the digital computer, automata theory was use to study *abstract machines*. Turing began the study of these "machines" in the 1930's publishing *On Computable Numbers, with an Application to the Entscheidungsproblem* (Turing, 1936), and with the *oracle machine*, or o-machine, in his doctoral thesis *Systems of Logic Based on Ordinals* in 1938. Turing's goal was direct; to determine what machines could do, and what they could not do.

How efficiently can a problem be solved? The solution to any computational problem requires an algorithm of sorts, but is this algorithm the most effective for the problem? This is an open problem of computer science[9] and the Turing machine is a simple solution by means of an abstract machine for an effective solution to an algorithm. The machine itself

[9]See the P vs NP problem.

is simple to explain and has been done so hundreds of times often with an accompanying drawing, but explaining *why* the Turing machine is significant is necessary.[10] From its base description and function it can be applied to a host of theories within AI philosophy and often can be misused. The explanation of its function comes in the way of the Church-Turing thesis - itself also highly misused.

Imagine an infinite length of tape on which are an infinite number of boxes which we call 'steps'. Each step marks a given location which contains a symbol, `0`, `1`, or it is blank. The machine's operation is that it manipulates the symbols according to its table of rules. Despite the simplicity of the machine, it can simulate any computer algorithm regardless of its complexity. For example they may read; 'AT s1,735, PRINT `1`'. At this point the machine will erase the symbol and print a `1` or any other symbol from a finite alphabet which its rules state for that step. At each step the machine, may write a symbol or leave the symbol, before then moving to a different symbol. The machine has three basic operations which it may perform:

1. Read the symbol on the square.
2. Edit the symbol by righting a new symbol or erasing the current symbol.
3. Move the tape left or right a certain number of squares and repeat the process.

The rules may read similar to the following:

1. AT s1,735, IF `0`, PRINT `1`. MOVE TO s9,336.
2. AT s9,336, IF `0`. NO ACTION. MOVE TO s57.
3. AT s57, IF `1`. PRINT `0` MOVE TO s9,336.

Our machine beings with a tape on which it has a string of `0`'s and `1`'s and following its rules it will produce a different string of `0`'s and `1`'s. Eventually our machine will halt, having completed the computation.

As we have said that a Turing machine can solve any computable algorithm, it must be

[10]It is of note to mention that each author provides a slight variation on the Turing machine but its core principles remain that same.

noted that there is no single Turing machine. Where we say a Turing machine can solve any computable algorithm, we refer to the collection of Turing machines, as in to say; there exists some Turing machine which we can construct to compute the algorithm. Due to this, each description of a Turing machine will differ. Some authors will describe it as only moving one step left or right, others will have it move to any step in any direction depending on the rules. So long as it obeys the same basic principles, there is no problem. There is however a catch to what is computable. Despite a Turing machine having an unlimited tape (memory) and theoretically unlimited time, computers cannot solve all mathematical problems.

1.8 Unsolvability in Computation: *The Entscheidungsproblem and Algorithmic Decidability Problems*

Turing's interest of the *Entscheidungsproblem* began sometime around 1935 from a lecture by M. A. Newman at king's College. The problem, as Feferman puts it, is this:[11]

> [T]he question [as to] whether there exists an effective method to decide, given any well-formed formula of the pure first-order predicate calculus, whether or not it is valid in all possible interpretations (equivalently, whether or not its negation is satisfiable in some interpretation).

While this problem has been solved in the affirmative for certain classes of formula, the general problem remained open.

This simple question is among the most complex open question within modern mathematics; P vs NP. It asks simply if P = NP, or if P $\neq$ NP yet stands as among the most difficult questions in mathematics perhaps taking over from Fermat's Last Theorem solved in 1995 by Professor Andrew Wiles.

What P versus NP asks is for every problem where the can be effectively verified in P-time, it can also be quickly solved in P-time. P complexity in computational complexity theory, also given as PTIME contains all decision problems which can be solved by a deterministic Turing machine by application of polynomial time. NP is nondeterministic polynomial time

[11] Feferman, 2006; 3

which is the set of decision problems where the “yes" answer must have efficiently verifiable proofs. NP problems are those which can be verified quickly yet not solved.

The travelling salesman problem (TSP) is the most famous example of an NP problem not to be in P; it is an NP-hard. It asks that given a list of cities, and the distances between those cities, what is the shortest route which visits every city once that the salesman may take, and then return to the original city.

This problem could be made more complex through the addition of certain variables such as gas limit, monetary allowance, speed limits, but it remains a NP-hard problem. If a solution was given we could easily check to verify if it is correct, but to find the solution it would require more time since one would have to try every possible route to verify each against each other.

Problems of this nature take time regardless of undertaken by a brain of with the use of a machine, and the problem has a long history.

While it was formally defined in 1971, and it most frequent definition comes from Stephen Cook in his official statement for the Clay Mathematical Institutes's Millennium Prize Problems.

In the early 20th century, mathematicians where fascinated by the question of how much mathematics we could do by following an algorithm. A Turing machine is not a physical computer; it is a model of computation. The P vs. NP problem is a central problem from computer science and a mathematical solution that P = NP would hold profound implications on computer science and mathematics with Gödel writing:

> [I]t would obviously mean that in spite of the undecidability of the Entscheidungsproblem, the mental work of a mathematician concerning Yes-or-No questions could be completely replaced by a machine. After all, one would simply have to choose the natural number n so large that when the machine does not deliver a result, it makes no sense to think more about the problem"[12]

[12]D.S Johnson *a brief history of np completeness*

1.9 The Church-Turing thesis

The mathematical work on computation in the 1930's established the groundwork for much on AI philosophy, including the Chinese Room, and the L-P result. The Church-Turing thesis is an indirect result of independent mathematical work by Church and Turing, concerning *effectively computable* algorithms. It is also highly complex and specialised to its subject[13]

It's not required to give a full explanation of the Church-Turing thesis as it can be easily be found elsewhere in extensive detail, but it is required that that it be given a context within AI and the confusion which it has caused.

The thesis[14] concerns computable functions which, at the time, where referred to as effectively calculable, and continues from Gödel's theorems. These functions where computable by the pencil and paper method. The thesis itself refers to proofs from Church and Turing independently where they proved three defined classes of computation. Simply, the thesis states that a function is λ-computable if it is Turing computable and if it is *generally recursive*. Broadly, the Church-Turing thesis concerns what can be computed by a human mathematician unaided by a machine, and using the pen and paper method. Whatever can be computed by a mathematician - including those idealised - may be computed by a Turing machine. This is not to say that there are certain limitation on what a machine may perform since the machine has the capacity to carry out computations which a human mathematician could not unless they were aided by such a machine.

While these mathematical theorems may seem like a far-throw from the Chinese room argument, they are closer that you might think. J. R. Searle bases the architecture of his Chinese room on Von Neumann architecture and believes his room to be a Universal Turing machine. This shall later (§5.3) be given an extended discussion.

[13]We return to this in §4

[14]This is not an actual paper but a general name to refer to research from both Church and Turing in the late 1930's. As a result it is also referred to as Church's thesis, or Turing's thesis.

Chapter 2

Imitation and Knowing

I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it.

- Alan Turing, 1950.

In the wake of the modern computing, and particularly following Turing's *Computing Machinery and Intelligence* (Turing, 1950), philosophy on AI became divided into two camps. Of these, Searle's Chinese Room argument (Searle, 1980) is perhaps the single most widely discussed argument in contemporary philosophy of AI behind Turing's paper. Searle is part of a group of academics whose argument towards AI is grounded in consciousness; the claim that a machine cannot be intelligent, or have intelligent understanding due to lacking a consciousness (Lucas, 1961, Putnam, 1973; Putnam, 1975; Penrose, 1989, Penrose, 1994).

This 'two camp' view is narrow but does encompass the broad philosophic discussion. This is where we see the problem of the Turing test emerge. In defining actions of the brain as algorithmic we are able to bypass all additional problems including how such problems can be programmed. What we find with Searle is that the use of a machine which can talk is deeply flawed since it neglects discussion on the operations of a machine and deals only with a specific algorithm. What is evident from these two camps is the central point of mechanism where one side is in defence and the other is the anti-mechanist. The problems of mechanism have been explored extensively outside of the (anti-)mechanist debate, yet they have not been explored within the debate in the same manner.

The application of Searle's paper (Searle, 1980) has stretched into a number of separate

fields of study and Searle has stood by his argument since 1980, publishing further extensions in (Searle, 1984; Searle, 1990a; Searle, 1990b) In the years following, it has been criticised, defended and included in more papers that we could list. Searle's argument primarily concerns philosophy of mind and while it is within the field of AI philosophy and research, many consider it to be irrelevant. Russell and Norvig write that many "don't care about the strong AI hypothesis - as long as the program works, they don't care whether you call it a simulation of intelligence or real intelligence" (Russell & Norvig, 2003; p. 947).

The argument doesn't reach its peak until 1990 with the publication of *Is the Brains Mind a Computer Program?* (Searle, 1990a) and *Is the Brain a Digital Computer?* (Searle, 1990a), both of which develop the more formal versions and introduce greater thought on prediction and simulation. As such it is necessary to detail the Chinese room first, not least because it was published first, but because it is Searle's argument against the claim that the brain is a digital computer.

At most, a machine can be said to replicate certain basic functions of the brain; for Searle, a machine can replicate the basic functions of communication with respect to a natural-language conversation (Searle, 1980). The basis of Searle's work is that a symbolic text-based system which can pass a Turing test is not intelligent since it has no understanding of the language it outputs.

Searle's result here has more in common with the Gödelian argument that it initially appears since it concerns the brain as a Turing machine. Where results based on Gödel's incompleteness theorems (Lucas, 1961; Penrose, 1989;Penrose, 1994) prevent the machine from replication of certain mathematical function due to any Gödelian system being unable to recognise its own consistency, both systems require an algorithm, and it is this which is of significance.

Since a brain is able to recognise its consistency, it is therefore of a higher power than a Turing machine, preventing the brain from being a Turing machine. What Searle argues concerns a 'complete' system. By this, it is meant that the system does not adapt since it follows an algorithm and and is incapable of acting independently. While the L-P result

makes no similar absolute claim, Lucas does highlight free-will as being a problem of note concerning AI systems. While the similarities between Searle and the L-P result, the architecture of both systems is radically different, yet they contain similar problems, such as those of free-will and deterministic systems from a finite alphabet.

Viewing Searle's room purely as an analogue of a computer system doesn't work and as a result Searle later extended the operations of his machine out into a formal system (Searle, 1990a; Searle, 1990b). This system is both confined only to its own logic of the rule-following basis Searle gives. This architecture will be expanded upon shortly (§.2.1.1).

Beyond the Chinese room, Searle's *Rediscovery of the Mind* is a sustained attempt to locate the mind in both the brain and the physical realm. Consciousness is defined as being a biological feature of the world (Searle, 1992) and something which is clearly present in the brain. By 'clearly present', it strictly means that it something which we could locate through the use of brain scans, but that it is a part of the chemistry of the brain; "Consciousness. . . . is a natural biological phenomenon" (Searle, 1992: p.9).

Both works are in direct relation to a potential artificially intelligent system which could produce "simulations of human cognitive capacities" and systems which could "answer questions about the story even through the information that they give was never explicitly stated in the story" (Searle, 1980; 417). These are what Searle defines as among the goals of 'Strong' AI. Both these claims with the addition of Searle's Chinese room are three separate claims each with little linking them in a concise manner. Schwering (Schwering, 2017) presents a system of first-order logic which allows for a system of limited belief which allows for representation and reasoning within statements which are expressible in first-order logic while Schmidhuber (Schmidhuber, 2006) presented a class of "rigorous, general, fully self-referential, self-improving, optimally efficient problem solvers" built on Gödel's theorems. What these results allow for is a system of reasoning - the Chinese room - to exist where it may express statements as propositions of first-order logic and then re-write themselves so as to adapt to new proofs. Any such system can theoretically act as a middle-ground between Searle and the L-P result with the addition of not being bound to its initial program.

Of the arguments given within the *Minds, Brains and Programs* paper (its central being the Chinese room argument illustrating Searle's argument that syntax is not sufficient for semantics since both are highly different) Searle writes the argument he gives "would apply to. . . any Turing machine simulation of human mental phenomena" (Searle, 1980; 417). This claim is deeply problematic (see §3 & §4).

What we study here is Searle's argument against what he claims to be the goal of 'strong' AI; what could more generally be referred to as artificial general intelligence (AGI), which is generally accepted to be a machine with the full capacity of human intelligence. Searle system does not however reach this capacity as it is not given to the system. The L-P result allows us greater freedoms to determine what is logically possible concerning a specific system whereas Searle's system is only a thought experiment with the explicit purpose of demonstrating his original argument.

We take issue directly with Searle's use of language and the nature of his argument. It is our intention to demonstrate that the Chinese Room argument presents a highly simplified version of AI, that it fails to involve a number of critical questions, and that Searle plays fast and loose with a number of key terms. Specifically in that his Chinese Room is in-line with a formal system[1], yet he makes no effort to present the room as being such.

We highlight four major problems with Searle argument. The first is the restrictive nature of this argument, the second concerns his theory of *biological naturalism*, and the third contains the error concerning the Church-Turing thesis. The fourth error is not something which has been given direct attention but it is my belief that many of those who have replied to Searle's room are aware of this problem; Searle gives two versions of the room (Searle, 1980) and uses them interchangeably. We shall detail this in §.5. Combined, we aim to show that these areas are incompatible with Searle's central claim, highlighting that the problems of Searle's room run deeper that an objection to his syntax vs. semantics argument. What we aim to show is as followed:

(i) Searle's claim that syntax is not sufficient for understanding the semantics is flawed.

[1]See Searle, 1990a; Searle, 1990b.

It is Searle's belief that the there must be an clear understanding of the words the system uses for it to be deemed intelligent yet this neglects a vast array of literature on first language acquisition and avoids what could be a substantial discussion on how a machine *could* understand or utilise language. Further this claim is based within *biological naturalism* and deals with natural language as a method of communication not as representative. There exist a number of reasoning systems based within logic where the application of which could allow for a system to make judgements, answer questions, and reason but Searle keeps it solely within the confines of a conversation and a biological understanding.

Where it the case that Searle presents a theory for an intelligent language-based system complete with a set of axioms or preliminaries before the presentation it would allow for a greater analysis of his argument in addition to the establishment of a more concrete foundation. The total lack of this means that what we are left with is an argument which fails support itself before it resorts to a *reducto ad absurdum*.

Critically, the lack of doing so combined with no detail concerning what sort of questions the room responds to, or the axioms of the Chinese room. It is only inferred as to what the room can speak of. Since Searle talks of the biological phenomena of attaching meaning to words, we are left with the impression that it concerns only a natural-language conversation and makes no judgement on reasoning or mathematics.

This becomes convoluted through Searle's own use of the Church-Turing thesis which deals with *functions*, not natural language, where he writes of his room that it would "apply to. . . any Turing machine simulation of human mental phenomena" (Searle, 1980; 417)[2]. How are these phenomena represented? This claim is firmly rooted in biological naturalism while giving no thought toward the distinction between symbolic and numeric functions.

(ii) 'Biological naturalism' and the theory of consciousness as being a biological phenomena is an incoherent theory of the mind. There exists no single accepted theory concerning consciousness and all scientific empirical studies have worked on separate aspects of the phenomena (Płonka, 2015). There is also the problem of reconciling the philosophic view

[2]Emphasis my own.

of consciousness with the scientific. Searle's theory is no exception to the rule that purely philosophic responses to the phenomena don't work.

In discussing consciousness, we encounter a number of problems concerning the language, and there still remains the main question of 'is consciousness necessary for intelligent agents?' and if so, to what extent does this self-awareness exist, and precisely what about itself the machine would know[3].

From this we can see a clear difference between certain cognitive abilities and what can be collectively referred to as consciousness while it also demonstrates the failure of the attempt to categorize cognitive actions and determine which categories each action belongs so[4].

As a final point concerning biological naturalism, despite Searle's frequent insistence that he is not a dualist, this does constitute as a form of dualism, and nor does it solve the mind-body problem. This theory of consciousness is the heart of his Chinese room argument as it provides an answer for *why* syntax is not sufficient. This problem will therefore be addresses across all sections concerning Searle.

(iii) The Chinese room is defined as being a universal Turing machine capable of running any program. As Searle's arguments are said to all be applicable to any Turing machines' simulation program, they are immediately retorted by the Church-Turing thesis[5]. This could be true if this were what the Church-Turing thesis claims, but the thesis does not claim that universal Turing machine can simulate the behaviour of any machine, or that for any algorithm it can be simulated by a Turing machine. In believing the room to be a universal Turing machine, Searle fails to alter the output of the machine and determine what outputs they would be, or how human outputs could be represented by the outputs of a Turing machine since the machine deals with algorithm all while resting his argument on a misunderstanding and false assumption. We detail this further in §4 & §5.

Searle's simple arguments against AI, and simple solutions to the mind-body problem and consciousness are fraught with ambiguity and loose definition whilst being based on

[3]See Carlson and Alexander

[4]See Dennett, *1995* for extended discussion.

[5]See *Is the Brain a Digital Computer*, Searle, 1990b.

his adaptation of Cartesian ontology. In contrast to the traditional dualism, Searle's replacement consists of a multi-layered hierarchical system where beginning with bottom level basic particles gradually ascending to atoms, molecules, unicellular organisms, multicellular organisms, before finding organisms in possession of intentionality and consciousness.

The successful unification of these is of immense difficulty since each is distinctively unique from the next. An argument against AGI based on human understanding of language is possible - this could be achieved through application of Chomsky's linguistic models - however the strength of said theory leaves much to the investigation. The use of a theory of mind is problematic; which do we take? Do we begin with a theory of mind or work towards one? Do we take a specific theory and work from it or do we adapt it around the argument? As we can see there is much to be said concerning Searle's approach and it is my belief that any single theory towards AGI will inevitably fail; there is too much which can be said. While thought experiments such as the Chinese room do well to illustrate a point, often they can encompass too great a subject and collapse under their own weight.

Shortly (§5.2) I shall present a response to Searle's formal argument (Searle, 1990a; Searle, 1990b) by presenting a formal variation of the Chinese room based on automata theory. This shall allow for each state to be more accurately represented, while additionally allowing the argument to adhere to formal logic thereby bypassing any bias which Searle may have imparted onto the original argument. What shall be demonstrated here shall be that Searle's misunderstanding of the Church-Turing argument where he says the Chinese room can run any program is a false assumption, which adds confirmation that his room cannot be Turing machine, and that he initially presents two versions of his Chinese room in his original paper, before later presenting a third version of his room (Searle, 1990b).

2.1 The Chinese Room Argument: *The argument and an overview of the broad problems*

Searle's argument asks us to imagine a man is placed in a room with a rulebook containing an immensely detained set of instructions. The man has a basic task; through a slot in

the room's wall, a card enters on which is a string of characters to which he must response. Unbeknownst to Searle they are in Chinese. Since he does not understand a word of Chinese, he consults the rulebook and is able to produce an accurate Chinese response. The rulebook provides him with no detail other that 'IF -', 'THEN -' symbols. Searle has provided a concise description of his original argument, writing:

> "A digital computer is a device which manipulates symbols, without any reference to their meaning or interpretation. Human beings, on the other hand, when they think, do something much more than that. A human mind has meaningful thoughts, feelings, and mental contents generally. Formal symbols by themselves can never be enough for mental contents, because the symbols, by definition, have no meaning (or interpretation, or semantics) except insofar as someone outside the system gives it to them." (Searle, 1989.[6])

Minds, Brains and Programs (Searle, 1980) sits between Lucas' *Minds, Machines, and Gödel* (Lucas, 1961) and Penrose's first work on consciousness *The Emperors New Mind* (Penrose, 1989), and while Searle's work is not mathematical taking no use of any initial state, Q, or an alphabet of $\Sigma \in \Sigma^*$ it argues a similar case against strong AI due to how machines produce certain computations. We argue this not to be the case.

While Searle expanded and made his argument more formal (Searle 1990), it remains as narrow as his original work. As earlier mentioned (§2) the use of consciousness affords Searle a degree of difficulty which he either chooses to ignore, or is unaware of; both of which prevent his argument from developing a solid foundation. Searle's lack of awareness shall become clearer as we discuss further and from his misuse of the Church-Turing thesis[7]. There is an ambiguity (Searle, 1980) which causes a significant apprehension; it's simplicity is its difficulty as it becomes challenging to know precisely which of Searle's claims deserve the most significant attention.

[6]Searle, "Artificial Intelligence and the Chinese Room: An Exchange", *New York Review of Books.*

[7]See §4.5

Rather that present a thought argument directed towards a specific claim, Searle builds a complex web of side-points, each of which one may devote substantial attention to.

The sole use of a natural language conversation is used to denounce the entirety of the mechanists' claim, with Searle making no comment the mathematics/logic and giving only a brief mention on Turing machines (Searle, 1980), instead writing that syntax and semantics are two fundamentally separate linguistic qualities. What Searle argued from this is that the computer cannot understand the semantics by following an algorithm, and that this inference of meaning is a biological phenomenon.

There is the rather interesting discussion which can be had as to what the core of Searle's argument is. Chalmers believed it to be that consciousness is the root of the matter (Chalmers, 1996) and McGinn wrote that Searle's argument serves the purpose of providing a case for the argument that the hard problem of consciousness is unsolvable (McGinn, 2000).

Searle's construction of his room causes it major problems. It is important to first state Searle's paper to be a response to the Turing test, and the claim that satisfying this test is sufficient for a machine to be deemed intelligence, which is not entirely accurate.

A highly notable critique was given by Abelson[8] (Abelson, 1980) which concerns how mathematics is understood by the brain. Beyond providing a concise demonstration as to the narrow scope of Searle's argument concerning rule following and understanding, Abelson's point highlights a clear failure on Searle's part.

While Abelson presents a concise objection to Searle's rule-following, the problem of consciousness is neither solved or left remaining; it is almost disregarded. Brains as biological organisms are more critical to Searle's argument. It can easily be postulated that Searle's brains are elevated to an ontological principle of a *mind*.

Beginning with a basic story, Searle's questions concerning understanding and intentionality do work to a degree as a response to the Schank-style AI story test while confining him to a specific test. Dennett (Dennett, 1980) writes of Searle's challenge as being fair

[8]Peer-review in Searle, 1980.

play due to Schank having “allowed enthusiastic claims of understanding for such programs to pass their lips, or go uncorrected". While Searle is right to challenge this story method, questions concerning linguistic programs do no necessarily lead one to determine mechanism - or some extension, or similar theory - to be incorrect. Dennett does accept this or similar writing Searle’s challenge to also be “cheap shot, since it has long been a familiar theme within Al circles that such programs tackle at best a severe truncation of the interesting task of modelling real understanding." (Dennett, 1980; p. 429).

This criticism of Searle shall additionally feature as a criticism of the L-P result. In the case of Searle, what shall be argued is that a natural language program such as the Chinese room is highly narrow. Since the program has no understanding of Chinese and from Searle’s description of its operation method, it can only provide a response to a question.

In the case of mathematics where there are no syntax to semantics relation, similar to what Searle talks of in the linguistic room, the methodology of following an algorithm of sorts to provide an output is the same, and remains correct so long as the rules and axioms are correct. This is what is meant where Dennett refers to Searle’s cheap-shot and the room being too narrow.

Searle takes no issue with the claims of weak AI, and of strong AI writes: “[It 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 said to understand and have other cognitive states" (Searle, 1980: p. 417) Despite being an attack on ‘strong’ AI, Searle room adheres more to a narrow AI system since it deals only with one task, and of that task, a very specific version of a question-answer based system.

In both instances of mathematics and linguistics, the syntax is equal; it is the basic rule(s) of production for an output, or the starting point, more broadly speaking. The method of production which Searle gives, while not complex or detailed, is a method most would view as satisfactory. Between both, there exists no similar understanding of the output in the case of mathematics like we see with linguistics since we cannot understand the content the same; the linguistic may refer to physical objects, and the mathematical to the abstract.

Thagard (Thagard, 1986) argued that the *principle of inductive adequacy* is critical to the success of learning AI, and that the application of this refutes the claims of Searle that AI systems lack the ability to have semantic understanding. While this may appear as extensive criticism, Searle succeeds in forming a well-constrained argument where language and mental states are clearly and constructively linked. Fundamentally his argument rests on the "simple logical truth that syntax is not the same as, nor is it by itself sufficient for, semantics" (Searle, 1990b: p. 21). Additionally, minds cannot be computer programs since minds - and by extension, brains - have mental states.

The Turing test as a measure of intelligence versus an intelligent system is indirectly explored by Thagard where he writes of LISP code that atoms such as BATTER in relation to a field-batter only present the illusion of semantics but it is non-the-less evident that this atom holds no understanding of the meaning of 'batter' versus a system which has a function called UNDERSTAND can actually be said to understand. Thagard notes Dretske's claim that the only meaning which symbols of a system can hold are ones which are given to them from the programmer (Thagard, 1986; p. 140).

These atoms of LISP code hold no understanding since they provide no substitute of meaning or relation to the world

and this is not a problem. Thagard acknowledges this arguing instead that for a system to hold meaningful symbols it is necessary that "it be able to build upon new constructing the the same ways that people do, expanding meaning beyond what is given in experience." (Thagard, 1986: p. 140)

Here Thagard demonstrates the futility of Searle using the Chinese room as an argument against AI and systems of meaningful symbols. This perhaps stems from a misreading of Turing's paper on Searle's part and in conjunction with a challenge of the linguistic story machine - the problem of which Dennett is so easily able to notice and therefore call into question the complete work - attributes heavily to Searle failure.

2.1.1 Architecture of the Room

The operation of room is less formal that it would initially appear from Searle's writing - particularly as he progressed the argument over the coming decade (Searle, 1990a; Searle, 1990b) - yet it follows the basic principle of a formal system.

While it has been noted that the borrowing of the terms 'syntax' and 'semantic' from linguistics is problematic (Boden, 1988), it is not of great concern in relation to the construction of the room. The borrowing from linguistics shares a relation with Chomsky's work[9], and a relation which can be applied to mathematical logic should one desire to do so.

Further, the architecture of the room is built on von Neumann architecture, and a reply to Turing's *Computing Machinery and Intelligence*. The subject of which being on intelligent replication of conversation between human and machine. Turing's paper and Turing machines of all types (o-machines, Turing machines, and Universal Turing machines) naturally have a close relationship with his work on computability, Turing completeness, and the Church-Turing Thesis.

The function of a digital computer is given by Turing (Turing, 1950) as being able to carry out any operation which could be done by a human computer, and to follow its set of fixed rules; it cannot deviate. Turing's machine, being purely mathematical, was theorised for the solution of mathematical algorithms thereby giving it an alphabet of numeric functions. As we have stated prior, this is not strictly a problem, but there is a problem of how outputs of the brain would be represented; a problem Searle neglects.

While the functional operations of the room are vaguely Turingesque, the architecture is distinctly von Neumann architecture. Of this architecture, the man in the room follows the rules, produces a response, yet does not speak Chinese. The man who computes these inputs with the aid of a rulebook, computes the same as the computer; he acts in accordance with a set of rules. Either the computer understands Chinese, or it does not. We have the program (the rulebook), memory (pens, paper, the symbols, etc.,), and the CPU to follow the instructions (the man). Adjusted, this breakdown shows the extensions of the man, and

[9]See *Syntactic Structures,* 1957.

the extensions of the computer. Since the man and the computer share the same function of computation, we have to ask on the function of its extensions.

Linguistically our syntax and semantics operate as we would expect; they are rules for how to construct the language. In formal systems, they serve the same purpose as we have prior explained. In both languages, our axioms must be correct to construct the semantics and we must also have a set of rules of inference. With a formal system, these are less complex since we are often dealing with a highly specific subject, whereas with natural language, we require a greater number of axioms, axiom schema, and rules of inference. Precisely what questions the room - with the man being a part of - can answer are not provided yet Searle often makes references to language grounded in the physical. Since our room is given no way of being embedded into an environment to receive any form of input, we can assume it answers only basic questions of pre-determined answers; there can be no variables or else the person outside could catch out the machine and since Searle has no way of learning as modern chatbots do it would be foolish for this inclusion to be added by Searle.

There is a persistent dualism throughout Searle paper, with him arguing that computational models provide little insight into the workings of the brain. These computational theories are purely formal, and cannot explain mental processes, and programs are not like neuroproteins, so cannot allow for casual powers of a brain.

In his original paper Searle identified and replies for the following:

1. The Systems reply: intelligences resides in the totality of the system. Since Searle is part of the system he does not require a complete understanding of the room, and neither does any single component of the system since it operates collectively.

2. The Robot reply: if the isolated system is replaced with a robot which has audio-visual sensors, it is able to interact with the world directly rather than though a textual input.

3. The Brain simulator reply: instead of having the program, we simulate the actual neurones of a native Chinese speaker. As such we would be able to develop a intelligent system since we have a perfect replication of the process of language understanding and speech verses the implementation of a pre-programmed code.

4. The Combination reply: a combination of each above argument amounts to a refutation of the Chinese room. (Searle, 1980: pp. 419 - 421)

Searle's response to the 'systems reply' is that suppose the individual internalises all components and memorises the rule-book, they would be of equal power to the original system while having no separate components and yet he would still not understand: "he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system that isn't in him. If he doesn't understand, then there is no way the system could understand because the system is just a part of him." (Searle, 1980: p.419) This reply shall be returned in §5.5 where the internalised version and the systems reply are shown to cause two version of the room to emerge.

In response to the brain simulator argument, Searle writes of losing conscious awareness:

> "You find, for example, that when doctors test your vision, you hear them say 'We are holding up a red object in front of you; please tell us what you see.' You want to cry out 'I can't see anything. I'm going totally blind.' But you hear your voice saying in a way that is completely out your control, 'I see a red object in front of me.' . . . [Y]our conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same." (Searle, 1992, in *Russell & Norvig, 2003*: p. 957.)

This response appears to have little to do with AI and consciousness; it confuses matters by speaking of a dualism between conscious experience and external behaviour. What Searle intends to demonstrate here is that in the case of the brain simulator argument, the gradual replacement of single neurons will cause a deterioration of consciousness and affect reportability. It is difficult to know why Searle considers this of importance since any neuron which gives a false report of a registered sense is a poor example since it naturally takes us away from any concern of significance. Searle's reply is similar to the ship of Theseus paradox which asks if all components of a ship are replaced, is it the same ship? The brain simulator reply sits in-line with connectionist replies which are perhaps the single most useful in regard to the Chinese room reply.

The *robot reply* is of some interest since it talks of where meaning comes from; that it acquires meaning and semantics from this direct relation between the room and world yet it fails to satisfy Searle's definition of understanding since the robot has no brain. I wish to dedicate a brief moment to clarify why this reply holds a unique interest despite appearing to belong more towards a problem of epistemological ontology.

This robot is much more sophisticated that the room as it is able to walk, and interact with the world in a highly direct manner. This veers from Searle's argument but does remain somewhat in-line with it. The room's program - let us call it the 'Chinese program' - cannot any longer be thought of as being purely a program implemented by a machine since it has the additional property of direct interaction; it changes its function and method.

The 'robot reply' is problematic in that it only complicates matters further; it takes an already inadequately defined system and complicates its interaction with what could be defined as meaning full interactions with objects which possess what could be described as an *intrinsic* meaning.

However, if we put aside these problems and remain with the reply, it can be found that it has little difference from Searle's isolated system. The reasoning for this is that if the system observes an object, it must be programmed to interact with it in a specific manner; if it encounters a cup, it must know how to hold and use it. Without the knowledge of the function of the object, the robot would be pointless.

The use of such a robot does not confirm Searle's result, yet it does cause one to alter the Chinese room program since the program can no longer by viewed as complete due to its change of environment. Any reply which has subsequently been given regarding a work-around encounters the problem of being a response to an argument which holds the brain to be the sole origin of meaning.

While it is certainly true that the robot can do what any child would do by learning through interaction, it still abides by a code which prevents it from passing Searle's test. This calls into question the fundamentals of human understanding. Since the robot follows a code to interact, it must be pre-programmed to engage in each interaction. This pre-programming

issue is where Searle's problem with the reply lies since the room is no different from the room as a program is used, not a brain.

Taking the example of a cup, if a machine is programmed to use the action of 'GRAB' it doesn't need to understand the function of a cup *a priori*. As has been recently seen with educational robots, they are able to perform these actions, but they are far from perfect.

It presents itself as being clear that the robot must implement the *exact* same program as the Chinese room, yet this cannot be true since our system is now able to progress audio-visual data along with take use of physical interaction/stimuli. From this, Searle's argument is concerning the robot is no longer purely about symbol manipulation. If we have a robot which is able to speak, then it remains that Searle's point is confirmed since its lacking of a brain prevent the robot from attributing the meaning. While this point can be made, Turing's *other minds* reply (Turing, 1950) returns; there is fundamentally no difference in these. Since any vocal reply of the system, while equal to the room's original symbol-manipulation argument, by having a robot which has interacted with the world there is a compelling argument here which has been made in multiple subsequent responses.

Searle's response is convoluted. Writing that the reply "concedes cognition is not solely a matter of formal symbol manipulation, since this reply adds a set of casual relation with the outside world." (Searle, 1980: p. 420) Pinning cognition down to pure symbol manipulation is difficult since it theoretically makes sense. This being the view of those in defence of mechanism, it is natural that it is Searle target, but his approach bypasses the commonly held view that mechanism is flawed[10].

While pure symbol manipulation *could* be argued to be a central part of cognition, it is uncommon to read of this in isolation from an additional cognitive action[11] since the nature of the outputs and the symbols is vague. What is done by here by Searle is reduce a complex process now to a conversation.

Searle's reply is less of a reply to the robot and more towards an adaptation of the Chinese room *with* a robot. Writing:

[10]This shall be explain in §.6

[11]See Chalmers, *1992*.

> Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case. . . Suppose, unknown to me, some of the Chinese symbols that come to me from a television camera attached to the robot and other Chinese symbols that i am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these other facts. (Searle, 1980:p. 420)

The main issue here is that Searle further bloats, and detracts from his original argument. He readily acknowledges himself to be the symbol manipulator: am receiving "information" from the robot's "perceptual" apparatus, and I am giving out "instructions" to its motor apparatus without knowing either of these facts." (Searle, 1980: p. 420).

In this sense, the robot adheres to no form of self-autonomy, nor does it follow a code. Questions of robots and interaction, and simulated brains, or actual disembodied brains, and stimuli are interesting with regard to post-modernism and epistemology, it does little for the Chinese room.

All replies he details, he dismisses. Much of these arguments which include some form of dualism are rejected by Searle, particularly dualism of strong AI where the mind is concerned that the brain does not matter. I shall argue that this dualism isn't the case, and that the dualism Searle talks of between the brain as a machine and the mind as a state of consciously aware mental states is irrelevant to AI and indeed much of the neuroscientific community, particularly research into the human connectome (Sporns, Tononi, Kötter, 2005)

Concerning the robot reply, there is little difference for a simple reason; it is not a brain. Limitations of symbolic text-based systems have been prior recognised, e.g., D. Michie and R. Camacho (Michie & Camacho, 1994), and the relation between computers and formal systems explored, e.g., Yee (Yee, 1993). Michie & Camacho note:

> [G]iven sufficient sampling of a trained expert's input–output behaviour, machine learning programs have been found capable of constructing rules which, when run as programs, deliver behaviours similar to those of their original exemplars. These

'clones' are in effect symbolic representations of the subcognitive behaviour."
(Michie & Camacho, 1994: p. 385)

Substantial extensions are however required to allow for the test to account for important parts of thinking and Yee (Yee, 1993) highlights how Turing machines are static.

2.1.2 Understanding from the System

Nasuto et al (Nasuto, Bishop, Roesch, 2015) give Google translate as an example of a complex rule-book and appear to avoid the clear issue of relating this to any formal system/Turing machine, thereby binding it to the CTT, by referring it to high-level systems with high-level rule. The example of high-level rule - *Google* Translate - is a system with a process analogous of the Chinese room. The gradual process of understanding begins from the rule-book. All English to Chinese rule processes are contained within the English language rule-book. This rule-book doesn't necessarily have to be in English, by the language of the computer. In this case it is English since we have the man in the room. Each rule required for the operations are programmed in English, the man forms a slow understanding. Searle brings to the room the knowledge of the English language, just as the computer brings the language of its code to the system. Each 'system'[12] holds a language which it knows and utilises in the translation process. This is liked (Nasuto, Bishop, Roesch, 2015) to Boden's *English Reply* (Boden, 1988)

2.2 Machine Reasoning

Dealing with the ambiguity is a long-term goal of machine learning, and there exists a number of methods by which systems can deal with reasoning in natural language (Xiong, et. al., 2016) (Hermann, et. al., 2015). These deductive systems are able to produce and accurately answer to question where chaining facts are not always required information for the agent. What is proposed here is that recent deep learning methods allow a machine to successfully answer questions from a script by reading, and also to reason when the information is not

[12]Be it biological or software.

made explicit. Solutions to Question Answer (QA) systems, especially those containing incomplete knowledge, often require a highly expressive - or declarative - language to be used so that incompleteness can be represented and reasoned with.

Such systems will not make jumps in logic in cases of indefinite information but can be trained to provide an answer which reflects the missing information. Weston, et. al. demonstrated a series of methods by which an agent may reason and answer questions with chaining facts, including supplying and answer when one is not made clear.

2.2.1 Schank's Story: Reasoning Systems

Searle's reply to Schank's restaurant scenario is that with human reasoning it is a straightforward task of answering if a man does, or does not, eat a hamburger based on how it is served:

> "A man went into a restaurant and ordered hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip." Now, if you are asked "Did the man eat the hamburger?" you will presumably answer, "No, he did not." Similarly, if you are given the following story: "A man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill, ' and you are asked the question, "Did the man eat the hamburger?, ' you will presumably answer, "Yes, he ate the hamburger." (Searle, 1980; p. 417)

Story A provides a fairly clear narrative; the hamburger was burnt and the man stormed out. Naturally we assume the man did not eat the burger since we assume him not to want burnt food.

Story B is similarly straightforward in that it offers two lines of information yet has the same subject of story A; reasoning towards an answer. Unfortunately the story is reliant upon an understanding of the culture of tipping and, unfortunately, manners; it is not clear

that we can say either way. While Searle says that we would probably say the man ate the burger, it is only implied the man ate the burger since he was satisfied, and both paid and tipped. It can be assumed with high probability this is true, but the story does not contain the information to give a certain answer. Were the latter of tipping and paying removed, it would remain reasonable to assume the man ate the burger since he deemed his order to be satisfactory.

Both stories given by Searle hold indefinite knowledge in that they don't mention if the burger was eaten. Intelligence extends beyond giving a correct answer, or just relaying what has been said, such as if asked if the burger was eaten when the statement makes clear the burger was or was not eaten.

This introduces a social contract. By established a set of rules towards constructing a family-tree of the operation of a social contract, we can allow for a system to reason and work through the process to produce an answer.

In the case of a restaurant, if a system is equipped with the knowledge that a person orders a service from a restaurant with the agreement of the exchange of services for money. By having it established that payment is only given if the services are good we can write:

`PAYMENT if FOOD{burger}` = GOOD.

They give a script to a task, **Task 10: Indefinite Knowledge**, as the following:

> John is either in the classroom or the playground.
>
> Sandra is in the garden.
>
> Is John in the classroom? A:maybe
>
> Is John in the office? A:no
>
> (Weston, et. al., 2015; p. 4)

Writing of its purpose as being that "Task 10 tests if we can model statements that describe possibilities rather than certainties" (Weston, et. al., 2015; p. 4).

If we considers Schank's story in the form of a **Task 10**, we can give:

> A man when into a restaurant and order a hamburger.
>
> The hamburger was burnt.

The man left the restaurant without paying.

Did the man eat the hamburger? A: Maybe

This surely would not be enough to convince Searle since the machine cannot determine yes-or-no, but it is by no means a vast leap to develop a system which could answer 'no' since the reasoning behind it is not complex. If we had a story which read "A man ordered a hamburger and when it arrived it wasn't burnt. The man left the restaurant" it is not possible to give an answer here since a significant piece of information was missing. The good qualities of a hamburger extend beyond it not being burnt.

Further research into these test is required to extend the capabilities of an agents learning methods and minimise supervision. Ma et al (Ma, Cui, Wei, Sun, 2018) developed a model of machine commenting which is trained in an unsupervised manner by proposing a neural topic model.

2.3 The Problem with Rule-following

Here I wish to consider a highly simplistic example of rule following which could be done by a machine, and view this as the context of Searle's argument.

The complexity of language is the first difficulty. Initially the complexity of the Chinese language as a whole is of note; it is not one single language with only slight regional differences, but a collection of different languages such as Mandarin, Wu, Min, and Yue. To claim that one could produced a single rule book of any language and have it able to perfectly converse is a highly primitive understanding of languages. As such the specificity of the language would need to be agreed upon prior to the conversation, such as both converse in traditional Mandarin. The programmer would be required to be fluent in the language on all aspects.

Consider yourself in the room. A slip of paper comes through a slot and you are given a brief equation. It asks you to solve for F. You have zero knowledge of physics yet you consult the manual on the desk. On the slip it states a number of varying factors such including the required mass, m, and acceleration, a. The manual states that $F = ma$. From here, it would

be possible to present a correct solution to the 'solve for F' question.

This is a highly basic example, and more complex mathematical variations could be given but since the solution to F is dependant on the correct equation methodology, the rule is required to be correct. For this, no biological understanding is required since we have no reason to apply any mental understanding, or give semantics to the response.

The person would need no knowledge of physics or mathematics, only a detailed manual, accurate questions, and the correct information on the card. The card could easily involve factors placed there as a test to see if the person in the room has accurately followed the manual. We can extend this to much of mathematics since it is all easily defined and follows a set of very regimental axioms.

So if a correct answer is given, what has the person done? Naturally Searle's argument is that this is just rule following and no meaning is given to the string but one must acknowledge there is a significant distinction which can be made between mathematics and natural-language conversation yet the methodology remains the same.

2.4 Conclusion

Here we have established the base-version of Searle's argument concerning the imitation of a behaviour. Searle's correlation between this imitation and the intelligent action it is said to perform is presented as accepted under the view of 'strong' AI. This is problematic due initially due to its origin in McCullough and Pitts as being the computational theory of mind (McCullough and Pitts, 1943) and due to 'strong' AI beginning a definition of Searle's; the position was essentially formed my Searle.

Searle's version of the CTM takes a divergent approach through then neglect of neural networks and remains solely with a traditional philosophy of mind. A standard response to Searle here, may take issue with this form of the CTM, yet the Chinese room itself is equally as problematic since it has a reliance on a simple, and very direct form of rule-following. Critically here the issue is this ambiguous form of rule-following. As we have explained (Abelson, 1980; Nasuto, Bishop, Roesch, 2015) there is a serious conflation between rule-

following of language, and high-level programs of such nature.

The formal version of the Chinese room (Searle, 1990a) conflates formal systems into consciousness as hits upon a number of significant errors, most obvious of which is where he writes:

> Nobody supposes that the computational model of rainstorms in London will leave us all wet. But they make the mistake of supposing that the computational model of consciousness is somehow conscious. It is the same mistake in both cases.
>
> - Searle, 2002: p. 16

In our critique of Searle's 'weather simulation' argument (§3.1) we shall explain how these two versions of a simulation - due to being radically different - cannot be used to explain either. The indeterminate nature of weather prediction, as confirmed by Lorenz's result (Lorenz, 1963), are utilised to show the distinction between chaos theory and data versus simulation based on a concise algorithm of first-order logic for the function of intelligence and reasoning.

Chapter 3

Digital Minds

Searle presents the Chinese room as supporting a larger argument than it actually does. As such, the obvious problem that any such rule-book would need to be immense in both detail and physical size, shall be ignored and the focus shall on the concept of a digital mind. There is a difficulty in remaining concise here since Searle digital mind is build on the work of Church (Church, 1936) and Turing (Turing, 1936). Effective computation is clearly a part of the Chinese room

While this was the focus of Searle's original paper, a strict concept was absent with a formal premise argument established in his later works (Searle, 1984; Searle 1990), and with the Chinese room being supported by Searle's third premise; syntax is not sufficient for semantics.

Having already addressed the claim of syntax versus semantics, we shall not tread the same ground too closely, and instead take Searle's formal argument as a bridge towards the L-P result and the question of brains/machines equivalence. While traditionally not mentioned in papers concerning (anti-)mechanist claims, Searle falls into the anti-mechanist camp in its address of formal systems and Turing machines[1]. The concept of 'digital' mind works well with reference to Turing's definition of a digital computer as being anything which can produce an output in accordance with some algorithm (Turing, 1950). Through the use of the man in the room, we find the argument to become rather circular yet Searle has no

[1]Mechanism will be given a great deal of focus in §5 where we shall be concerned with the Lucas-Penrose argument.

reason to find this true since biological naturalism breaks the circle. It can't be that these following axioms can be separated from biological naturalism, but it is not a requirement that we speak in terms more akin to traditional philosophy. While Searle remains committed, the formal version affords greater specificity, and despite not breaking the circle, it allows for more room to breath. However it still stands on a poor axiom.

In these axioms[2] it is clearer to see the implications which Searle directs towards conventional theories of mind, such as connectionism. Through analysis of these axioms, it becomes clear that Searle's argument is not based on any sound logic or neuroscience, but that his response towards AI is born of his theory of biological naturalism. The rejection of Searle's theory of consciousness does not necessarily cause his argument to breakdown, yet its claims cannot be treated with the same level of strength.

There is a marked difference between Searle's original argument and the formal version, yet there is little adaptation between each, only deeper clarity. Johnson-Laird's gave three independent positions on the mind and computation. The alternative to these, he proposed, is that consciousness is not scientifically explainable.

> (1) The human brain (or, variously, mind or mind brain) is a computer, equivalent to some Turing machine.
>
> (2) The activity of a human brain can be simulated perfectly by a Turing machine but the brain is not itself a computing machine.
>
> (3) The brain's cognitive activity cannot in its entirety be simulated by a computing machine: a complete account of cognition will need 'to rely on non-computable procedures' (Johnson-Laird 1987; 252).

Copeland states these not to be the only three and proposed extensions. The brain can be what Turing referred to as an o-machine, or that it may be that the cognitive activity of the brain may be perfectly simulated by an o-machine, but the brain is not an o-machine, and that this simulation cannot be effected by a Turing machine (Copeland, 1998, p. 129).

[2]As they are given in *Searle, 1990*

3.1 Operation and Function of the Brain

Searle (Searle, 1990b) gives the following three questions:

1. Is the brain a digital computer?
2. Is the mind a computer program?
3. Can the operations of the brain be simulated by a digital computer?

Only of *question 3* does Searle provide the answer as being a clear 'yes'. This question can be altered to ask if the operations can be simulated by a Turing machine; the mechanist claim. "The operations of the brain can be simulated on a digital computer in the same sense in which weather systems, the behaviour of the New York stock market can be simulated" (Searle,1990b: p. 21). This was posited by A. Friedman as being: "Are our brains merely Turing machines, where our conscious minds are simply algorithmic programs?" (Friedman, 2002: p. 28)

The use of weather predication and stock markets are poor examples to use. While weather prediction is built on the computational model, whereby a computational resource is used to study the behaviour of a complex system, it differs from the standard computational theory of mind. Neural networks belong to the computational model but weather systems and stock markets are a different entity in that these are not simulations in the sense of 'intelligent' actions

Imagine a travelling storm, weather systems operate on prior data, and give an accurate prediction on how it will develop, or its travel pattern based on large amounts of data, variable wind speed in the storms path, pressure levels, or if it will encounter any other variables which may cause it to change. Standard prediction here satisfies most people, but weather forecasting is notoriously inaccurate.

Mathematician and meteorologist E. N. Lorenz used a Royal McBee LGP-30 digital computer to run a weather simulation (Lorenz, 1963). Starting the simulation in the middle, it was his hope that the cycle would repeat on the middle-previous calculations thus repeating the cycle making weather prediction accurate. Rather than act this way, the predictions

began to deviate from the prior calculations. Where the computer worked on a six-digit precision, the printout work on a three-digit rounded variable; variable 0.506127 printed as 0.506. This minute difference was theorised to have no effect on the outcome. Lorenz writes:

> "Two states differing by imperceptible amounts may eventually evolve into two considerably different states. If, then, there is any error whatever in observing the present state - and in any real system such errors seem inevitable - an acceptable prediction of an instantaneous state in the distant future may well be impossible... In view of the inevitable inaccuracy and incompleteness of weather observations, precise very-long-range forecasting would seem to be nonexistent." (Lorenz, 1963: p. 133)

It is irrelevant to compare weather prediction to the simulation of intelligent action based on unknown variables; the data it receives from the outside world is variable and then replicated, it differs vastly from Searle's language room. This difference extends into how language simulation is built upon rules of grammar and the program works in accordance with rules of production and inference. Weather simulation is chaos theory.

Weather simulation is purely about data, not intelligence. It has no reason to learn, requires understanding, or knowledge about information, neither must it be able to extract data from a string as it presents all available data and someone records it. The two systems are comparable only at a basic level that they simulate some basic action from some given data. Beyond this highly basic similarity, they are radically different. With the stock market, it is just data of stock prices and the change in value. It's barely a simulation because it is an efficient data stream. In trading of this level, a human cannot possible process all information in enough time, communicate this information, and make a decision if required.

Searle (Searle, 1990b) only chooses to address question 1 (Searle, 1990b; p. 21), since he sees the answer to both 2 and 3 to be highly obvious. Of the mind as a computer, Searle views the answer as a clear negative since programs are formally, or syntactically defined, whereas minds hold "intrinsic mental contents. . . the program by itself cannot constitute a mind" (Searle, 1990b: p. 21). If the mind can be explained as a computer, or if "the mind is

to the brain as the program is to the hardware" (Searle, 1990b: p. 21) are two sides of the same coin. The brain as a computer does not require a dualism of mind and brain yet is does not dismiss dualism. According to Clark (Clark, 2016) the computational model encounters several problems such as an information bottleneck, yet he is not against the model, only that he presents a variation on the traditional model holds the brain to be a in a passive state until it receives data, which it then must decide if or how it must act on the data. Clark's model holds the brain to be constantly active, processing large amounts of data each second and then making predictions on this data.

Searle's address here is into a different realm; into post-modernism and *simulation*. We shall address this in a later section but I wish to draw the readers attention to this prior to progression since again we find a blurring of subjects which pervades Searle's work.

3.2 Searle's Axioms

Axiom 1: Computer programs are formal (syntactic).

Searle's description of a program is a fairly simple - and unnecessarily convoluted - way of describing the process by which the process of information processing is that it is first encoded in the symbolism used by the machine before it is then manipulated in-accordance with a set of precise and pre-determined rules.

The constant use of a dualism between the symbols and the machine by Searle is unnecessary and not accurate; it's just confusing trying to keep up with since there appears to be little regard for the questions any rational computer scientist would ask. Searle completely bypasses a number of possible questions about the nature of interaction between the algorithm and the program. In Searle's mind, there appears to be the program and the machine, with the program being what manipulates the symbols. This serves as a rather neat analogy for the dualism between the mind-body problem - something Searle believes to have solved. Since the mind is where the syntax is understood, the body is merely a vehicle for interaction with the input. Searle's computer functions the same, with the machine taking in the input, and the program 'understanding' the symbols. Here the computer cannot understand

anything since it lacks mental contents. Giving the mind as part of the brain is not strictly wrong, but it can just as easily be given as a part of the body since it is physical whereas the mind and consciousness are not. 'Essence' is not a term which Searle appears to be a fan of since it is too similar to Cartesian 'thought' in that the essence of the mind is conciousness. Despite Searle frequent insistence against him being a dualist, in this case, it rather aptly sums up Searle's belief.

Algorithms and the computer are both difference kinds of "substance". If an algorithm is considered as a mathematical abstraction, as is most likely when dealing with a Turing machine and type-0 grammar - those which are recognised by a Turing machine - then it is an abstraction in the sense of numbers, sets, etc., as it is composed of these. If it is considered as a representation of language - as in the case of Searle's machine since he makes not notation on the machine dealing with mathematical abstractions - then it must represent the text in some clear manner. It appears obvious that it would be a textual representation of language since mathematical abstraction would cause issue. Rapaport states that an algorithm may be viewed in these manners while adding an addition that an algorithm may "even be - and indeed ultimately is - "switch settings" (or their electronic counterparts) in a computer." (Rapaport, 2007; 2)

Rapaport earlier (Rapaport, 2007: p. 1) says that the brain cannot be a computer in that there is a programming language which is executed by the brain causing action, but that there it could mean that through a combination of bottom-up reverse engineering neuroscience and top-down cognitive-scientific investigation we could write a program to allow a computer to exhibit mental behaviour. This, Rapaport acknowledges to be different question.

Prior to Searle's summary into what is *axiom 1*, he writes that what goes for Chinese goes for other forms of cognition, and that symbol manipulation is not enough to guarantee cognition, perception, understanding, or thinking. Note that here Searle extends the actions of symbol manipulation out beyond understanding (Searle, 1990; 26) into more general cognitive actions. The claim that symbol manipulation is not sufficient for semantic under-

standing can at least be understood as a solid line of inquiry and can be given the time, but symbol manipulation on to thought seems like a leap.

There are multiple aspects to thought, from thinking between two options, to making an intelligent action, all the way up to thought on philosophy and mathematics, yet Searle places the collective actions under that same general action of thought, along with having symbol manipulation as insufficient for cognition.

While this is against the claims of 'strong' AI, it is important to note 'strong' AI to be a proposition of Searle's own naming, and to some extent, creation. The claims of 'strong' AI are found scattered across much of AI philosophy to different degrees with many authors not totally accepting the claims but providing commentary on methods for which certain levels of specific claims may hold true.

Boden (Boden, 1988) claims Searle takes for granted the assumption that computations are purely syntactical, and can be defined as "formal manipulation of abstract symbols, by the application of formal rules" (Boden, 1988: p. 89). Boden's assumption is here is sound, and a claim repeated across the literature. Searle's argument is directed towards both functionalism and computationalism. The rulebook contains no detail of what the symbols mean, how they operate in a sentence, or anything else which would allow for any meaning to be extracted from them. Searle doesn't provide any preliminaries or detail of how the machine functions, but it is understood by all and that the rules only state along the lines of: IF -symbol- THEN -symbol-. This process in in no-way analogous to the operation of the brain which means that 'strong' AI and the CR-system cannot be an accurate model of the brain. This is extended to state that brain processes and mental processes can be simulated. It is perhaps problematic for this definition to take precedent. In defining it to be 'equal', it is not defined to be complete replication of the neuro-process.

Axiom 2: Human minds have mental contents (semantics)

Axiom 2 is the "point that the Chinese room demonstrated" (Searle,1990a: p. 27) and is a more general version of Searle's third axiom.

When we are given a question - an input string - the brain will analyse and produce an accurate response. How these 'axioms' are fully utilised by the brain is not clear yet there is a distinction between mathematical response and a language based response. While there exists a specific process of analytic breakdown of the individual components of the input string, there is a difference since they are dealing with different forms of response such as to mathematical abstraction such as numbers, fractions or sets of something which we can only refer as opposed to physically refer to.

In comparing the axioms of the brain and the Chinese room, we can only say that the axioms of the room are axioms of language, but not in the case of Chomsky's grammar. the reason for this is simple; Searle writes that if you are in the room and a card has a symbol, you find the symbol in the rulebook which says "IF -', 'THEN -'". This is not constitutive for any form of understanding since there is no context to the words; not even a human would understand or attach any meaning to these symbols since they interact with no grammatical structure. The function of the Chinese language, and indeed any other language is radically different to English and without a definition, translation, and grammatical structure it cannot be learned. As such, the room's axioms are not grammatical axioms. They differ even from the very thing it is meant to be; a computer.

Axiom 3: Syntax by itself is neither constitutive of nor sufficient for semantics.

Searle's third axioms is most problematic since it is what the Chinese Room intended to prove. Unlike the previous axioms, it is difficult to use as a starting ground for an argument since the Chinese Room must be deemed as a successful argument.

Syntax is said not to be "constitutive" of semantics, yet we must disagree. Due to Searle's frequent referral to some 'biological phenomena', and mental states, in the biological sense it refers to an enzyme or enzyme system which is produced continuously in an organism. While this is certainly reaching to an extent, Searle under this axiom refers to a difference between formal elements and phenomena of intrinsic content, yet contents one may define both syntax and semantics differently. As we know, without a constructive definition of our

primary terms, axioms cannot be constructed since they fail.

He follows with Conclusions 1-4:

Conclusion 1: Programs are neither constitutive of nor sufficient for minds.

"Firstly, I have not tried to prove "a computer cannot think." Since anything that can be simulated computationally can be described as a computer, and since our brains can at some levels be simulated, it follows trivially that our brains are computers and they can certainly think" (Searle 1990a: p. 27)

Rather conflictingly, Searle's claim here appears to be in direct contrary to the point of the Chinese room argument where the intention was to demonstrate that machines cannot think (Axiom 3) since simulation by symbol manipulation does not put thinking to be equivalent to formal symbol manipulation.

He also claims not to have intended to show brains to be systems to think, and that there may be other systems in the universe capable of producing conscious thoughts. In a rather bizarre manner, Searle writes that "we might even come to be able to create thinking systems artificially" (Searle, 1990a: p. 27).

Conclusion 2: Any other system capable of causing minds would have to have casual powers (at least) equivalent to those of brains.

Searle provides and illustration of this conclusion by writing: "if an electrical engine is able to run a car as fast as a gas engine, it must have (at least) an equivalent power output." (Searle, 1990a: p. 29) The comparison between power outage of a vehicle versus the capacity of a brain vs. a computer is foolish; the two are in no way comparable.

Firstly Searle's uses the example of a gas engine to electrical engine where there exists a measurable power wattage between the two engines. Since the two do not work together, Searle use of this as a comparison between equivalent systems is a weak method for counter. Searle writes of this conclusion that is "says nothing about the mechanisms (of the equivalent systems)"

Immediately he turns our attention back to cognition, again clarifying his position that cognition is a biological phenomena. As is clear, this is not necessarily wrong; there is undoubtedly a problem with the use of language and the constant need to define all terms, something which everyone accepts, but there appears to be a slight error in his logic.

As a biological phenomena, and mental states and processes are caused by minds, Searle does not believe this phenomena to be exclusive to biological systems, but does imply that "any alternative system, whether made of silicon, beer cans or whatever, would have to have the relevant casual capacity equivalent to those of brains." (Searle, 1990a: p. 29) This allows for him to derive conclusion three.

Conclusion 3: Any artifact that produced mental phenomena, any artificial brain, would have to be able to duplicate the specific causal powers of brains, and it could not do that just by running a formal program.

Of *Conclusion 3* Searle only writes he can derive conclusion 4, as such we may as well consider both *conclusion 2* and *3* together as logical extensions. We have mentioned the error in Searle's logic and it is as follows.

Searle states that from the biological phenomena of cognition it does not follow that only biological systems could think, yet does imply that any other system must have the same casual powers. As Searle has prior argued, machines lack these casual powers due to these powers being a product of biology (Searle, 1980). *Conclusion 3* states that any additional system would have to duplicate these casual powers, but could not do so via the running of a formal program.

Returning to Searle's error, he claims cognition to be biological phenomena, yet it is not only biological systems which can think. If and only if cognition is a biological phenomena, then only biological systems must have the capacity to carry our cognition. If we assume cognitive action to be a product of consciousness then we can say that if a system exhibits a cognitive action then it is not conscious since only biological systems are capable of cognition.

So if the biological phenomena is not exclusive to biological systems, then it is not true

that only biological systems can exhibit cognition. The correct set of casual powers are attributed to the biochemistry of the brain (Searle, 1980), an therefore a product of the biological brain.

Machines can think because we are machines, but we are a specific kind of machine, and we can think because we have a biological process. Following Searle's logic, only biological 'machines' can think; this (Searle, 1990b) therefore contradicts his earlier point (Searle, 1980).

For an artificial brain to exhibit these casual powers it would require some form of biological process, which Searle has prior rules out in his original paper thus making his point in conclusion 3 redundant. We see further issue when Searle presents conclusion 4.

Conclusion 4: The way that human brains actually produce mental phenomena cannot be solely by virtue of running a computer program.

This here affirms the following of consciousness (Searle, 1992: p. 97):

"[It] is a causally emergent property of systems. It is an emergent feature of certain systems of neurons in the same way that solidity and liquidity are emergent features of systems of molecules. The existence of consciousness can be explained by the causal interactions between elements of the brain at the micro level, but consciousness cannot itself be deduced or calculated from the sheer physical structure of the neurons without some additional account of the causal relations between them."

Earlier in *Rediscovery of the Mind* Searle breaks down all entities to be made up of smaller, and smaller entities until we reach the atomic level. The collection of which forms a system-like base for interaction and functional operation - only should it be required or a property of the object - so we find the brains' cognitive action to be a product of reductionism down to the neurons, and as (Searle, 1992: p. 97) explains, conciousness is the result of casual interaction between elements of the brain at the micro-level.

This is a clear incompatible with the second and third conclusion since any other system which holds the product of thought, must hold the phenomena of cognition.

While (Searle, 1990a) does pre-date *Rediscovering*, it is a continuation of Searle's writing in *Minds, Brains, and Programs*. His theory of biological naturalism and these arguments, - along with those found in *Minds, Brains, and Programs* - are highly incoherent. Despite his frequent insistence that he is not a dualist, Searle does have a dualism between the brain and the mind. This leads to his forth axiom: *Brains cause minds.*

The issues here are fairly self explanatory since the axiom is the logical conclusion. It's not necessary to provide an extended critique into the axiom *per se* but there are extensions which are worth noting.

Searle began with a disclaimer that the room is not against the goals of AI since the room does not limit the level of intelligence which it can display. While he does not give specifics, the use of the Church-Turing thesis provides the answer[3].

3.3 Simulation and Duplication

In relation to $B \leftrightarrow M$, simulation and duplication are more evident but the precise nature of the distinction between is not so clear. Consider again the weather simulation (Searle, 1990), in this instance it is clear that a simulation of weather is only a simulation, in the same way a simulation for digestion is not a duplication of the actual physical event. Searle's reasoning for this is perfectly sound, but the extension into thinking is a sudden leap since it is no longer dealing with a physical event.

It seems sound but causes problems to emerge. The first and most obvious is that there is a difficulty determining between simulations. Baudrillard (Baudrillard, 1981) argues similar that a simulacrum is not a copy of a true thing, but becomes real because of its likeness. Earlier Nietzsche argued - without the term "simulacrum" - that the use of language to describe our sense experience produces a distorted copy of the real.

As part of Chalmers' definition of conscious actions, the reportability of internal states is defined. If one proceeds with Nietzsche's line of thought, consider a human being is exposed to a series of lights of a random wattage between 30W and 120W and asked to report on

[3]This shall be explained in §.5.1.

them as to the brightness. Naturally distinctions will be made between those wattages of the greatest range but between 45W and 47W would be harder as the senses don't register small changes with as great an accuracy as they would between 45W and 85W. This is the same for the use of an sensory glove which allows patients who have lost sensory receptors.

In these examples, machines can measure and record this sensory data highly efficiently by measuring the photons received or the pressure exerted. By having a human brain report on this, it will be less accurate thereby giving a relatively distorted version. The distorted version, to Searle, is the real version since it is observed by a brain versus a machine. The reason this has significance is because it is a version of the brain simulator reply as found in Russell & Norvig (Russell & Norvig, 2003). While not exact, the brain is still conscious and measuring a real phenomena, just as a computer is; there is no simulation or duplication here since a real phenomena is recorded.

The human brain knows its *program* for the production of an output[4] in the same way as a machine does. By 'program' it is meant that the brain has knowledge of, and how to utilise, a problem-specific process by which a solution may be given where each solution is a response of $\Sigma \in \Sigma^*$.

This internal process is continuously at play, but it is perhaps impossible to really understand it[5]. When we speak of mathematics, we cannot refer to something in the same way as we can with language. If we have four apples, the apples do not have the property of 'four' since they are singular and have no property of number since the having of a number property will depend on if there are other objects of the same grouping - a grouping which may be anything defined - and how many of those there are.

Our thought process on mathematics cannot just be grounded, so how do we know we *know* our programming? We don't. We only know that we have it and that it works because we have tested it by solving number problems, or answering questions. In the case of the room, it has a program which it follows, therefore in a sense 'knowing', yet Searle maintains

[4]We generalise 'output' to be any action from physical such as movement or speaking to the mental such as making a judgement or reasoning.

[5]Particularly introspectively

that the computer follows this without understanding. It is never made explicitly clear if the room - or exactly what constitutes the room; where is the line between the room, the program, and the man? - understands the program but one is inclined to say it doesn't[6]. I wish to present two forms of knowing; the first a logical knowing, and secondly a phenomenal knowing. The significance of this knowing is that to Searle, a philosopher working in the philosophy of mind, the internal state of the mind is critical. Biological naturalism has been mentioned but the importance of this to Searle must be acknowledged before continuing. It is this theory of biological naturalism which separates minds from machines.

3.4 Conclusion

Here the distinctions which can be made between Searle original Chinese room argument (Searle, 1980), and the discussed revised axioms, demonstrate the fundamental flaw in Searle's argument. The reason for the fully breakdown of these two arguments serves two purposes; those which were given under our abstract and introduction.

First, is the basic failure which stems from a reductionist approach towards mechanism and the broader CTM argument. Our purpose here is only to respond to Searle as opposed to presenting an analysis of *how* it fails. That shall follow shortly in §5 where it shall be opened out to encompass the more precise problems which are encountered once the room is followed and put into practise.

Under the functions of the brain, Searle's argument presents itself as holding a close link to the Lucas-Penrose argument. It is necessary that attention be diverted from the Chinese room (Searle, 1980) prior to continuing towards the Gödelian arguments since Searle provided no context of how to breach into the Church-Turing thesis, despite his mentions.

Our secondary reason is of greater significance due to the attention towards Searle's axioms (§3.2). Here the purpose is to present a fully critique of Searle's revised argument without having to resort towards a critique *from* the point of an alternate argument, as is often the case with the Chinese room (Searle, 1980). Here the use of the CTT result and

[6]Based on Searles' writing and argument in general.

the overall proposition of Brains/Minds equivalence is introduced in a manner such that the shift towards the Lucas-Penrose argument is provided with context. Due to the significance of Searle's argument in the Computational theory and mechanism, it would be a mistake to avoid this.

What shall now follow is restricted to formal systems and automata; that which underlies Searle's argument. It is my intention that this following chapter shall highlight how fundamental Searle's omissions are.

Chapter 4

Formal Systems and Mathematics

4.1 Chomsky Hierarchy and Automata Theory

Automata theory deals with abstract machines and automata (a self operating machine designed to follow a predetermined sequence of operation). Closely related to formal languages, an automata is a finite representation of a formal language, which may be an infinite set.

These automata are frequently classified by the class of languages they may recognise which may be illustrated by the Chomsky hierarchy. Describing the relations between the languages and the logic, Chomsky's containment hierarchy, first described in 1956[1], contains 4 types of grammar given as Type-0 - 3 grammars.

This essay will not focus on all of these grammars since they will not be relevant, yet I feel it necessary for this to be included so that the reader may refer back to these at any given time. We give the type- grammar, its language, and its automation. Chomsky's automata will be here introduced and defined before we return in §4.

Both the computational theory of mind and the mechanist claims deal with automata yet it is rarely described as being so. The reasoning for this is perhaps that automata deals with abstract mathematical machines versus CTM and dealing with the mind. In being built around these theories of automata, both doctrines hold the operations of the brain to be computable. By 'computable', it is essential that it be clarified that 'computer' and all variations concerning operations of the brain, refer not to modern computers which we use day-today, but to the operation of computing;

[1]. *Three models for the description of language*, Chomsky, 1956.

4.1.1 Concepts of Automata

Much of automata theory will not apply, so we shall only detail those parts relevant. We have already stated S to denote *set*, and F denotes a certain function. Our core concepts are the *alphabet*, *strings*, and the *language*. General notation of lower-case symbols at the start of the alphabet denote symbols, and letters at the end *x, y, z* to denote strings.

Alphabet

Σ denotes a finite set of symbols called an *alphabet*. Any *alphabet* is a finite, non-empty set of symbols.

There are a number of possible alphabets which may be constructed with the most common being:

1. $\Sigma = \{0, 1, 2\}$, a binary alphabet,
2. $\Sigma = \{a, b, ...z\}$, the set of lower-case letters.

As with any system, or language, it is constructed from a finite set of symbols. There is nothing more to the alphabet than these basic symbols which form the language.

We have provided 2 sample languages each denoted by Σ. Each automata will be built on one alphabet. If two alphabets are used where each will denote the same symbol, such as if we us binary representations of letters and numbers, then we would have a pointless exercise. These two Σ alphabets only serve the purpose of illustrating the different types.

Strings

A *string* is a finite sequence of symbols from some alphabet such as 110101 is a binary string of a $\Sigma = \{0, 1\}$, just as 'go' is a string of $\Sigma = \{a, b\}$.

Strings are additionally classified by their length, with our binary string being a string of 6, with standard length notation being $|w|$; $|110101| = 6$.

In the instance of binary, it is accepted in a colloquial nature to say that our string is a string of 6 symbols. This is not strictly true since there are only 2 symbols: 0, 1. Length

is not then the number of symbols, since this denotes *individual* symbols, but the total collection of *all* symbols in the string.

Language

A *word* is a string of the symbols of Σ, and Σ^* is a word of the alphabet of Σ. We consider Σ to be finite, with Σ^* additionally being finite.

Each $\Sigma^* \in \Sigma$ is part of the language L of the machine M. Each case of M can be viewed as a physical production of the function of some formal system or automata machine; an abstract machine.

Operation of Automata in Relation to Formal Grammars

Each type has a set of production rules, or constraints of the following:

a = terminal;

α = terminal, non-terminal, or empty;

β = terminal, non-terminal, or empty;

γ = terminal or non-terminal;

A = non-terminal;

B = non-terminal.

Type-0 Unrestricted grammar: *Turing Machine*

Constrains:

$\alpha \rightarrow \beta$; no restriction.

Definition:

Each automata and finite state machine (FSM) can be defined as an n-tuple machine which satisfies the basics conditions of $M = \langle Q, \Sigma, \Gamma, q_o, \delta, F \rangle$[2] where

[2]Symbols from Hopcroft and Ullman (1979: p. 148).

Q, is a finite, non-empty set of state;

Σ, is a finite set of symbols called the *alphabet*;

q_o, where $q_o \in Q$ is the *initial state*;

F denotes the *accepting state*.

$\delta : (Q \setminus F) \times \Gamma \to Q \times \Gamma \times \{L, R\}$ is a partial function called a *transition function* were L, R[3] are left shift, and right shift[4]. Should δ be undefined for the current state and current tape symbol, the machine will halt.

Language Recognised: *Recursively enumerable*

Type-0 unrestricted grammar included all formal languages where no restrictions are made on the right or left side of the grammars production. These are capable of generating all languages which can be recognised by a Turing machine.

We have already provided a definition of a Turing machine, and there is little difference here beyond the basic function. A grammar is given as being regarded as a device which enumerates the sentences of a language (Chomsky, 1959). The language L is a collection of finite sentences each of finite length, constructed from the alphabet Σ.

Type-1 Context-sensitive : *Linear-bounded automata*

Constrains:

$\alpha A \beta \to \alpha \gamma \beta$

Definition:

Abbreviated as *LBA*, a linear bounded automata[5] is a restricted Turing machine. Originally developed as a model of an actual computer, rather that a model of computation, it is a *non-deterministic Turing machine* satisfying the conditions that its alphabet contains two special symbols serving the purpose of right and left endmarkers, that transitions my not print other symbols over the end markers, and that transitions don't move left or right

[3]Italics removed.

[4]There is and N variant which allows no shift, but this isn't relevant.

[5]Its plural term.

of these endmarkers (Hopcroft, Motwani, Ullman, 2007).

Computers are finite machines; any *LBA* must have a finite tape the length of its output, and has restrictive storage and memory. As a Turing machine has no endmarkers, it can have no memory or output restriction. Any *LBA* is restricted through the use of endmarkers which constrain the output length and therefore determines its memory length.

Language Recognised: *Context-sensitive*

Linear bounded automata are *acceptors* for the class of context-sensitive languages (Hopcroft, Motwani, Ullman, 2007; 225) Acceptors, or recognisers and sequence detectors, indicated whether or not the received input is accepted. At each state the input from the alphabet is either accepted or not accepted until all inputs of the string are accepted.

Γ of Σ is either found or not found, and the machine continues along these steps until the string is complete. At step 1, the machine starts, it progresses on the input 'H' of $\Gamma \in \Sigma$ to step 2 where 'H' is found, from here it processed to its next symbol. Assuming each symbol is found, this repeats on until it produces its output, in this case it could print 'Hello'.

Type-2: Context-free: *Non-deterministic pushdown automata*

Constrains:

$A \to \gamma$

Definition:

Pushdown automata (PDA) are used in theories about what can be computed by machines. While more capable that finite-state machines, they are not as capable as Turing machines (Hopcroft, Motwani, Ullman, 2007).

Type-2 is recognised by *non-deterministic pushdown automata* which can recognise all context-free language versus *deterministic pushdown automata* which recognise only deterministic languages which are a proper subset of context-free languages (Sipser, 1997: p. 102).

Standard notation applies where Γ^* denotes a set of strings of the alphabet Γ and ε denotes and empty string:

$M = (Q,\ \Sigma,\ \Gamma,\ \delta,\ q_0,\ Z,\ F)$ where

Q is a finite set of *states*;

Σ is a finite set which is called the input alphabet

Γ is a finite set which is called the stack alphabet

δ is a finite subset of $Q \times (\Sigma \cup \{\varepsilon\}) \times \Gamma \times Q \times \Gamma^*$, the *transition relation.*

$q_0 \in Q$ is the start state;

$Z \in \Gamma$ is the initial stack symbol;

$F \subseteq Q$ is the set of accepting states.

Language Recognised: *Context-free*

The formalism of context free grammars from Chomsky gives them as being *phrase-structure grammars.* This provides a simple and precise mathematical mechanism for the description of the methods by which smaller phrases build a larger phrase. Chomsky referred to this syntactic description as “immediate constituent analysis" (Chomsky, 1956: p. 116) where this grouping of phrases into smaller and smaller phrases until a general morpheme is reached (Chomsky, 1956; p. 116). These phrases are classified as ‘noun phrases’ (NP) and ‘verb phrases’ (VP)

Type-3: Regular: *Finite state automata*

Constrains:

$A \to a$ and $A \to aB$

Informal Definition:

A finite-state automata (FSA), at its simplest, it the easiest to understand. The most common example is a coin operated turnstile. Beginning in a **locked** state, the input of a coin unlocks the machine and then pushing the arm completes this rotation through the **unlocked** state before reverting the machine to its **locked** base-state. An even more basic form is a light switch. (Hopcroft, Motwani, Ullman, 2007) write the critical distinction

among classes of finite automata is between deterministic and nondeterministic states; the automata cannot be in more that one state at any time and the automata may be in several (Hopcroft, Motwani, Ullman, 2007: p. 37).

Nondeterminism does not let us define any language which cannot be defined in deterministic languages (Hopcroft, Motwani, Ullman, 2007: p. 37). Deterministic and nondeterministics languages can be used in conjunction. A nondeterministic language allows one to program solutions to problems by the use of higher-level languages. This would not be possible in deterministic languages, and from there the automata compiles this into a deterministic automata.

Of these automata, there are several conditions which they must satisfy;

1. The system must be describable by a set of finite states,
2. There must be an initial state for the system. That is to say is its 'start' state must be describable.
3. There must be a finite number of inputs which can trigger the transitions between each state and the reversion to the original state.
4. There must be an action which causes each state and the behaviour of said state must be describable by the relation between states.

(Hopcroft, Motwani, Ullman, 2007) give a *Deterministic finite automata* (DFA) as consisting of:

Q, a finite stet of states.

Σ, a finite set of input symbols.

δ, a transition function between states.

A *start state* denoted by q_0 where $q_0 \subset Q$.

A final accepting state F, where $F \subset Q$.

Of an finite-state automata A, they give:

$$A = (Q, \Sigma, \delta, q_0, F) \tag{4.1}$$

Language Recognised: ***Regular language***

Regular languages are exactly all those which can be decided by an FSA. Chomsky defines finite-state languages as being those which are a finite or infinite set of strings or symbols which can be generated by finite set of rules (Chomsky,1958; 91). Of these strings of the language L, not all are spoken. Chomsky's example was the nonsensical "Colourless green ideas sleep furiously".

4.2 Meaning and Form

The relation between the form of the mathematics and the direct meaning of it is seldom mentioned. It is often not really required to be asked, and 'meaning and form' is rather ambiguous, I confess.

D. Hofstadter addressed this (Hofstadter, 1979) where he posits an isomorphism between language and mathematics which allows for an understanding of the form. This was achieved via the pq- system[6]. It's construction is simplistic and concerns mathematical notation with the explicit purpose of demonstrating as to how meaning emerges. I do not wish to focus or even talk at length of Hofstadter's system, but I do feel that it is relevant to discuss this briefly and provide the problems with it, particularly those which would be highlighted by J. R. Searle.

4.2.1 The pq- system

Hofstadter introduces a formal system known as the pq- system which is comprised of three different symbols: 'p', 'q', and '-'. Proposed in his 1979 work *Gödel, Escher, Bach: The Eternal Braid* (*GEB*), the pq- system begins as being composed of only three symbols and a basic decision procedure. The system is made more complex gradually, yet always maintains

[6]Verbally 'pq *hyphen*'.

a wonderful simplicity and asks the important question; where does meaning lie in strings? At first glance it is apparently meaningless, but the symbols are revealed to contain meaning due to the form of the theorems in which they appear. This is given as being an isomorphism of language by Hoftstadter (Hoftstadter, 1979).

Hofstadter's intent here is to show how even the most primitive strings can contain meaning. Hofstadter acknowledges this to be of no importance to mathematicians or logicians and to be just a simple invention of his. Indeed, this will serve little benefit in this thesis, but I wish to include a brief commentary on the overall ideas of the system as it provides a wonderfully simple and inclusive system for the demonstration of how such strings are not inherently meaningless, while acting in direct opposition to Searle's claims since we with Hofstadter's system we engage in 'meaningless symbol manipulation', yet still produce meaning, and allow the string to infer that meaning.

Our rulebook must be highly detailed and complex so as to avoid overlap or possibly of error, as with any formal system. "A mere description of the axioms might characterise them fully and weakly", and each axiom must be defined such that there is "and obvious decision procedure for axiomshood of a string composed of p's, q's, and hyphens" (Hofstadter, 2000: p. 46)

4.2.2 The Meaning of the Strings

There are two fundamental questions raised from Hofstadter's system which are of importance to our inquiry concerning the isomorphism and of active vs. passive meaning: Are the isomorphisms between language perfect? and what is the difference between symbolic representation and and meaning within human language?

There is a third question which is to ask can *all* theorems be represented by a set of p's q's and hyphens? In this case, the answer is no. Allow the pq- system to prove theorems of natural number $\mathbb{N}$. We may write

- p - - q - - -

$$1 + 2 = 3$$

since there is no deviation from the symbols. Beyond basic addition there aren't enough symbols to accurately represent all true theorems of $\mathbb{N}$ and be able to discern all non-theorems of $\mathbb{N}$. What are the inherent limitations of representing knowledge through a formal system? If we consider the mechanist claims concerning atrophic mechanism, the recognition of ones own consistency is called into doubt through Gödel's theorem. From a formal system of first-order logic it seems not to be possible to represent all outputs as being purely mathematical.

Hofstadter's system is unique and possibly the most primitive example of a workable[7] formal system of a finite alphabet, yet it is not without problems. As complex organic systems - or simply 'brains' - we are easily able to follow the system and infer the meaning contained within since we posses natural language abilities and mathematical thought but can this be true of artificial systems?

4.2.3 Bottom-up vs Top-down

One starts with a collection of axioms (the bottom) and from there works upwards deriving all provable theorems from the set of axioms. For this to effectively work, one must then be able to follow the theorems back down from the top to reach the basic axiom.

DEFINITION 1:

'xp - qx - is an axiom, whenever x is composed only of hyphens.'

Here x must stand for the same in both occurrences.

The rule of production is that the statement establishes a "casual connection between the theoremhood of two strings, but without asserting the theoremhood for either one on its own" (Hofstadter, 2000: p. 46). The decision procedure is that the first two hyphen groups must add up to the third.

[7]Most formal systems belong to abstract mathematics.

RULE 1:

Suppose X, Y, and Z all stand for a particular string containing only hyphens. And Suppose that XP Y PZ is a known theorem, then XPY-QZ- is a theorem.

These theorems currently operate as basic addition, yet can also serve as a statement such as '1 plus 2 equals 3'. Hofstadter states to have utilised P and Q to remind the reader of 'plus' and 'equals'.

- - p - - q - - - - is a theorem since 2 plus 2 equals 4, whereas - - p - - q - is not a theorem since 2 plus 2 does not equal 1.

4.3 The Problem of the Isomorphism

Hofstadter's intention was to show how meaning can arise from a formal system. Of the string '- - p - - - q - - - - -', Hofstadter gives this as a theorem since 2 plus 3 is five. This may also be read as a statement whose meaning is that 2 plus 3 equals 5; Hofstadter writes this to be an odd notation but it is quiet clearly a mathematical statement as much as it is a theorem.

This isomorphism is described as between a discovery of meaning when two complex structures are mapped to each other which is what creates the meaning according to Hofstadter.

For the human brain, this is all perfectly fine since the process described in the meaning of 2 plus 3 equals 5 demonstrates how we infer a meaning from a simple mathematical notation.

J. R. Searle's work[8] presents a fairly rigorous obstacle to the notion of artificial systems being able to truly understand meaning[9] but it is not clear as to which concepts Searle is referring to. Semantic concepts of physical objects are referred to and understood by their relation to the world through the observer[10] but this is not so with mathematics; we cannot point to a number, only a physical representation of a number. Here the object

[8]I refer here to his Chinese room and - albeit less so to - biological naturalism, both of which have almost become his lives work.

[9]This I shall detail further in §3.

[10]See Wittgenstein, 1951

has no concept of the collective number only the singular; i.e. three apples has no concept of the number 'three', each is singular and therefore described as being 'one apple' where the collective sum of all apples referred to is described as being 'three apples' with each remaining singular since they may also be describes as such: $A \supset a_1, a_2, a_3$.

Clearly this works best for basic mathematics of numbers where we can point to or relate to objects of that specific value; those of the properties of a natural number. Mathematics would therefore seem to be a meaningless symbol manipulation where no real understanding of the actual form is ever considered - few ever delve into meta-mathematics - so is mathematics the exception to the rule concerning intelligence and our language?

There is a difference between *doing* mathematics and *understanding* mathematics. A common joke among mathematicians is that it works because "it just does." If we measure the mass of an object and the acceleration then we can easily find the force, as so-on-so-forth.

This critique is not new. Abelson's Open Peer Commentary on *Minds, Brains, and Programs* in its original publication made a series of highly similar claims. Writing of the basic arithmetic with a child learning to count: "When a child learns to count, what does he do except apply rules? Where does "understanding" enter" [p.424]. Obviously meta-mathematics is of little concern to many, least-of-all the child. He further states that many don't really understand the "transcendental number e, or economic inflation, or nuclear power safety."

Of rule following, Abelson writes:

> "Searle is misguided in this criticism in at least two ways. First of all it is no trivial matter to write rules to transform the "Chinese symbols" of appropriate answers to questions about the story" (Abelson, 1980: p. 424)[11].

Knowing where manipulation becomes understanding is still something of a mystery, yet at the same time it ought not be of great importance since we function with manipulation perfectly well.

[11] In peer-review of *Searle, 1980*

As we have been alluding to - but never explicitly stated - Searle's use of Chinese is that is a symbol based language, which I wish to comment on briefly.

Searle takes an almost reductionist approach to machine intelligence/AI, and to language which doesn't work once it is given direct thought. Concerning outputs of a machine, Shapiro asks how they would be given, and what outputs they would be. Indeed there is a vast collection of questions concerning how these outputs would be represented but that is getting ahead of ourselves.

Hofstadter's pq- system begins basic, and in a sense it is basic; there is nothing complex about it, it requires no understanding of formal systems or of complex (meta)mathematics.

These strings can - and Hofstadter does [geb]- be expanded upon and made more complex yet the basic arithmetic form allows for rapid and simple solutions to be given. If we imagine the CR to read these symbols and produce answers it is no different from the human brain doing the same, as we have just explained. We are required to make this more complex and ask further questions of Searle's machine. At this basic level this is rule-following which Searle would see as lacking some 'internal states', or view mathematics in this nature as being calculation which he would attribute similar arguments to. This is problematic, and not the last time we will discuss this.

By taking language as symbol manipulation, the process of thought is reduced to a rule-following state. It appears to make sense since for language to work there must be rules for its production (Chomsky, 1957) since some sentences may be syntactically well-formed, but still grammatically incorrect. By taking it beyond rule-following for grammatically correct - it is assumed that the rule-book is *always* correct; Searle gives a semantic understanding as being critical, yet he fails to extend the capabilities of the machine beyond providing an answer to the question it receives: of course this system is not intelligent.

Mathematical workings form an almost paradoxical counter to Searle in that it conforms to Searle's model of the room, while demonstrating its failing; the line between understanding and the following of the method is not so clear-cut.

There is certainly a validity in what Searle says and this is propped up by Chomsky.

The manipulation does not work, and it cannot be guaranteed. Since they would have no way of knowing what was asked of them, it cannot be accepted that the rule-book would provide the correct answer. For example, if the question read “How is the weather today?", the person would not be able to answer since they do not know how the weather is outside their room and they would not know what weather condition was which.

4.4 The Location of Natural Language

From these automata, we can see how a number of automata process different languages and states. Within computer science this has vast application, but where natural language fits into all of this is not strictly clear. It is only of increased difficulty when one attempts to unify this with artificial intelligence. In speaking of AI, once again, our parameters must be clearly defined.What we are dealing with here are abstract formal systems and asking questions no different from Gödel and Tarski with what can we know and prove from a system; it relates to the systems limitations.

Chomsky (Chomsky, 1957) pointed out that English is not a regular language, and the inquiry into language through automata and formal systems demonstrates certain limitations of AI in a concise and logical manner.

A question that can be ask from this concerns the Chinese room argument. Rather than ask if understanding depends on a mind, it is of more use to discuss the hypothetical machine.

As it deals with language from an alphabet, strings and symbols, and rules, it acts like an automata. Searle could not deny this; he refers to his room as being a universal Turing machine [ref]. The use of an automata-like machine in Searle’s later works shifts his argument from rhetoric and a theory of mind into a more grounded argument build on computability.

Within the philosophy of artificial intelligence, a number of problems persist. As we have said, this thesis is concerned only with a very small number of these problems. We are not concerned with the human side of AI; can machines feel, have emotions, or how can we make machines act ethically. These ‘more-than-machine’ machines - while they have their place -

are vast open problems with application to autonomous machines, anthropic robots, or more general AI computer systems which could be implemented in hospitals to allow for the mass sharing and access of data by doctors to improve patient comfort and treatment.

Our focus concerns, at its core, formal systems. What we intend to show is that current literature such as Searle's Chinese room, and the L-P result fail to provide a compelling argument against the potential of knowing agents. A number of computational problems persist, that is a fact. The application of number theory onto formal AI systems demonstrates there to be things a system cannot know, such as questions related to consistency, and knowing everything about its nature. We shall discuss these is much further detail but this serves as our beginning. Where these are utilised by Searle, Lucas and Penrose for the explicit purpose of showing AI to be impossible, we shall claim that these problems do not demonstrate the futility of AI research, only that there are certain limitations which inevitably will arise in AI systems. Concerning Searle, I argue that his theory of biological naturalism which takes precedent over his logic is not a sound theory of mind which solves what he wrote to be the problem with Cartesian dualism.

The central argument towards formal systems is built upon the Gödelian argument, properly formulated independently by J. R. Lucas and R. Penrose. The base-argument concerning the Gödel sentence contains a highly notable problem: if systems can be independent[12].

Much has been written on Searle, Lucas and Penrose's writings, and they themselves have provided extensive responses to these critiques. Certainly Cartesian dualism can be seen as being among our central inquiry, yet this isn't wholly the case. As much as one may re-evaluating Cartesian philosophy, our study extends beyond the mind-body/machine debate and into the realms of knowability and self-awareness. This debate is heavily intertwined with mechanism where we see Hobbes' and Descartes philosophy of the mind and the nature of artificial systems which provides an analogy for explanation of the biological system.

Before we reach this point, we begin[13] with Chomsky's 1957 work *Syntactic Structures* which began a revolution in cognitive philosophy and a shift into more formal methods of

[12]See §2.4

[13]This serves as an introductory measure.

research concerning the philosophy language. By attempting to provide a formalised theory of linguistic structures, the language becomes more mechanical; it no longer needs to be thought of in terms of a natural and adaptive manner which cannot be pinpointed. Much of Wittgensteinian linguistic philosophy is against this where language is see as a game: "consisting of language and the actions into which it is woven" (Wittgenstein, 1953). Concepts don't require concise definition in order to be meaningful, and that forms of language are connected by a set of family resemblances.

There is a distinct difference in that Wittgenstein's language is centred on communication between A and B, while Chomsky's focus in on the structure and formation on the language and its syntax in a formal manner where we can see language as a process of symbol manipulation of the object and actions into the form.

(a) X $\rightarrow$ Y $\rightarrow$ Z.

(b) The man $\rightarrow$ kicked $\rightarrow$ the ball.

Earlier in 1956, Chomsky wrote:

> "The first step in the linguistic analysis of a language is to provide a finite system of representation for its sentences. . . . By a language then, we shall mean a set (finite or infinite) of sentences, each of a finite length, all constructed from a finite alphabet of symbols." (Chomsky, 1956: p. 114)

Chomsky's sentence is almost equal to a mathematical sentence; each is composed of symbols of pre-defined meaning, constructed from its rules of inference and axioms. There is a clear distinction between the understanding of words and of mathematics. Chomsky's now-famous sentence "Colorless green ideas sleep furiously" (Chomsky, 1957: p. 15) illustrated this point in highlighting the specific difference. Both sentences appear to be correct in that they are composed of the correct symbols in that they are true and definable words. Structurally they make sense since they describe actions of something in the object language.

The sentence takes the structure of Adjective $\rightarrow$ Adjective $\rightarrow$ Noun $\rightarrow$ Adverb. This structure is sound, it is the symbols which are incorrect. The determination of its truth-

value; i.e. is it logically correct (it is grammatically correct), is only done if accurate meaning can be extracted from it

Second, how is the language understood? Is there a difference between understanding natural language which refers to objects and concepts which we in the physical world can have experience and of understanding of mathematical concepts and arithmetic which we have no physical experience of. Critically, is this purely representable concept understood in the same manner as natural language.

4.5 Conclusion

The understanding of language is complicated, even within the field of natural language acquisition. The emphasis on a natural language conversation can be seen as stepping too far forward without a sufficient address of the underlying problems which exist towards all formal systems of this nature. These systems are the extent of the the L-P result if the Orch-OR theory of Penrose (Penrose, 1989; Penrose, 1994, Hammeroff & Penrose, 2013) is discounted given that is is not required.

Representations of knowledge within the Chinese room are never given a clear explanation, but if these are utilised, then it is possible to construct a theoretical base for how such a system can learn.

Chapter 5

On the Problems of the Chinese Room

Searle's foundational axiom addresses the significances of the intuition behind Searle's arguments. Chalmers deems each of Searle's axioms to be "superficially plausible" (Chalmers, 1992: p. 5) and his foundation axiom represented a "distillation of the main thrust of the Chinese room argument." Chalmers, 1992: p. 5)

M. Boden's address of Searle's use of formal syntax and manipulation is on of the most direct challenges towards Searle's errors, and the confusing nature of his often changing explanations of the difference between minds and machines, and the significance of what Searle refers to as casual/causal powers. This critique - and many others - don't often focus too greatly on the limitation which Searle applies to the CR-system.

The axioms given (Searle, 1990) don't deal with any form of learning, or set of outputs from its alphabet; little of the CR-system does. Searle places no higher level which the system cannot exceed - thereby limiting its operations - theoretically giving the brain and the machine an equivalence concerning its output functions. By this, it is not meant that consciousness, or any form of casual/causal power is the upper limit since it is not an function of the system but an biological occurrence given as an absolute limit. There limits are not the same. Consciousness is not a function since is not an expression, nor does it define a relation between variables; broadly, it is not definable in the system and its axioms and rules.

Within Searle's argument, one finds that there is a lot going on which calls attention to the original argument and *exactly* what the Chinese room is. It may strike the reader that there appears to be two or more variations of the room and this is correct; Searle presents

two versions of the room under the guise of one room.

5.1 Knowledge Representation and Connectionist Networks

Adaptivity is the core of the issue. Searle's room has no adaptability since it is constructed on pure rule-following and Thagard's objection towards this is the beginning of how meaning can emerge. If we are concerned only with mathematical statements, it can be asked if meaning is ever needed since so long as it works, we encounter no problems. Ultimately this only takes us so far and fails to provide a response to language isolating itself to a small subset of formal/first-order logical statements.

Knowledge representation can be considered as a part of ontological engineering. Connectionist networks can equally be considered in the same line of inquiry since both provide a counter against Searle's symbol manipulation in that they do not deal only with this process. The connectionist networks mapping of the neurons and their functions combined with knowledge representation extend beyond a basic symbol manipulation of corresponding symbols, which therefore allows for a knowledge base which contains a semantic web between concepts. These provide a welcome addition into the intelligent nature of imitation, along with an alternate perspective into psychology and the casual powers which Searle describes as being akin to an innate intuition.

5.2 Towards a Formal Variant of the Chinese Room

Before continuing, I wish to present a formal version of the Chinese room so that it shall be easier to deconstruct and display the problems of Searle Turing model. I wish to be very clear that I do not hold this formal version to be correct, only that I shall proceed with the assumption that the accuracy of the Chinese room and the TM model is unconfirmed. Any errors between the Searle's model and actual formal systems shall be shortly discussed and fully acknowledged.

Recognising a formal language, the language L of M holds a syntactical structure for its production which we call R. L is comprised of its alphabet Σ where each correct string

comprised of Σ forms a word w.

At q_o, M receives an input string of words, w which comprise a correct sentence, W. Since W is correct of both w and R, M may produce an output string in accordance with R.

Since this is a naturally language conversation there are a countably infinite string of words $w_1, w_2, ...w_n$ where only a finite number of W inputs and outputs are correct following R. It cannot be enough to say that W is correct only because it is in R; R must be comprised of a rule which does not just list every correct string. R must contain the rules of grammar G.

We assume this system to be finite since we assume grammar to be a finite set of rules and therefore of a finite structure, which in-turn presents a finite set of input/output statements. As Searle speaks of input and output functions as being a question and a response, the system may not contain w and R may contain every possible correct string of w which produce a correct sentence W, yet make not mention of the singular word, w. If this is the case, and Searle does little to dispel one from believing this to be the case, then W is correct only because $W \subset R$; it requires G. Without G, and every w the system cannot be intelligent.

5.2.1 Subsymbolic Computation

Mizoguchi & Bourdeau (Mizoguchi & Bourdeau, 2016) write of adaptivity that "It comes from the declarative representation of what the system knows about the world it is in." (Mizoguchi and Bourd, 2015: p. 108). Thagard (Thagard, 1986) notes that certain scientific concepts such as black holes and electrons don't correspond to anything which is directly observable (Thagard, 1980: p. 140). Black holes are a radically different from intractable objects The understanding of these concepts comes only through the constructs which the human brain does; these concepts are understood through the use of a specific language and through a network of relations. Any such system which could 'understand' these concepts would be required to learn concepts and build upon prior known concepts and language.

Butler notes the claim on how connectionist networks are not typically symbolic by being syntactically structured - as in the case of the CR-system - meaning the networks are not

a form of syntactic symbol manipulation. he early days of cybernetics and contemporary neural networks from those proposed by W. McCulloch and W. Pitts (McCulloch & Pitts, 1943) which showed on/off formal neurons in feed-forward networks could compute logical functions onto 'neuron' states represented by '0' or '1' where each state could be identified as a neuron of formal truth or falseness (McCulloch & Pitts, 1943; Kaufman, 2012).

The CR-system uses it alphabet to represent knowledge and the axioms represent the knowledge structure, including the conditions and actions; "IF", "THEN" statements. In frame-based systems, knowledge structures are used to represent and categorise object or actions. Common Logic, for example, uses its axiom, which is a sentence that is assumed to be true which allows for others to be derived. An atom is defined as a "sentence form which has no subsentences as syntactic components" (INTERNATIONAL STANDARD, 2007: p. 2), where a sentence is a unit of logical text which is either true or false, and assigned a truth-value. Systems such as these allow for formal representation of the symbols which the system uses.

If any form of consciousness is set aside, and focus is directed solely on a knowledge base and knowledge representation then the CR-system remains fairly unchanged in its basic operation. Containing axioms which allow for additional sentences to be derived is not enough that we can make strong declarative statements about the knowledge representation of the CR-system since from Searle's definition (Searle, 1980) the system does not have this since it only follows an "IF" "THEN" process where each initial sentence is known by the system which holds a corresponding end sentence; the input to output process.

The rule-based CR-system does not mention it is able to create new rules, making its axioms and knowledge sets fixed. Without being able to learn new rules it seems that such a system would not require a knowledge representation between knowledge sets since the axioms and initial-output structure is pre-defined by the rulebook.

Computations at a neuron level are deep-level, verses the high-level of symbolic computations which is deemed too high to be a good model of the brain (Rumelhart, McClelland, & the PDP Research Group, 1986). Smolensky (1988) wrote this to possibly be too deep:

> "The terminology, graphics. and discussion found in most connectionist papers strongly suggest that connectionist modelling operates at the neural level. I will argue, however, that it is better not to construe the principles of cognition being explored in the connectionist approach as the principles of the neural level" (Smolensky, 1988; p. 3)

The requirement for going to a deeper level is given by Chalmers as being that "when viewed at the semantic level such systems often do not appear to be engaged in rule-following behavior, as the rules that govern these systems lie at a deeper level." (Chalmer, 1992: p. 3)

The plausibility of Searle's axioms is established. Any manipulation of symbols without a deeper-level computation towards the actions is sure to prevent meaning from arising since the rule-following manipulation is to high to allow for the meaning since it is dealing with fully formed concepts (Rumelhart, McClelland, & the PDP Research Group, 1986; Boden, 1988; Mitchell & Hofstadter, 1990; Chalmers, 1992.)

5.2.2 High-level Axioms and the Physical Symbol System Hypothesis

Earlier, McCarthy (McCarthy, 1980) outlined programs which represent information of their "problem domains in mathematical logical languages and use logical inference to decide what actions are appropriate to achieve their goals" (MaCarthy, 1988: p. 297) For reasoning, these systems much be able to accurately store facts, reason with a precise language, and have a precise definite of how the the reasoning may be derived.

The atomic to sub-atomic concept symbols are found in the LISP programming language[1] such that they can represent concepts such as penguins by having an axiom refer to the complete concept and each sub-axiom refer to a property of the axiom. The method is akin to bundle theory in describing the collective properties and relations of an object.

The physical symbol system hypothesis (PSSH) formulated by Newell & Simon (1976) which states: "A physical symbol system has the necessary and sufficient means for general intelligent action." (Newell & Simon, 1976: p. 116). Both the L-P result and Searle's result

[1]Developed and developed by McCarthy, Steve Russell, Timothy Hart, and Mike Levin in 1958

contain some form of physical symbol system in a formal system, and a digital computer. In a digital computer, it is possible to interpret each string of 0's and 1's passing through the machine as being information where each definable set of 0's and 1's - such as a binary alphabet - is deemed a symbol.

By having each symbol designated to a specific function it may express anything the system and its rules allows:

> "That is, given a symbol, it is not prescribed a priori what expressions it can designate. This arbitrariness pertains only to symbols; pertains only to symbols; the symbol tokens and their mutual relations determine what object is designated by a complex expression." (Newell and Simon, 1976: p. 116)

By use of the rules of inference it prevents the symbols from holding any "magical properties hidden in their physical realization." as put by Touretzky & Pomerleau (Touretzky & Pomerleau, 1994: p. 2)

The production of new information, or formulas, requires a production of reasoning and as Chalmers (Chalmers, 1992) states, this cannot be done from an isolated system of high-level symbol manipulation. Searle's blanket refusal to acknowledge connectionist networks or allow his rule to implement a low-level network of definitions across the sub-properties of concepts means that each definable concept in the CR-system is independent of each other concept. The relation between only exists in the form of an output production, yet this does not make sense since there must be a lower-level relation in order for the system to answer questions.

I do not mean to disagree with Chalmers' point here, since it is valid. Searle uses only high-level symbol manipulation of formed concepts such as TABLE and PENGUIN, for example. Imagine you are asked how many legs a Penguin has. For Searle's system to take no use of a lower-level computation it must be programmed such that it may answer every possible question concerning all concepts it contains. If we assign the question of how many legs a penguin has to be *Q3n7* the answer would correspondent to *A3n7*, and so on for each separate question. This corresponding method can certainly allow for a system to answer

every question it is asked so long as its axioms will allow it, but it will obviously never understand them because compound concepts are reduced to a single concept. The clearest example is of the term *bachelor*, a high-level concept of the lower-level concepts of *young*, *male*, and *unmarried* yet as syllogism and the analytic-synthetic distinction has taught us, it is not try that all unmarried men are bachelors.

So are descriptive concepts the same as physical? Atomic symbols, which are those high-level concepts of any kind, contain lower-level *sub-atomic* concepts:

BACHELOR: ((unmarried) (young) (male)).

PENGUIN: (Emperor Penguin: (Colour: (black/white/yellow)) (Height: 1.1 - 1.3m)).

Detailed sub-atomic concepts give rise to a complex network of relations between objects and properties. From this a system may differentiate between different breeds of penguin thereby supplying an accurate answer since if the question is *where do penguins live?* the answer varies across breed.

If the system manipulated only the atomic concepts then Searle's argument is fairly convincing since it is not manipulating, or holds any understanding or information concerning the sub-atomic concepts.

Epistemic arithmetic is a modal theory developed independently by W. Reinhardt and S. Shapiro. Interested in the philosophic ramifications of Gödel's incompleteness theorems and required an axiomatic setting for his analysis. Shapiro's purpose was in finding common ground between classical and constructive mathematics.

There is a relation to Searle here; a counter which he would not accept. The counter is elementary. Carlson (Carlson, 1999) begins with Reinhardt's conjecture that a formalisation of *I know I am a Turing Machine* is shown to be consistent with Epistemic Arithmetic, a theory of extending basic number theory of Peano arithmetic. Here the notion of knowability is added.

Implying a positive answer to Reinhardt's conjecture (RC), Carlson's result is "there is a machine which knows it is a machine." (Carlson, 1999: p. 52) Further formalised in EA, the Epistemic Church Thesis, or ECT is a formalization of the statement "For any function I

can compute, I can find a Turing machine which computes the same function." Of knowing, Carlson states that it does not mean "known at that moment", but should be taken to mean "can eventually be known." Carlson establishes the following basic conditions of knowledge for all statements ϕ and ψ:

1. If I know ϕ and I know ϕ implies ψ then I know ψ.
2. If I know ϕ then I know ϕ is true.
3. If I know ϕ then I know that I know ϕ.

These conditions must be both true, and known.

These conditions are formalised in EA, where two possibilities for determining how knowledge is too be represented: as a predicate, or a modal operator. When the conditions are treated as a predicate, a contraction appears. One therefore is forced to represent knowledge by a modal operator. If ϕ is a formula in the language of EA, then "I know ϕ" will be formalised as $K\phi$.

For now, it is necessary to put aside Searle's thought that syntax cannot cause semantics, and deal with Carlson's method of knowing. The significant deference is the function from Searle linguistic. The Chinese character is symbolic in that it is the syntax which refers to the semantics, Carlson's function operates in a similar manner in that it provides meaning and reference to a semantic counterpart, in that the symbol ϕ represents some knowledge. In this instance, ϕ is the syntax and ψ is the semantics since it is inferred from the syntax.

Clearly it cannot be followed strictly that both symbolic functions are equal as syntax and semantics, but the relation they share between themselves as signifier and signified must be given. If, as according to Leibniz's identity of indiscernible the machine and the program are identical; $M \leftrightarrow P$.

Searle's counter towards claims of this nature is as one would expect; that it does not concern the mental phenomena of consciousness. For Searle, consciousness goes deeper than being an awareness, it relates to the total function of the brain concerning awareness,

understanding, and meaning. As one could present the argument that if a computer were to replicate the exact functions of the brain in the exact way that the brain functions; what is referred to as the 'brain simulator' argument. Searle replies that meaning has to do with the "neuro-chemical interactions within the brain, and that this mental phenomena is dependent on actual physical-chemical properties of actual human brains." (Searle, 1990: p. 29)

In continuing along this path, Searle states that the monolingual man looks up the rules in an English language book. We return to this issue of understanding the rules. A direct reading of this argument tells us that we have a machine which is able to follow the process of the rules, that it is able to apply these in the correct instance to produce an output of what it does not understand. If a human is to perform a real-life experiment similar to this, then what happens when they produce the output? There is still - as Searle calls it - a mental phenomena taking place between the interaction of the Chinese symbol to correspondence with the rules to the output. This opens to a multitude of questions which are difficult to answer. Without accurate neuro-imaging, studies, and a whole host of neuroscientific research, these questions cannot be answered, and even then it is not certain.

A novel variation on this argument is the 'brain replacement' scenario where we imagine a computer capable of the exact replication and simulation of one neuron. This would do nothing to alter the consciousness, and if we proceed with this by replacing each single neuron, then we end up with a simulated brain. A number of Searle's critics (Russell & Norvig, 2003) claim that it is not possible to know where consciousness ends and the mindless simulation begins. This critique is broadly a core issue of the problems of consciousness. Searle misses this point almost entirely and rests his argument solely on what we have discussed; if we view AI as a computational task of replication then we shall no doubt encounter problems.

5.3 Searle's 'Turing Problem'

Searle's problems with the Church-Turing thesis are layered. Originally, (Searle, 1980) didn't focus on Turing machines, yet his later works (Searle, 1990a; Searle, 1990b) introduced Turing

machines and the Church-Turing thesis. Regardless of the extend to which Searle applied his Chinese room to these, there are a number of problematic situations Searle unbeknownst to himself, enters into. This is particularly due to his dismissal and response towards the 'robot reply', or any environment-based system.

The basic confusion between algorithms and functions (Searle, 1990b; pp. 22 - 23) and that the room can run any program is clear, and has only been given attention from a select few, most notably by Copeland (Copeland,1992) and Spravak (Spravak, 2007). There is an equally significant flaw concerning the inputs; Turing machines are static.

If the room is to be a formal system, it is not so simply the case that the machine can be receiving external information On Searle's part there is no satisfactory explanation towards how this would work functionally, and the TM model is not adequately discussed. Additionally the use of a TM model does not include a sufficient discussion towards formal grammar.

5.3.1 What it all Means

It would not be sufficient to consider some programs to be able to produce understanding based on a specific architecture which thereby requires Searle to demonstrate strong AI to be false covering *all* programs including those which can be simulated. Where prior I have argued against Searle foundational refutation of strong AI, I shall proceed in similar.

Penrose argued a similar version to *axiom 2*, and Searle maintains that it follows from the Church-Turing thesis that the brain may be simulated by a Turing machines. By utilising a program which can run the steps for the understanding of Chinese, the claim that this does not produce an understanding of Chinese is flawed, as we have prior detailed, but the argument Searle puts forward is that we cannot imagine *any* rule-book which the person could run that would ever produce and understanding of Chinese: "I can have any formal program you like, but I still understand nothing" (Searle, 1980: p. 418)

Jahren (Jahren, 1990), in his defence of Searle's Chinese room, writes: "I am referring specifically to the thesis that a computer can (in principle) have a human mind (i.e., a mind that functions as does a human mind)." Jahren places Searle's argument more in-line with

AI that Searle since he feels that one has no reason to consider biological naturalism and may still present a philosophic/scientific inquire into the machines/brains equivalence problem.

In *Minds, Brain and Programs*, a computer is described as being a machine which can run the steps of a program for mental capacity for understanding Chinese, yet it would not understand Chinese. There is sufficient claim that Searle has based this form of system on a Turing machine (Searle, 1992), thereby meaning the Church-Turing thesis.

It's clear from this that Searle is rigorously against a computer understanding since it is a result of a biological phenomena. By the man not being able to understand Chinese, and run any program for the production of a Chinese output, the room is a universal Turing machine:

> "For our purposes, the Church-Turing thesis states that for any algorithm there is some Turing machine that can implement that algorithm. algorithm. Turing's theorem says that there is a Universal Turing Machines which can simulate any Turing Machine. Now if we put these two together we have the result that a Universal Turing Machine can implement any algorithm whatever." (Searle, 1990b, p. 22 - 23).

Nasuto et al (Nasuto, Bishop, Roesch, 2015) state:

> "[I]t is equally clear from 'Minds, Brains and Programs' that Searle intended the CRA to be fully general - applicable to any conceivable [now or future] AI program (grammar-based; rule-based; neural network; Bayesian etc): *'I can have any formal program you like, but I still understand nothing'*. So if the CRA succeeds, it must succeed against even the most complex 'high-level' systems." Nasuto et al (Nasuto, Bishop, Roesch, 2015: p. 7)

This is computable in the sense of Turing where it can be accomplished by a human using the open and paper method.

Avoiding Searle for a brief moment, consider the rooms symbolic outputs. As a word-based language its output is functions acting on words; *symbolic* functions. Turing's proof

was for *numerical* functions acting on natural numbers, or integers. As such, in the case of language it cannot be acting in accordance with Turing's proof.

Of Searle's argument, it can be formally given as "I have outputs indistinguishable from a native Chinese speaker". The program which enables these outputs does not understand the initial input, nor the output.

Of any alphabet Σ where R is its set of rules, each function ϕ can be given by a Turing machine TM in the language of Σ and in accordance with an algorithm from R. For Searle's room to hold as sufficient, it must be able to compute any operation of TM. If P is given to be the algorithm[2] and ϕ to be a function expressible from R in Σ, the simulate of Σ from R is incorrect since the thesis does not hold that a separate machine Q may compute P, only its function.

For Searle (Searle, 1980) Sprevak gives the following three statements of the Chinese room, which we shall give as being (G):

> The man inside the Chines room cannot understand Chinese stories.
>
> The man inside the Chinese room can run any program.
>
> No program can be sufficient for, or constitutive of, understanding Chinese.
>
> (Sprevak, 2007: p. 757)

Sprevak (Sprevak, 2007) considers the additional proposition that running a program *is* sufficient, or constitutive, for understanding Chinese. He gives the proposition of the man in the Chinese room being able to run any program (G) as being absolute; it is necessary for Searle's argument. both arguments depend on the room being able to run *any* program, or the argument fails to work. Searle states that the room cannot understand Chinese regardless of what program it runs, and since it is a universal Turing machine (UTM), we can conclude that no program is sufficient for the understanding.

The language of our system is employed by the individual in the room; the pen and paper method. In this situation they operate the same as the program. If Searle is correct, then the person cannot understand Chinese no matter how detailed the rule book.

[2]This by extension means its Turing machine.

Searle would be correct but since the Church-Turing thesis does not state that for any algorithm there exists a Turing machine which can implement that algorithm, he is wrong. For Searle, the Chinese Room is a universal Turing machine when it is not. Secondly, this statement of the Church-Turing thesis is incorrect. What it states is that or any computable *function* there is some Turing machine that can reproduce that function. A function and an algorithm are not the same. A function is a set of input-output pairings, so for Searle's argument to work, the two require an equivalence they do not hold. Concerning the simulation, in the context of Turing's theorem, it is technical rather than an exact replication and implementation of the same algorithm.

Since the Chinese Room is not a universal Turing machine as indicated by Searle, it is not justified to state that it may run any program, only those of the set of computable functions. Even if we assume it it be a universal Turing machine, it is still unjustified since it is not supported by the Church-Turing thesis. The more basic o-machines also fails to satisfy Searle's use of the Church-Turing thesis since o-machines are defined both formally and syntactically.

What Searle believes to be asserted by the Church-Turing thesis is not the case, Copeland wrote: "Turing had no results which entail this. Rather, he had results which entail the opposite" (Copeland, 1998: p. 133).

Where semantics and programs are defined syntactically, Searle gives the following: "The development of proof theory showed that within certain well known limits the semantic relations between propositions can be entirely mirrored by the syntactical relations between the sentences that express those propositions" (Searle, 1990b: 23). The difference between the mechanical computer and the Turing human computer is that there is "a program level intrinsic to the system and it is functioning casually at the same level to convert input to output." (Searle, 1990b: p. 23).

5.4 Finite-state Language and Formal Manipulation

The grammar of the language L is finite giving rise to a finite language. The total set of grammars G are finite, where it follows that only a countable set of grammars can cause a countable set of L. L is not a metalanguage, but the total set of all possible languages. Where any grammar of G which forms a language is used, L is finite in what it can express since it is finite in its natural state; constructed of a finite grammar. Each grammar of the set G - denoted as $G_1 . \; . \; . \; G_n$ - is the tool for the construction of all strings of its corresponding language. L_1 only differs from L_2 if there is G_1 and G_2. If there exists a blank slate of no currently constructed grammar, then any constructed grammar G_n will correspond to L until a different grammar constructs L_2 where L becomes L_1, and so on so forth for each additional grammar. In a meta-state we have $G_n \subset G$ where each n-grammar is finite within the finite set G. Since all forms of grammar allow for the construction of a language, it is therefore meaningful to say there exists a theoretical set of countable languages of the set $\mathtt{L}$[3].

The language of the room, which we denote as being L_1, is constructed in the standard manner; as an algorithm of the basic sense where there exists a logical process from the axioms and rules of inference which can produce all available strings.

Inherently problems arise from each argument where solutions can and have been given, but Yee has found the lack of a clear address of Turing machines directly has produced unproductive discussion. There is perhaps a great deal of truth here which one finds hard to ignore given the volume of literature produced surrounding particularly Searle's argument.

While Yee (Yee, 1993) writes of Searle's argument to be an attack only on universal Turing machines, and argument of which Searle wrote to be supported by the Church-Turing thesis. Yee has written of the computational theory that it effectively reflects the view that the Church-Turing thesis covers the mind-brain process (Yee, 1993). If Searle is to utilise a universal Turing Machine, then it must be accepted that the CR-system is deterministic. The failing of the mechanist, and by extension all similar against the computational theory

[3] $\mathtt{L}$ is the total of all possible sets of L: $L \subset \mathrm{L}$

of mind, was in the non-specificity of the machine/computer. The errors of these claims are addressed in §4 and §5.

5.5 Manipulation as Thinking

Boden (Boden, 1988) claims that Searle's misuse of formal syntax, along with the borrowing of the terms 'syntax' and 'semantics' affords him great difficulty (Boden, 1988). 'Thinking' is given as symbol manipulation, and the mind is to the brain as the program is to the hardware. Machines cannot think, since they cannot possess minds or the properties discussed. Minds, Searle grants, are machines (Searle, 1980: p. 422). Of the question if machines can think, Searle writes: "The answer is, obviously, yes. We are precisely such machines." (Searle, 1980: p. 422). Later, he writes:

> "My own view is that *only* a machines could think, and indeed only very specific kinds of machines, namely brains and machines that had the same casual powers as brains. And that is the main reason strong AI has had little to tell is about thinking, since it has nothing to tell us about machines. By its own definition, it is about programs, and programs are not machines." (Searle, 1980: p. 424)

There is no tailored response from Searle here, only the general. The reply is only of limited use since it presumes that 'meaning' bypasses the program. By removing the notion of 'strong' AI and dealing with symbol manipulation we are still left with language versus mathematics. Chomsky has demonstrated the caution one must accept with rule-following for well-formed sentences and Abelson's open question on mathematical ability leave each argument isolated and failing to work together. If Searle maintains that rule-following will not produce understanding since the system cannot *understand* the output's, he cannot be sure that the method employed is correct since he may give a mistake which causes the system to produce incorrect outputs. There is no difference here between this and a mathematical method since if it is not employed correctly it will not work.

If our system is only capable of manipulating symbols, we are given no reason to believe that it employs any form of verification system. Computers are able to produce highly accu-

rate simulations based on massive amounts of data and complex mathematics, thereby giving them a level of understanding. They are doing nothing different than a human mathematician would do since it is built with the same equations as any human would employ. Is the human mathematician performing a more *aware* form of mathematics? Does this matter?

Boden's point of Searle's language borrowing begins to present itself more clearly now. Thagard used "semantics" to refer to the general philosophic concept of relation between symbols and what they represent (Thagard, 1986). While this is understandable, it differs from how it is generally used in computer science.

If semantics is the relation between the symbol and the representation, then in any AI system there still remains the problem of *how* semantics arise.

The subsymbolic manipulation method provides a system with a deeper representation of symbol manipulation, that Searle basic form. Carnap (Carnap, 1950) made the attempt to construct empirical definitions of scientific terms, but failed. Thagard proposed meaning of a symbol to be understood in "terms of the computational mechanisms that led to its construction" (Thagard, 1986: p. 141).

From the axioms, we are able to de-construct the language to extract its meaning. Since our system has complete language from the symbols, it is already constructed and we may say that it is complete so long as it is correct. Our language is composed of symbols, but it is not the collection of the totality, rather a finite string of the n-symbols constructed from the axioms. For a language to be 'complete'[4] it must be comprised of the defined symbols - the alphabet - and constructed from the axioms, along with being correct. The difficulty here is that it must not have a contradiction, yet language can still hold contradictions and be complete: Liar paradox. Instead we say that it must abide by its rules of inference; the strings can function as both theorems and statements since they refer to something, or more specifically *say* something.

We would expect the outputs of the Chinese room to hold some grounded-relation in the world; it must relate to something which we can comprehend and speak. This may seem

[4]I am hesitant to refer the these languages as being complete in the manner we shall discuss concerning Gödel.

Wittgenteinian, and to a certain extent it is.

The rulebook operates much the same as the code of a computer program yet "the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations since the symbols don't symbolize anything" (Searle, 1980: p. 422). Obermeier calls this intentionality a "*conditio sine qua* for understanding", (Obermeier, 1983: p. 340). For Searle, intentionality is a biological phenomena and writes it to likely be dependent on specific biochemistry.

The language of the Chinese Room is composed of symbols. We need not refer to these symbols as Chinese, but only that they are symbols which hold a relation to a word or letter; they refer to something yet they are not definable in a spoken language. We may further state that they are non-phonic and similar to binary.

Each symbols differs from each other that we may say each individual symbol S holds a specific bi-relation to an 'object' of the world. The object does not necessarily have to be a physical thing and may refer to a concept; a thing in the natural language.

If we were to begin with language, we could too easily take one stance and have that confirm or reject Searle's underlying statement. We may view language communication to be structured around the purpose of interaction, in that much of who what we say is irrelevant to the topic, and only serves the point of seeking to invest the additional party in the dialect. This may also mean that when we speak, our individual personality comes through in how we attach a level of emotion to what we say.

If this is how we build our language communication then we may comfortably explain how Searle's argument holds true since the machine will not have an emotional attachment to the language used; its understanding will be formal and built around a network of definitions and relations. Hofstadter spoke of isomorphism within language (Hofstadter, 1979: pp.49 - 51) and between "rule-governed symbols and things in the real world" (Hofstadter, 1979; p. 82) and stated that the more complex that isomorphism is, the more is required from the system to extract the meaning.

With mathematics, we don't need a full understanding of the symbol - in this case the

individual hyphen which represents '1' - and from there we must deal purely with mathematical notations. These notations can be greatly expanded out - which we shall arrive at shortly - but with basic arithmetic we must only operate in accordance with the rulebook. So what is the difference between our own mental arithmetic and the arithmetic of a machine? Each symbol represent as specific thing such as a number and the rule of addition is not contained within the axiom so a correct response is dependent on our knowledge of it.

Since Searle accepts that machines can work in accordance with rules, we find no issue there. So when we extend this onto ourselves and the machine, we find that we are both following the same rule book. Since there is no other way to interpret a string of mathematics in any way other that what is correct; 2 plus 2 *must* equal 4, or it is wrong.

The main problem with Searle's Chinese room concerns translation and effective natural language bots. In discussing the progress made by such chatbots, little needs to be said concerning intelligence. In the entirety of the papers Searle dedicated to the Chinese room argument, the room's system is never adapted for presented as a hypothetical, and is sometimes presented as a *reductio ad absurdum*.

The adaptation from the basic argument found in Searle's 1980 paper into what it has now become what holds that the best program which could run a program constitutive of understanding, the Chinese room will be able to run since it is a universal computer, are those which I wish to challenge.

In that way it is very much an extension upon the Turing test. As such, it is difficult to construct any response which can follow wholly with the argument since there is much that can be asked, which he fails to include. The most obvious being questions about our own understanding of the language we use. Do we *really* understand each word, or do we understand a clear, concise definition in relation to objects, and the way in which it is used within the context of the given string, i.e., does it contain metaphor, or is it attempting to convey something in a more understandable manner?

This is a basic failing of Searle's Chinese room. While this would not be required of every intelligent agent, Searle's room is built solely on a natural language question and

meaningful understanding of language. Its individual operation eradicates the potential for an exploration of discussion beyond what can be given as symbol manipulation, or even if this operation occurs in the human brain. As with asking if we truly understand, Searle's room has no space for mathematics.

Much is missing concerning first-language acquisition, and Searle includes no notation on research or basic theories. Lidz, et al., concerning early childhood language acquisition writes: "The question . . . is more properly understood as the question of how a learner takes the surface forms in the input and converts into abstract linguistic rules and representations." (Lidz, Waxman, Freedman, 2003: p. 66) The complexity behind language acquisition puts Searle's argument into difficulty since there is little to the system beyond meaningless manipulation as everything beyond is disregarded or not even considered as a possibility which could be explored in a potential system. Rather than construct a working argument against a specific set of theoretical intelligent agents, Searle formulates one against a specific type of natural language based system which can engage in a conversation. Clearly this is an attack on the Turing test, yet misses the point of Turing's paper, and takes a misunderstanding of the Church-Turing thesis.

5.6 The Two Rooms

In he original version, Searle states "I have inputs and outputs that are indistinguishable from those of the native Chinese speaker" (Searle, 1980: p. 418). In this version, it is not the room which produced the output, but Searle, yet this is not correct since the room has the program. Where Searle not in possession of this program - if the program was removed from the room - then the room would fail to output accurate since Searle is incapable of doing so as he clearly states across the paper: "but I still understand nothing." (Searle, 1980: p. 418). Searle proceeds here to state that it is the reason Schank's program doesn't understand the stories since "in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing." (Searle, 1980: p. 418).

This analogy fails since the computer and Searle are not the same: the computer has the

program, and Searle only performs it. Now, while Searle's sees this to not be the case[5], the conclusion cannot stand since Searle does not run the program the same as the computer. The issue of subsymbolic computations concerns the program run by the computer, but not the program run by Searle. Since it is not stated that the room has the sub-atoms within its program, it is likely that Searle processes the exact same symbols.

However, since Searle has no understanding of the Chinese atoms it would be equally impossible for him to understand the sub-atoms, nor would he be in anyway aware that these sub-atoms exist. If it is believed that the program has the sub-atoms, and Searle runs this exact program, he understand neither the atom or its sub-atom. The outputs, while being symbolically identical, the concern towards understanding does not hold, since Searle does not possess any part of the program and would be totally oblivious to its inner workings.

The problem remains with Searle's original version: "*I have inputs and outputs that are indistinguishable* from those of the native Chinese speaker" (Searle, 1980;:p. 418). Searle does *not* have indistinguishable outputs, the *program* has indistinguishable outputs. This is clear from what has just been explained; Searle will not function without the program. While it is certainly not true that this proves the room understands, that requires something more complex, but it does prove the room to have something which produces indistinguishable outputs.

5.6.1 The Internalised Argument

This version is contained within the systems reply, where the program being a product of the computer is now a product of Searle. The failure here is more simplistic as Searle fails to provide a reason for why he would not understand Chinese if he has internalised the program, yet this is not the end of the argument. By having Searle memorise the entire program, he is now indistinguishable from the program to the point that the room could be removed: "If he doesn't understand, then there is no way the system could understand because the system is just a part of him" (Searle, 1980: p. 419).

Searle finds this reply rather embarrassing to supply since it seems to be obvious. Here it

[5]See *The systems reply* in *Searle, 1980.*

is Searle who produces the output from a perfect rulebook - which is clearly grammatically perfect - and this is done by a brain. For Searle to give a convincing argument, one would be required to believe that he has memories an entire new language which he can now converse in fluently and yet not understand a word of the language. No reason is given to this, and fails to support this conclusion and reply.

5.7 Conclusion

From this philosophic approach, the mathematical shall make more sense. The Lucas-Penrose result deals with these same themes, yet both differ in their foundations. There is in the following a notable shift towards a more precise argument, yet it remains constrained to the mechanism argument, and related to the same question of equivalence.

Searle's version of what could be referred to as a 'Turing mind' is further presented in the Lucas-Penrose result with greater accuracy and detail. This, combined with the mathematical variant of the argument and the Gödelian argument is our final point. Here the notion of equivalence is given extended attention with a focus on formal systems and a precise alphabet of the system. What is of significant note, it that this coming section allows for an examination of how *truth*, *knowability*, and *provability* hold up under the Lucas-Penrose result. Much of this shall relate as similar to the paradox of knowabilty. While this shall not be a direct focus, there is much that can be described as problematic under the Lucas-Penrose result related to provability and knowabilty. This invariably leads to the open problems of free-will versus determinism in formal systems.

Chapter 6

Turing Machines and Turing Minds

It is within mathematics that we find the clearest evidence that there must be actually be something in our conscious thought process that eludes computation.

- Roger Penrose

J. R. Lucas and R. Penrose's Gödelian arguments deal with a similar end-point as Searle; the human brain cannot be explained as a machine. Despite this, there are a number of notable differences. Penrose, like Searle, has an original theory of mind which is proposed as both an objection and a counter to the standard mechanist claim. Penrose's *Orchestrated-objective reduction* (Orch-OR)[1] does not so much concern the outputs of the brain, but the internal operations of consciousness (Hammeroff & Penrose, 2013) and the non-determinism of the brain (Lucas, 1961). Both Lucas (1961) and Penrose (Penrose, 1994) work within the same theory that if the operations of the brain can be understood by some algorithm, a machine can be constructed to computer the algorithm, thereby creating an intelligent machine. While Lucas does not include a theory of mind, he does make a number of comments towards open-philosophical problems, most notable of which is free-will and humans as autonomous moral agents (Lucas, 1961: p. 127) Mechanism does remain however, very concerned with a theory of mind, yet it cannot be said that it concerns this over mathematics or logic due to the open nature of the thesis.

The L-P result has failed to provide a compelling argument against the computational theory of mind due to it addressing only systems of first-order logic; those where the in-

[1] See §6.3

completeness theorems directly apply. The result is essential the reverse of mechanism so it beings on an unstable foundation.

Where both (Lucas 1961; Penrose, 1994) conclude the mind cannot be a Turing machine, it is demonstrated that this is not all one may conclude from these theorems. J. L MeGill argues similar (MeGill, 2004) from the L-P result writing "the most one can conclude from the argument is that either we are not Turing machines *or* we are Turing machines implementing a non-classical logic." (MeGilll, 2004: p. 23). Despite the initial strength of the L-P result, it is widely viewed as flawed.

P. Benacerraf formed a response to Lucas in 1967, and before the publication of Gödel's 1951 Gibbs lecture. The Gibbs lecture contained Gödel's comments on a potentially unsolvable Diophantine equation and the brain as a machine, and Benacerraf (Benacerraf, 1967) give a result which mirrors Gödel's own writings, which I shall discuss shortly. Gödel's own difficulty to reconcile the equations with mechanism should allow provide enough of context for the reasons why the Gödelian arguments are in heavy dispute.

Where it has been shown that the use of the Church-Turing thesis[2] fails to provide adequate refutation of the mechanist claim due to the concern of numeric functions over symbolic functions (Church, 1936; Turing, 1939; Jay & Vergara, 2014), extensions of Gödel's theorem's shall be given greater concern. This extends to the use of alphabets over a tape of a Turing machine which, concerning numeric functions, do not produce a symbolic output. If Penrose's quantum alternative is to be believed, there exists a theoretical possibility that a machine may be constructed which each realisable probabilty may be simulated by a computer. A machine of this nature was proposed in principle by D. Deutsch, and rhythmic synchronisation in simulation has been demonstrated by C. Peskin (Peskin, 1975). Peskin's model of oscillators set to fire randomly, of a voltage between a baseline and threshold, have been shown to fall out of chaos and into rhythmic synchronisation within a simulation.

What Peskin's model demonstrates is the underlying complexity concerning chaos theory whereby the natural physical laws impact upon even simulation, which in application to brain

[2]This includes any form of Turing result which is utilised to either prove or disprove any form of mechanism. See *Searle, 1990*.

simulation present the notion that random variable functions will not always follow a chaotic path and fall into sync.

What is raised shall give focus to an objection originally made by Benacerraf (Benacerraf, 1967) in his address of Lucas (Lucas, 1961), and has further been made against the complete L-P result; knowing oneself to be consistent. It shall also be raised that the mechanist thesis of the brain as a Turing machine, and the statement 'I am a Turing machine and I know this' restricts the system to one of first-order logic (MeGill, 2004). What shall not be discussed here are the minor details of what Lucas means by a mind and what is meant in relation to what others believe, or even if a distinction is to be made between minds and brains.

While MeGill's argument (MeGill, 2004) is relatively sound providing a solid response, its use extends beyond the L-P result and the Gödelian theorems with Agudelo and Carnielli (Agudelo & Carnielli, 2008) describing a method to axiomatize computations for deterministic Turing machines where para-consistent logic may be used to solve the Deutsch and Deutsch-Jozsa problems.

This wider problem is highlighted by Lucas where he writes of machines not acting on their own, only in accordance with a program. As is evident from this, the ramifications place strain on the notion of free-will if the mind is a machine. This line of logic has prior been taken by Bringsjord (Bringsjord, 2012) where he presents a proof of the mind as Turing machine and avoids the problem of free-will. These problems have to do with the 'alphabet' of both brains and machines; that which concerns and produces all outputs. Lucas doesn't expand on his concerns in his paper, but the basis of them are broadly similar to those found in Searle's model. This is highlighted in greater depth where I shall discuss Penrose's hypothesis (§6.5.1) before presenting a number of propositions concerning the mind as a Turing machine and its varying implications.

Under Searle, I feel that it is adequate to focus on the more philosophic side of a theory of mind, but under that L-P result the intertwining logic and mathematics don't require it to take precedent, Lucas's concerns towards free-will is deeply philosophical in that it deals with determinism versus free-will and as shall be demonstrated shortly under Penrose's

hypotheses (§6.5.1) and Bringsjord (Bringsjord, 2012) that determinism is favoured under these such systems.

While it can be argued either way of the mind as a Turing machine[3] due to a Turing machine being an idealised abstract machine of an often infinite memory/tape, the brain cannot be compared to the standard model of a Turing machine (Van der Velde, 1992) yet certain highly specific models do allow for the brain as a Turing machine (Weng, 2015).

The direct simulation of a brain by a Turing machine such that $B \leftrightarrow M$ produces a complexity. By having $B \leftrightarrow M$, all functions of the brain must be known in order for their simulation. This gives an uneasy and unfamiliar version of free-will, such as in the case of Lucas (Lucas, 1961) where all possible outputs are determinate.

If it is believed that brains have free-will, then the simulation lacks the free-will of a brain such that the brain and the machine are not equal; knowing states prevents free-will. It is my belief that while this is not strictly true, the hypothesis of a Turing machine simulation of a brain leads to something of a paradox: It makes logical sense that we could define and program all outputs for simulation, but in doing so makes our machine determinate.

As is clear, the use of Turing machines in relation towards AI and the understanding of the brain gives rise to an array of problems which are heavily intertwined. No current research deals with this - perhaps due to the often circular logic of free-will/determinism - but it is given a surface-level mention (Lucas, 1961) (Bringsjord, 2012) which I wish to address since even if the philosophy of free-will is not discussed, much remains which can be said concerning the brain as a Turing machine, or finite-state machine; if the brain is algorithmic. While these problems cannot be ignored[4] the most problematic aspect of their argument concerns their foundations; the belief in consistency. B. S. Anand writes both have placed faith in, and followed "standard expositions of classical theory in overlooking what Gödel has actually proven in Theorem VI" (Anand, 2006: p. 2) and it is further highlighted

[3]However convincing one finds these, the level of research towards a critique of the Lucas/Penrose argument cannot be doubted.

[4]The extent of Penrose's argument makes it difficult to present a concise reply while brushing over much of what he talks of.

LaForte, Hayes, Ford, 1998; MeGill, 2004) how Penrose's argument rests on ambiguities and on Peano arithmetic being consistent.

6.1 The Limits of the Mathematician and Gödel's Conjecture

Do Gödel's theorems tell us anything about human intelligence? From these two highly complex theorems of consistency in mathematical systems, can we learn anything about the brains' creativity or how the brain functions with mathematics? To which we can add; does this differ from how a computer operates?

Having demonstrated the incompleteness of formal number theory, Gödel's theorem can be seen as much about number theory as it can about computation. The implications of the theorem on AI and much of philosophy of mind haven't been given time-over, with the theorems - along with the Church-Turing thesis - often being discussed in logical-philosophic relation over the direct implication of the theorems.

Gödel's theorem brings to light a number of questions concerning the brains mathematical abilities. Penrose has stated (Penrose, 1994) that from the gödel sentence, the brain must operate on a system of reasoning beyond first order logic, which is where he proposed his Orch-Or theory (Hammeroff & Penrose, 2013). While this is Penrose's solution, it is not required to discuss this as the arguments can be discussed and alternatives proposed without a new model of consciousness.

J.R. Lucas' 1961 paper *Minds, Machines, and Gödel* states Gödel's theorem proves mechanism to be false as the mind cannot be explained as a machine. Due to Gödel's theorem holding for all consistent formal systems strong enough to produce basic arithmetic, the mind cannot be explained as a Turing machine since it may recognise the Gödel sentence to be true; the mind cannot be a formal system of first-order logic.

> "Gödel's theorem applies only to consistent systems. All that we can prove *formally* is that *if* the system is complete, then the Gödelian formula is unprovable-in-the-system. To be able to say categorically that the Gödelian formula is unprovable-in-the-system and therefore true, we must not only be dealing with a

> consistent system, but be able to say that it is consistent. And as Gödel showed in his second theorem - a corollary of his first - it is impossible to prove in a consistent system that the system is consistent." (Lucas, 1961; p. 120)

Since Lucas has not included a theory of mind into his result, it is clear that little is lost by not focusing on a non-algorithmic model. In standing without a theory of mind, Lucas' argument remains fairly open as it avoids filtering one towards a specific claim of mechanist. It's very clear from Lucas' base claims that there are a number of routes which were Lucas to follow, would cause his argument to fail. Beside the Gödel theorems, Lucas' most stable affirmation comes from Gödel himself in his 1951 Gibbs lecture (Gödel, 1992: pp. 304 - 323), yet this doesn't fully confirm Lucas' argument. That there exists a set of functions for the recognition of a Gödel sentence functions which Turing machine of first-order logic (FOL-TM) cannot compute, gives a firm basis for critique towards mechanism, but the placement of consciousness as the heart of the problem by both Searle, and Penrose is perplexing.

A difficulty is found in precisely working in their implications partly for the obvious reason of working with abstract machines, and partly due to not working on a specified definition of the philosophy of the mind. This is evident from the repeated critiques of the L-P result and a line of discussion which is found to be highly prevalent in review of Penrose. Of the relation to philosophy of mind, Ernest Nagel and James R. Newman write:

> The human brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines.... Gödel's proof [the first incompleteness theorem]... mean[s] that the resources of the human intellect have not been, and cannot be fully formalized, and that new principles of demonstration forever await invention and discovery.... The theorem does indicate that the structure and power of the human mind are far more complex and subtle than any non-living machine yet envisaged," (Nagel, Newman, 1958: pp. 98 - 103)

What can be inferred from Newman & Nagel's reasoning here differs from that of Penrose who holds the brain not to process mathematics in this algorithmic way. The existence of

this side of the brain was conjectured by Penrose on the basic of Gödel's theorem, which Zizzi (Zizzi, 2012: p.1) suggests to be a form of metathought by the brain; the process of thought by thought. This can easily be understood as a form of consciousness, and indeed the argument can strongly be made that this metathought is the origin of consciousness since it allows for an awareness of the bodies sense beyond a nerve-reaction[5].

This is not to say the brain does not process mathematics in a non-algorithmic manner, only that it extends beyond first-order logic. Additionally it refers only to the contemporary machines, and places no restriction of the possible power of future machines or systems as Lucas, Penrose, and Searle have each done.

Penrose utilised the same anti-mechanist proposition as Lucas, the same which was established by Putnam (Putnam, 1960), yet developed the human mathematician in greater detail speaking of a totality of outputs of the mathematician. The totality of outputs is considered by Lucas since it concerns the extent of the brains *casual powers* and deductive capacity. Both Lucas and Penrose place the totality of provable theorems as being these which are solvable by a human mathematician H. Since H is able to prove more than a formal system, the powers of H far outstrip the powers of any machine. Practically all arguments laid out in Lucas' paper (Lucas, 1961) are a variant of the described, with some being more complex to follow than others, yet he additionally writes the theorems allow for distinguishing between conscious and unconscious beings. Lucas also touches on morality, allowing us to "begin to see how there could be room for morality without its being necessary to abolish or even circumscribe the province of science." (Lucas, 1961: p. 127)

Lucas' paper is problematic. Putnam (Putnam, 1960) argues that mechanism ignores the issue of consistency, writing instead that the human brain is inconsistent since it makes mistakes. Benacerraf (Benacerraf, 1967), similar to Putnam (Putnam, 1960), stated that to know the Gödel sentence is truth, it must first be known id the formal system is consistent. This remains central to Penrose's argument.

In 1989, R. Penrose published *The Emperors New Mind*, which he followed up with

[5]See §6.7 for extended writing

Shadows of the Mind in 1993. Both books deal with mathematics, physics, and consciousness while providing a unified theory of consciousness with physics. Penrose declared that from Gödel's theorem it is clear that the brain and consciousness is non-algorithmic towards mathematical judgements. Since we are able to see the Gödel sentence to be true, it can be proved *outside* of the system. It will not be necessary to discuss Lucas' result in expanded detail since much of the argument in contained in Penrose's version.

There is little which presents itself as 'clear' from the mechanist camp. The broad theories within computer science towards the brain as a machine do *broadly* encompass mechanism, but they in no way wholly reflect it; the mechanist claim is too open. What machine are we referring too? What are its outputs, and how are they represented? What are the limitations of the machine? It is just too obscure to be supported by mathematics and yet the anti-mechanist camp falls victim to the opposite; it is too narrow. The restriction of the machine being of first-order logic subscribes to a belief that the brain is consistent, and appears to ignore the mathematics which the mathematician employs. Lindström speculated that the issue at hand may be that our system in question (human mathematical thinking), is "not a well-defined question and so has no well-defined answer" (Lindström, 2001: p. 243).

Proponents of anti-mechanism encounter much the same issues - as does Searle, which we discussed in the chapter prior - especially with the use of the Church-Turing thesis. It is well known that computers cannot solve all mathematical problems even if given unlimited time or computational power[6], and in knowing the limits of the human mathematician *HM* it is often determined that the brain cannot be a Turing machine or some equivalent. Much has been drawn from Gödel's theorem and despite the extent of the L-P result, Boolos accurately notes these arguments never to have gained significant academic credence (Boolos, 1990). Even with the basic claims of the L-P result, there are problems which arise from

Simply, Lucas and Penrose hold that minds cannot be digital computers and where they differ from Searle's account is in focus on the mathematical side. With their arguments being built upon Gödel's theorem with application to consistent systems of first-order arithmetic,

[6]See §1.1 for further detail of each of these problems

the central claim is that we can recognise truths which cannot be proved in the system, thereby eliminating the possibility of the human mind being a digital computer, or some Turing machine.

Naturally these can be as basic as if 7 is greater that 5, and if 7 is greater than 10. These are verifiable and our system may check and verify - assuming our system is capable of this task or has the set of natural number hierarchy pre-programmed. Both statements are provable and true. In this case there is no reason why this would not be true and provable in the system.

The general problem with mechanism won't be addressed much since treatment varies by author and with the principle remaining the same, the L-P result is directly applicable and there is less of a tenancy to deviate from a single approach as is common with Searle due to his frequent in-direct attacks towards multiple theories, gross simplification and misunderstanding of otherwise well-established theories, and a poor theory of mind.

In addressing Penrose, Shapiro (2003) has claimed that there are certain truths which an agent can know from its system[7] and while he does not go wholly against the L-P result, he does claim that there are further ways with which a knowing agent can be achieved. This is partly the line of logic we shall proceed down.

6.2 On Truth and Provability

Lucas builds on this consistency, and while Searle makes no mention of this, using only Church-Turing thesis as a support[8]. The only difference between Lucas and Penrose is in Penrose's *Orch-OR* theory. Due to its poor academic reception, it shall not be discussed further and it contributes nothing to the L-P result. Where it is necessary to talk of each in the singular, it shall be done so, but towards any comment which can be placed upon both, the L-P result notation will apply.

Redhead (Redhead, 2004) highlights a method by which one may bypass the L-P logic, and MeGill (MeGill, 2003) demonstrates how the use of paraconsistent logic allows one to

[7]See Reinhardt 198; Carslon; 20-, Alexander; 2006 for further

[8]See §3.2

work within Gödel's logic and still avoid the L-P result in order to posit the mind as a formal system. Redhead states we can "always go to stronger systems of axioms and to prove the Gödel sentence, but then we can re-Gödelize, and so on," (Redhead, 2004: p. 732).

Redhead presents two statements (Redhead, 2004) which he writes are the two base claims which are argued.

> (1) For any proposed Gödelian sentence there exists a machine which can deal with it (i.e. prove it)
>
> (2) It is not the case that there is a machine which can deal with all Gödelizations. (Redhead, 2004; 732)

Lucas and Penrose utilise (2) to state the claim that minds are not machines. Redhead claims (1) can be argued and that that is all that is needed to argue its reverse (Redhead, 2004: p. 733).

Before we continue with the L-P result it is important to make certain clarifications of Gödel's paper, despite it being of limited use to both Lucas and Penrose. Of the Gödel sentence, G, the following widely accepted claim is almost as far as Gödel's theorem extends in the anti-mechanist argument. Redhead writes:

> A1. If PA is consistent, the G is not provable.
>
> A2. If PA is consistent, the $\sim G$ is not provable.

Both given sentences are true; A1 formally is a theorem, and A2 is not to the best of our knowledge.

McCall (McCall, 1999) argues that any machine programmed to run and check - i.e., verify - all proofs of PA, will be able to enumerated all proofs and therefore know A1 to be true in the sense that it may *prove* the theorem of A1. This does not extend to A2 which the machine will be incapable of knowing since it will be "incapable of recognizing, grasping, or otherwise perceiving the truth of A2. The concept of "unprovable but true" is beyond it." (McCall, 1999: p. 525)

At its simplest, this is what the L-P result holds as its foundation. We may state:

If *PA* is consistent, then the prove of ϕ is contained in the formula.

If I know that *PA* is not consistent, then I know that ϕ is not provable.

However this would contradict 1. since we can only know ϕ iff we can prove ϕ. McCall gives (McCall, 1999: p. 525):

$$T(\text{If PA is consistent, then } \sim G \text{ is not provable}).$$

In all papers it is clearly stated that if F is consistent, then G is unprovable; true but unprovable. This extends into the negation, $\sim G$. This is not explicitly clear across Lucas (Lucas, 1961).

6.2.1 Between Truth and Provability

Consider any mathematician who utilises the algorithm A for the solution to any problem n. Further, A may be considered as a formal system, and in-line with Penrose's workings with Gödel we shall first consider A to refer to formulas of $\mathbb{N}$ where the Gödel sentence G is contained and true but unprovable.

Redhead clarifies that "truth in mathematics seems to equate with proving theorems but we need to be careful here. There are basically two sorts of mathematics." An example of the first sort is group theory, where one is really concerned with 'unpacking' the logical consequences 'locked up' so to speak in the axioms of group theory. The truths of group theory are analytic in the sense that the conditional statement 'If the axioms are true then the theorems are true' is itself a logical truth. One is not asking the question: are the theorems true *per se*? Rather one is claiming, if the axioms are true under some interpretation, then the theorems are also true under that interpretation" (Redhead, 2004, 732).

Where truth out runs provability - or 'definability', as Redhead also gives - concerns natural number theory and the hope that the axioms are true of "*intended* interpretations, i.e. the natural numbers 0, 1, 2..." and not being being able to prove, or define technically, all arithmetic truths (Redhead, 2004).

6.3 On *Minds, Machines and Gödel*

Beyond arguing mechanist to be false, Lucas (Lucas, 1961) writes the brain cannot explained as a machine and that these theorems allow for distinguishing consciousness. Where Searle later (Searle, 1980) presented a response to the computational theory of mind as a philosopher of mind, Lucas' response is that of a mathematician and philosopher towards the emerging theory of the computational theory of mind, which was proposed by H. Putnam in 1961. However both have acknowledged their respective papers to be a response to the Turing test (Searle, 1980; Lucas, 2002). Lucas addressed his reasoning for using the Gödel theorems over Turing's theorem - which applies with greater easy to Turing machines - as being that "it raises questions of truth which evidently bear on the nature of mind, whereas Turing's theorem does not; it shows not only that the Gödelian well-formed formula is unprovable-in-the-system, but that it is true." (Lucas, 2002: p. 1)

Chalmers' work (Chalmers, 1995) in consciousness has demonstrated the overwhelming difficulty which is present in work on the levels on consciousness, in a similar manner to D. Dennett (Dennett, 1991). While Penrose's Orch-OR theory provides a contrary claim on the standard models, it does not avoid this basic problem of distinction between conscious 'sorts'[9].

In the the case of Lucas, there is no distinction between *sorts*, and his 'distinguishing between' argument is not central since it is his believe that there is no reason for a distinction since only conscious beings may solve the Gödel sentence (Lucas, 1961).

What I shall argue here is that Lucas' argument is neither sufficient or complex enough to refute the entirety of B $\equiv$ M debate, yet it may still be strong enough to call into considerable discussion the mechanist argument. As shall also be the case with Penrose, I wish to highlight the failing of providing a critique against the mechanist argument while remaining of the position that there is unequivocally no equivalence between brains and machines; the stance

[9]By 'sorts' we refer to any different complex sort of consciousness. it is meant as being similar to Dennett's levels only in that is allows for direct referral to separate aspects without referring to the singular phenomena of consciousness

of Lucas and Penrose.

The reasoning for this is similar to the caution taken by S. Shapiro (Shapiro, 2003) in how the mechanist claims towards the mind as a Turing machine are deeply problematic. The mechanist thesis does not explicitly state the brain to be a Turing machine but substantial effort has been made to dispel this given necessity in more philosophic leaning papers. It's difficult to present a concise critique of the mechanist arguments and the respective responses, and instead it becomes a multi-layered critique.

Since the consistency of such systems cannot be prove within the system, such that consistency cannot prove consistency, the brain cannot be a consistence system. We are here in agreement with Lucas but this basic statement lacks any meticulous detailing. This is adequate refutation towards the general theory of the mind as being a Turing machine, but it fails to ground itself as a strictly sound foundation. Despite Penrose's vast adaptation of Lucas' foundation, the core premise remains the same with its concern on mechanism. Both Lindström (Lindström, 2001) and Shapiro (Shapiro, 2003) have taken issue with the ambiguity within the mechanist claims, but if we begin with the simple that the brain can be explained as a Turing machine then Lucas' argument is sufficient.

The holding of the mind to be a Turing machine is problematic due to the limitations of a Turing machine along with an infinite memory, or additional idealisations which can invariable come along with most models[10]. These show clear differences in power between the two. On an infinite tape and from a clear algorithm, theoretically a Turing machine is always correct, yet the human brain is not. There are a large number of differences which one can highlight to demonstrate the strength of one over the other in individual examples, and Lucas' highlighting of arithmetic does provide one strong critique, yet it does not hold up to all counters. Further, it is not widely held that the brain deals only in first-order logic. Alternate theories have been provided that claim then brain to take use of a non-classical model of logic with the implementation of second or higher-order, or paraconsistent logics. It is not necessary for either the computational theory of mind or connectionism to hold the

[10]Certain models are constructed with an infinite memory, but this is not required and they may have a finite memory.

mind as a Turing machine.

6.3.1 The Lucas Model

Due to Penrose's version being a greatly expanded version of Lucas' paper, I feel it would only lead to repetition if a full response was given to Lucas' work. The more complex areas shall be left for §.6.5, and give focus toward the most immediate concerns of Lucas' model concerning cybernetic systems and in/de-ductive systems and and follow from Benacerraf's reply with regard to the consistency of the brain.

The Lucas model and the collective subsequence response concerns two central claims: the first is Gödel's first incompleteness theorem (Gödel I) of there being a sentence G which is true but unprovable in in a consistent system F. The second concerns Gödel's second incompleteness theorem (Gödel II) that the prove of a consistent system F cannot be proven in the system F.

For the brain to recognise the truth of G, it must first know the system is consistent. For it to hold that the Lucas model is correct, the brain must be consistent, yet this is contradictory to Gödel I. If the brain is inconsistent is is able to belief false information and therefore false proofs. Minksy argued along this line of believing false information to demonstrate the inconsistency of the brain (Minsky, 1999), and LaForte, et. al., argued from Benacerraf's original critique against Lucas on the consistency of the brain. The reason for this is simple; the brain cannot prove its consistency as given by Gödel II.

Here Lucas provides a very streamline method for dealing with mechanism, and its a method not employed by Searle despite Chinese room despite it appearing as an effective method. Lucas' method is not to concern oneself with what a system can do, but what it can't. The method by which this is achieved is of immense difficulty; finding a Gödel sentence which is true but the machine cannot prove. By exposing this limitation Lucas' allows for a total bypass of all other concerns of mechanism by demonstrating mechanism to be incompatible as a theory of mind. while this technique may not be wholly sound - as Benacerraf (Benacerraf, 1967) and LaForte et al (LaForte, Hayes, Ford, 1998) argue - it requires certainly requires attention.

Despite referring to the machine as a cybernetic machine, Lucas employs a Turing machine. This is a perfectly legitimate place to start since for any formal system, there exists a theorem-proving Turing machine which proves all and only the theorems of that system. This, in a very simple sense is where Searle goes wrong with the use of the Church-Turing theorem and thinking his Chinese room is a universal Turing machine. Since the Gödel sentence is of first-order arithmetic, there exists no Turing machine which can produce the theorem or prove it. Since this is Lucas' argument of a theorem being true but unprovable in the system, this is the connection between Turing machines and Gödel and why it is logical to being with a Turing machine.

If we take a Turing machine which is dedicated solely to proving mathematical theorems, its total output of provable theorems and compare those to what is provable by a human mathematician, all theorems of the machine with be contained in the collective theorems of the human mathematician whereas the reverse is not true. Therefore the power of the brain is higher than that of the Turing machine since we care concerned only with what can be proven by each.

Lucas begins that Gödel's theorem must apply to cybernetic machines as this is "the essence of being a machine, that it should be a concrete instantiation of a formal system." (Lucas, 1961: p. 113) This presents Lucas' first failing which F. H. George [25] noted that cybernetic machines are not deductive systems, but are concerned with inductive systems; those which are "capable of producing the axioms from which deductive operations start, and these are obviously beyond the range of being formal systems in the sense that makes Gödel's theorem applicable to them." (George, 1962: p. 62)

Lucas is very much unaware of this and states "Human beings are not confined to making deductive inferences" (Lucas, 1961; p. 117) and that a "fair model of the mind would have to allow for the possibility of making non-deductive inferences and them might provide a way of escaping the Gödel result." (Lucas, 1961: p. 117)

Since both *truth* and *provability* rest on PA being consistent, demonstrating the consistency of the brain is not as simple as it is in a formal system. While this may seem like a

sound claim, from the Gödel sentence we cannot say that the human brain can know the truth of whereas the Turing machine cannot since the truth *and* provability of the sentence rest on the assumption that PA is consistent.

McCall writes similar of the Gödel sentence, saying that should PA "turn out to be inconsistent, G would be not true-and-unprovable, but provable- and-false." (McCall, 1999: p. 527)

Seeing two clear alternatives, he writes:

> "If PA is consistent, G is un-provable and true; if PA is inconsistent, G is provable and false. Whichever alternative holds, provability and truth part company. Either some truths are not theorems, or some theorems are not true. The domain of expertise of a Turing machine lies in the area of proof and provability, not in the area of truth. Human beings, on the other hand, are acquainted with both proof and truth, and also know of cases where the two diverge." (McCall, 1999: p. 527)

Shapiro notes that we take F, where G is true and unprovable and simply add G as an axiom to F for F_1 and continue each time adding a new sentence G to produce a new system: "The point is that the process is effective. A machine could carry it out as well as Lucas or Penrose, probably better. Thus, any attempt to effectively delimit the extension of arithmetic truth effectively leads to an arithmetic truth not so delimited" (Shapiro, 2003: p. 274).

6.3.2 Cybernetic systems: *Inductive machines*

We are assuming all formal systems of arithmetic to be implemented by a machine of some definable sorts. Lucas describes cybernetic systems as being machines which "performs a set of operations according to a definite set of rules" (Lucas, 1961: p. 113). Where we state 'cybernetic machine', we will be referring to the machine Lucas describes.

The suggestion here is that a machine should be programmed so that it entertains a set of propositions which have yet to be proved, and from its axioms consistently give a

solution to each. If we assign these to the unproven set of theorems **K**, it is unknown if they can (i) be proven, (ii) are provable by extensions of a pre-existing proof, or (iii) contain a contradiction. If we assign all proven theorems to **P** then we have a knowledge condition similar to what Carlson (Carlson, 1999) established concerning statements ϕ and ψ[11]. With this, all theorems of both **K** and **T** are limited to first-order logic or it is an inconsistent system.

Lucas writes it would be a poor model of the brain if such a system where to accept the first of each pair of undecidable formulae which it has now added to its list of axioms such that it would no longer regard its negation as undecidable

From Gödel's theorem, Schmidhuber constructs a system which follows the basic method of the above described system where the proof search is a "system that can rewrite and improve itself in arbitrary computable ways and in a most efficient fashion." (Schmidhuber, 2006: p. 2) Systems of this nature are therefore less-deterministic since each will develop a different set of axioms of a period of time.

By having a system which is able to prove theorems which are not known from the initial axioms, the use of a deductive system is employed for a solution diverts from Lucas' original claim of the program being a set of instructions for what to do in each eventuality. It seems odd that Lucas writes off a system of reasoning which includes unproven theorems since that appears closer to a brain than a determinate system of set axioms. While such a learning system would be incomplete, it would be a more complete deductive system since it must employ the formal system it holds to prove its axioms and by extension prove statements about its own nature, as with that presented by Carlson (Carlson, 1999) and Alexander (Alexander, 2006).

Lucas' description of a machine prevents it from following parallel to a brain since the behaviour of a machine is " completely determined by the way it is made incoming "stimuli": there is no possibility of its acting on its given a certain form of construction and a certain input of information, then it must act in a certain specific way. (Lucas, 1961: p. 113) A

[11] §.2.6, p.46

cybernetic model of the brain is given as being:

> "Composed of complicated neural and that the information fed in by the senses is "processed" acted upon or stored for future use. If it is such a mechanism, given the way in which it is programmed - the way in which "wired up" - and the information which has been fed into response - the "output" - is determined, and could, granted sufficient time, be calculated. (Lucas, 1961: p. 113)

Lucas' *definite set* of rules and operations as he calls them (Lucas, 1961: pp. 113 - 114) construct a finite machine; the action at every state is determined by the complete code. The limit of the program is its code, in that it can only do what it is programmed to do. Lucas writes that no matter how complete the rules of inference or axioms are, they can never produce the Gödel sentence (Lucas, 1961; p. 115). What Lucas fails to mention here is this applies only if our machine is of first-order logic, and that any machine of definite operations is finite, thereby making it deterministic. This is altered through application of learning systems (Schmidhuber, 2006; Weston, Chopra, Bordes, 2015)

As we see from the cybernetic model and Lucas' system of a definite set, both are deterministic. The acceptance of the implications this would have on free-will is readily acknowledged, but this assumes that brains have free-will, and a very specific kind of free-will. While Lucas acknowledges that it is "not to say that we cannot build a machine to simulate *any* desired piece of mind-like behaviour" (Lucas, 1961: p. 115) we cannot build a machine to simulate every behaviour.

In applying only to consistent systems of arithmetic, we know the human brain cannot be explained as a consistent formal system, or Turing machine. As it is impossible to prove any consistent system to be consistent[12]. The mechanist claim, in Lucas' terms is that there are "distinctive attributes which enable a human being to transcend this last limitation and assert this own consistence while still remaining consistent" (Lucas, 1961: p. 120). We have so far seen a number of different definitions of mechanism

[12]See §2.6 for explanation on consistent and inconsistent systems.

If we imagine an intelligent agent which is a system of PA, F, then it must not be consistent. Since:

$$PA(\neg Con) \tag{6.1}$$

$$F(\neg Con) \equiv F_{PA} \tag{6.2}$$

Lucas' machine is described as being programmed with what to do in each eventuality, and then the correct response, which is given as an 'output' is produced from the original input. It is stated that we may view the mind as a cybernetic system, yet from Lucas' writing it may be possible to have misunderstood a 'cybernetic' machine. Rather 'cybernetic' may be understood simply as a machine which is capable of computation where its 'environment' may be given as mathematics. By extension, if we take a mind to be a cybernetic machine, it is operating in an environment and adheres to the more conventional definition of a cybernetic system[13].

Lucas additionally provides a fairly weak response to Putnam's views of the brain being inconsistent[14] with Lucas writing:

> "The fact that we are all sometimes inconsistent cannot be but from this it does not follow that we are tantamount to inconsistent systems. Our inconsistencies are mistakes rather than set policies." (Lucas, 1961: p. 121)

> "Even though we might preserve the façade of consistency by having a rule that whenever two inconsistent formulae appear we were to reject the one with the longer rule would be repugnant in our logical sense" (Lucas, 1961: pp. 121-122)

In reply to Penrose, LaForte et al write:

[13]Due to this we shall avoid the use of the term 'cybernetic system'.

[14]These were views expressed in private conversation so it is difficult to say how detailed Putnam was since it pre-dates his reply (*Putnam,1962*) to Lucas.

> "Gödel accepted the possibility of a machine which might be equivalent to human mathematical intuition, but which we could never prove to be sound. Turing, however, according to Penrose, actually argued that the theorem indicates that human mathematical intuition is bound to be unsound, since we do in fact believe that we know ourselves to be sound. (LaForte, Hayes, Ford, 1998: p. 8)

6.4 On the Definite Set of Operations

This is where the use of formal systems becomes more complicated. We have the difficulty of deciding what our machine, or agent, must *do*. Must it compute arithmetic? If so it must be incomplete. Must it answer questions on its nature, its truthfulness, and what it can learn from the system? Surely here the machine must also be incomplete. Must it exist in a given environment and make predictions, or complete some other task? Or must it do something more, such as pass a Turing test? There is a difficulty in knowing where to stop with each of these, and at what point the machine can be deemed to have completed its task.

This was the beginning of a much discussed line; it stands as one of the few mathematical theorems to take direct application into the critique of AI and the mind together, and of the still discussed theory that the mind can be understood as a computer. There is no consensus on how true Gödel's theorem is on application to the mind, but on face-value, the argument is certainly convincing.

If we do not view the human brain as a Turing machine, the computational model does not collapse since we may alter the definition of our machine and have it as some other machine. It is important to first know what the machine is, or at least what sort of machine we mean. The brain in the L-P result is a Turing machine, and its definition follows the mathematics of Turing's work more closely that in Searle's argument. The Chinese Room has a design similar to a Turing machine and based on Von Neumann architecture, yet to refer to the Room as a Turing machine encounters a number of problems[15].

The most fundamental question which must be asked concerns Gödel's theorem and

[15]See §2.1.

representation. Penrose has written extensively on this theorem and you'll not find a more convincingly written critique outside of his works, however we must know *what* we are representing and ask if that can be represented by mathematics and theorems[16].

Lucas' system of acting in accordance with each eventuality builds an architecture similar to Searle (Searle, 1980) yet we have less reason to believe that each input has a corresponding output. If we assume the machine to solve only mathematical theorems, they we assume it would follow a definable process, in the same way which we would assume of a human brain. By this we can state that our machine, M, holds a set of finite input states, Q, where each state q_n holds a corresponding end state F. By this method each action is clear and algorithmic such that it can be computed from its start state. This may not seem like a big concern since human actions can be predicted when the brain is monitored; each action has a cognitive 'tell' as observed by Libet et al (Libet, Gleason, Wright,Pearl, 1983) and Libet (Libet, 1985). While Libet's results do no cause determinism, it is of interest to question how much choice an individual has concerning these actions, or if the 'tell' is down to a processing delay.

By acting with this code, there is no free-will in the system, M, and it appears to operate on a strict form of determinism. This determinism is not discussed at length by Lucas, but he writes as a mechanical model of the brain where:

> "[T]he choice between a number of alternatives was settled by, say, the number of radium atoms to have disintegrated in a given container in the past half-minute. It is *prima facie* plausible that our brains should be liable to random effects: a cosmic ray might well be enough to trigger a neural impulse" (Lucas, 1961: p. 114)

6.4.1 Against the Passive Model

It has been argued in more recent years that brains are prediction machines but obviously this is not a narrow definition. The eventualities model is a prediction machine based on passive

[16]See *Shapiro, 2003*

input. A. Clark (Clark, 2013) writes the function of the brain is as a prediction machine where it operates by attempting to match incoming stimuli to top-down predictions.

According to Clarks' model the brain is not passive, as we find in Lucas' cybernetic brain since it is dealing with a multitude of inputs[17]. This passive linear model is similar to what is found in Searle (Searle, 1980).Where passive brains wait for sensory input and then make a decision/prediction based on this input, predictive minds extend beyond this and have no passive state and remain constantly active; "The whole function of the brain is summed up in: error correction." While vague, there is much to suggest this statement to be true (Clark, 2013). The potential for a 'strong' artificial intelligence, as defined by Searle (Searle, 1980), is refuted by Lucas' claim under the Gödel theorems. However as demonstrated Searle's definition is not an accurate claim any AI research concerning finite-state machines since it is directly linked to a theory of mind which one must either hold - or hold some similar - or refute while not refuting all potential claims of 'strong' AI.

6.5 Penrose's Argument

Much of Lucas' argument is contained within R. Penrose's argument, and a familiarity with Lucas' paper serves as beneficial when it comes to reading Penrose. His first work on conciousness, *The Emperors New Mind* in 1989, and his follow up *Shadows of the Mind* in 1994, detail Penrose's working on consciousness, physics, and mathematics. It is uncommon to find a work of such detail and precision in mainstream literature, and it difficult to find a more compelling, and articulated argument, along with such staunch defences as one shall find in Penrose.

These works have surprisingly little to do with Gödel's theorems since they concern two fundamentally different subjects. While Gödel did express a similar argument[18], he acknowledged its limitations. The complexity of Gödel's theorem has lead to a problem in the literature where there are varying claims as to precisely what it shows and how exactly

[17]This is easily imagined as non-linear information inputs entering the brain from all directions at each moment (Clark-model) in contrast to the brain receiving a stream of singular, linear inputs (Lucas-model).

[18]See *Gödel, 1951*

we can see the Gödel sentence and how we solve this. While many remain within PA, the use of paraconsistent logic or inconsistent logic allows for solutions of inconsistency; there are a number of neat mathematical/logical trick around problems.

I suspect both Lucas and Penrose are well aware of these issues but I do not wish to make comment and claim that they have misread or understood Gödel's theorem[19] . What this issue concerns is the unsound notion of human soundness. Penrose, specifically refers to an 'intuition' of human mathematician, which is a form of idealisation.

Penrose's enquiry concerns what is mathematics, and how do we know what is true of mathematics? The level of metamathematics in Penrose is kept under control and this is reiterated by Penrose (Penrose, 1989; 129). As Chalmers wrote; "we are [as] concerned with a system's reasoning about its own beliefs, as we as about mathematics." (Chalmers, 1995: §3.8).

Penrose writes that "a good part of the reason for believing that consciousness is able to influence truth-judgements in a *non*-algorithmic way stems from consideration of Gödel's theorem" (Penrose, 1989: 416).

This truth-reasoning is not found in Searle's result but as we say , self reasoning concerning propositions expressible in *FoL* allows for more intelligent systems which are able to reason in accordance with a mathematical method (Schmidhuber, 2006; Schwering, 2007).

Penrose's conclusion is alluded to on a multitude of occasions yet it is not given expansive detail until later on in the book where he presents his argument as being a *reductio ad absurdum*: "Let us suppose, for the moment, that the ways that human mathematicians form their conscious judgements of mathematical truth *are* indeed algorithmic" (Penrose, 1989: p. 416). In line with Gödel and computation, it is perhaps an error to begin with questions of conscious mathematical proofs and work backwards.

While Penrose is not claiming this to be the way judgements are formed - "let us *suppose*" - is does appear to be an accurate assumption concerning mathematics. With understanding language, we may think of our cognition understanding as being the relation between

[19]Although it is perhaps slightly out of their respective fields of study.

knowledge of a definition and the relation of the word on to the object. With mathematics it naturally seems we form judgements algorithmically. As clear as this may sound, it is not clear cut. On multiple occasions, Penrose refers to the the mathematicians algorithm as singular; “We should need to know what the mathematician’s algorithm really is..." (Penrose, 1989: p. 416).

Idealisation may seem problematic, and it does lead to murky conclusions, yet there is a sound reason for idealising such systems. With machines, we can assume there to be that any error it makes is done to a programming error, which can be corrected, or the program is sound yet it exposes a limitation - precisely what the Gödelian argument claims - and with the mathematician we therefore assume the total output is the totality of all sound, provable, and true mathematical theorem. Naturally, with one system idealised, we must extend the same luxury to the opposing system. The reason for believing the mathematician to be unsound is simple; one mistake is inconsistent.

6.5.1 The Penrose Hypothesis

Over the course of Penrose’s work, he gives four hypothesis[20].

> A. The Human brain has abilities which no Turing machine could be capable of.
>
> B. Human consciousness is non-algorithmic, and therefore the human brain is not capable of being modelled on a Turing machine or digital computer.
>
> C. The understanding of consciousness is to be found in quantum mechanics. Specifically in the microtubles within the neurons support quantum superposition
>
> D. The objective collapse of the quantum wavelength of the microtubles is critical for consciousness. This collapse is the physical behaviour that is non-algorithmic and transcends the limits of computability.

Each of the preceding hypotheses are driven towards an anti-mechanist view of the mind, with hypothesis C, and D being Penrose’s answer to the problem of consciousness (Hammeroff & Penrose 2013), making no strong comment of the problem of mechanism.

[20]Penrose, 1994

In *Shadows of the Mind* Penrose give the following (Penrose, 1994: p. 12):

> $\mathscr{A}$. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computation.
>
> $\mathscr{B}$. Awareness is a feature of the brain's physical action: and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke consciousness.
>
> $\mathscr{C}$. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.
>
> $\mathscr{D}$. Awareness cannot be explained by physical, computational, or any other scientific terms.

Hypothesis $\mathscr{A}$ is what Penrose refers to as being the position of those agreement with 'strong' AI, or functionalism (Penrose, 1994: p. 13). Penrose's application of this hypothesis onto a Turing machine can just as easily be applied onto a digital computer; it is applicable to all algorithmic systems.

Penrose readily acknowledges Searle's room to be against the claims of $\mathscr{A}$ while also defending $\mathscr{B} - \mathscr{D}$. What he writes is that Searle makes no comment on the *outwards*, or *active* aspects of consciousness, but of the *passive* and *inward* aspects (Penrose, 1994: p. 41).

In direct contrast to the L-P result where the brain is a formal system or Turing machine is the direction of critique, there does remain a level of overlap in that both are working on the hypothesis of an algorithm for the brain. Where they divert is that Searle's room remains with its architecture of being built around passing the Turing test verses the brain being a Turing machine or some similar system. This similarity concerns a theory of consciousness, but there is little which deals with self-reasoning.

While Lucas does not posit one, it forms the core of Searle and Penrose's arguments which can be viewed as the reason that both arguments[21] are never advanced towards more direct AI research; connectionism and neural networks.

[21]This is particularly true of Searle despite it attempt at a more formal argument (Searle, 1990).

Theorems $\mathscr{B} - \mathscr{D}$ deal more with the problem of consciousness than $\mathscr{A}$ as such making it difficult to apply the Gödel theorems or Turing's work since it is not a feature of consciousness. How the recognition of mathematics and the process of human knowledge is understood as a conscious phenomena is rarely focused on. It is remarkably simple at first since it is not a natural ability or awareness in the way which sensory data is processed by the brain; there both Lucas and Penrose have an argument to make, yet it does not concern Gödel's theorems.

Hypothesis $\mathscr{B}$ collects three claims into one:

$\mathscr{B}_1$ Awareness as a feature of physical action of a brain.

$\mathscr{B}_2$ Any physical action can be simulated.

$\mathscr{B}_3$ Any simulation of the physical action is insufficient for consciousness.

Hypothesis $\mathscr{B}$ forms a dualism between the physical feature of consciousness since that causes awareness, and the actual physical action governed by consciousness. Penrose provides an additional implication of thinking and awareness as:

> (a) 'Intelligence' *requires* 'understanding'.
>
> (b) 'Understanding' *requires* 'awareness' (Penrose, 1994; pp. 38-39).

Working backwards from this, if it is determined that a machine is able to simulate any physical action then the action, when performed by a brain, has been completed with awareness and intelligence which are required of each other. The physical action is therefore reliant upon awareness since is is the catalyst for any intelligent action.

The lack of notation on the limitations of Turing machines in each hypothesis $\mathscr{B} - \mathscr{D}$ avoids a series of intricate hypotheses which in subsequent years have followed and, while they have not always adhered to 'strong' AI, each have introduced a number of novel approaches to viewing the mind as a Turing machine or finite-state machine of some power. As I have made clear, our focus is minimal and concerned with highlighting the problem of mechanism so I do not wish to focus heavily on Penrose's philosophy of mind yet each follow on from hypothesis $\mathscr{A}$ so that any detailed response to $\mathscr{A}$ which inevitable provide a response to $\mathscr{B} - \mathscr{D}$.

I don't feel it necessary to reply to each hypothesis individually since $\mathscr{B} - \mathscr{D}$ can be discussed under the same claim, yet I do feel that hypothesis $\mathscr{A}$ deserves an extended reply since it is the core of Penrose's, and the mechanist's argument.

Theorem $\mathscr{A}$:

Let M^1 be the human brain and M^2 be some Turing machine running a program P such that the operations of M^1 can be simulated by M^2

$$M^1 \text{ SIM } M^2 \tag{6.3}$$

The program P implemented by M^2 for simulation of M^1 must be of equal power to M^1.

$$M^1 \leftrightarrow M^2 \tag{6.4}$$

The computational process which can be carried out by some Turing machine, M^2, puts the brain, M^1, as Turing equivalent such that generally Q and P are equivalent, where Q is the operational program of the brain and P is the operational program of the machine, iff Q can be simulated by P and Q can simulate P.

$$Q \leftrightarrow P \tag{6.5}$$

Since the system of instructions can simulate the Turing machine it is both Turing complete and hold a set of pre-determined systems by which it may compute its states. Therefore all actions of the brain are completed in accordance with some algorithm.

Let Σ be the finite set of symbols of the alphabet of the brain, M^1 and let $\Sigma^{'}$ be the alphabet of our Turing machine M^2.

Let $\Sigma \leftrightarrow \Sigma^{'}$ such that M^1 may be simulated by M^2.

Formally:

$$M^1(Q;\Sigma) \leftrightarrow M^2(P;\Sigma^{'}) \tag{6.6}$$

.

Response:

Let M^2 be a machine which implements the process of 'hypothesis $\mathscr{A}$' so that operations of the brain can be simulated by M^2 allowing for B $\equiv$ M.

Assuming that we could replicate $\mathscr{A}$ as stated in theorem 1, our machine must be a closed-state system which is pre-determined. Direct and complete simulation, as given as $M^1 \leftrightarrow M^2$ running the program P, may not be viewed as a multi-level strength system. By this it is meant that there may be no small set of operations which M^2 may simulate of the brain to which it can be viewed as being an equivalent. Simply defining P to pass a natural language conversation will not suffice since P is of limited capacity. Any conversational aspect negates the mention of extended emotional response or logical reasoning[22].

This theorem raises a problem of free-will in that if this system is deterministic, of a set of determined symbols of its alphabet, then it seems likely that we would not be able to replicate the conscious actions of the brain.

Take M^2 to be some Turing machine of a pre-determined set of symbols Σ' - where Σ' is a set of finite pre-determined symbols - the simulation of M^1 by M^2 would require our Turing machine and brain to be equal in accordance with Leibniz's law[23] whereby

$$\forall F(Fx \leftrightarrow Fy) \rightarrow x = y \tag{6.7}$$

where F refers to function and x and y refers to the operation of the respective machine. In this case both x and y can refer to *any* function of the machine, its purpose is only to illustrate that each function must be the same as its machine twin.

For this to be consistent such that $M^1 \leftrightarrow M^2$, both M^1 and M^2 must have the same set of finite symbols. Gershenson (Gershenson, 2011) notes of UTM's that being closed makes their outputs deterministic. In theory any machine of brain-simulation must be deterministic

[22]Turing makes no mention of this in his test since the two tasks of fundamentally different. Searle merging of the two causes a fundamental problem to emerge from the start; his argument is built on an assumption.

[23]All mentions of Leibniz's law refers to the *Identity of Indiscernibles.*

such that all possible outputs can be listed.

It may appear that is does not make them predictable if it is running the algorithm from the start state q_0 since the machine is in a fixed-blank state. However in this state if the machine has no input it cannot be known what output will be produced since once the machine returns to state q_0, all possible states may be given. In humans, all states may be given at any time depending on the input since outputs may be reused, although depending on what the input is will determine at least the category of output; such as mathematical, grammatical, or action based.

From q_0 it may be possible to predict what the system may do from there making certain states[24]. From this closed state where Σ' is pre-determined, the simulation of the brain may only be possible if all outputs of the brain are known.

6.6 Idealism: *Performing computational operations*

Penrose writes hypothesis $\mathscr{A}$ and $\mathscr{B}$ to allow for a robot to perform human actions convincingly yet not display conscious awareness. As it ought be clear from theorem $\mathscr{A}$, I wish to avoid mention of awareness and computation, at least for the time, and make only the following brief comment in direct address.

There stands a vast spectrum of complexity, programming, correct and specific application of what the program is doing, and fundamental questions concerning the nature of mental states: What are they? How do we know they exist? Can we clearly define them so that they are exclusive mental states rather than just some other basic form of cognitive action? Penrose is aware of this and readily acknowledges it (Penrose, 1994).

The four viewpoints he presents which one may hold on computational thinking are narrow and only part of the problem. Penrose's mathematical community is the totality of mathematicians stating that what is true for one mathematician must be true for the next and the next, and who - when idealised - can know the totality of all provable arithmetic so

[24]By which I mean a complete set of functions for the completion of an algorithm of some finite length q_{0-n}before the machine stops and returns to state q_0

as to know and recognise the Gödel sentence in the formal system of Peano arithmetic. They are consistent and totally without error, yet their mathematical judgements are not build on a knowable sound computation and rely on a mathematical intuition (Penrose, 1994). If we hold there to be a single algorithm for mathematical computation, our argument is reduced to absurdity, yet to believe in Penrose's intuition we are relying on no sound mathematical process.

The 'community' as Penrose has it, is idealised[25] and stands on the assumption that mathematics is perfect. Where the mathematics not be perfect in that there does not exists an algorithm for each theorem for the result of the correct answer, Gödel I, Penrose's faith almost seems to contradict Gödel's theorem which exposed the limits of mathematics. This objection of the consistency of oneself has been raised in some form by nearly all who have responded in criticism of Penrose. LaForte et. al. writes "For *us* to be able to prove that $\Phi(n_0, n_0)$ . . . and hence to know this, *we* must also prove (or know) that $\forall e \forall x(A(e, x)) \rightarrow \Phi(e, x) \uparrow$" (LaForte, Hayes and Ford, 1998; p. 5).

By having the mathematical community idealised to prove all mathematical theorems allows their abilities to extend beyond first-order logic and PA. Since our machine cannot - according to Penrose and Lucas - know the Gödel sentence to be true since it cannot be proven within the system, our system is therefore limited only to PA. If we assume our system to take use of higher logics then we may assume that it can recognise the Gödel sentence. Indeed this applies to the human brain by assuming it to be paraconsistent[26]. Our mathematicians can never prove the Gödel sentence to be true if they remain within first-order logic since the consistency cannot be proved.

6.7 Propositions 1 - 3

In closing, I wish to propose three independent propositions. In a broad sense, these propositions encapsulate the mechanist argument, along wiht being the collective point of that which I have presented. The intention here is to provide a brief commentary to each, sum-

[25] It is perfect making zero mistakes.

[26] See MeGill, 1996.

marise, and present the core point or issue in a concise manner, with the extended purpose of offering my independent thoughts, and providing the reader with an alternative proposal to the open issues for the purpose of further thought.

Proposition 1

The human brain is not, nor can it be, some form of Turing machine.

Our first proposition is the Gödelian argument in its classic form. In the case of both Lucas and Penrose it recognises not the limits of a Turing machine, but the limits of formal logic. S. Kauffman (Kauffman, 2012) noted the limitations of a Turing machine therefore providing an alternative in quantum mechanics in the same manner as Penrose has done.

Kauffman notes the determinism of a Turing machine to be of particular problem:

> "Given the symbols written on the tape, and rules in the reading head, its behavior at each step is fully determined. This determined behavior is essential to the algorithmic character of the Turing machine. Because it is determinate, the Turing machine is bound by classical physics. However, Turing machines are discrete state and discrete time systems, while classical physics more generally is based on continuous variables and continuous time and is also deterministic, and can, since Poincaré, exhibit deterministic chaos" (Kauffman, 2012: §2.)

Proposition 2

The human brain is not directly a Turing machine but is some form of finite-automata implementing non-classical logic.

By 'some form of finite-automata' it is meant that the brain takes use of a separate logic. As Lucas and Penrose demonstrate the brain is able to solve problems of first-order arithmetic which by their model, means the brain is of a higher power, yet also consistent.

In a paraconsistent system, it could be possible to prove the Gódel system, although paradoxes of this logic are notoriously difficult since they don't rely on standard negations. While this is not fixed, MeGill argues that the possibility of paraconsistency is enough o

refute the L-P result:

> [T]he possibility that we are paraconsistent entails that the *most* that we can conclude from L-P is that *either* (1) we are not TMs or (2) we are able to see the truth of the G. sentence because we are paraconsistent TMs (as a paraconsistent TM would be able to decide the G. sentence). (MeGill, 2004: p. 24)

Proposition 3

The human brain can be directly simulated by a Turing machine where all outputs and internal states are exactly the same.

C. C. Gershenson (Gershenson, 2011) notes UTM's to be both closed and computing once they halt (Gershenson, 2011; p. 2). Of closed machines , the internal states are set and defined making all outputs deterministic. Bringsjord (2012,) by what he describes as a simple proof, demonstrates the human brain to be at least as computationally powerful as a Turing machine in accordance to Leibniz's law.

If the system - or tape - is deterministic, there is no data change over time, meaning that these systems do not alter their axioms (Schmidhuber, 2012). Once the tape is set, the outputs are deterministic. For *proposistion 3* to hold, we must accept that brains are deterministic, and there is no free-will.

It has been shown (Wegner, 1998) that the behaviour of interactive systems goes being algorithms, preventing TM's from being interaction machines because "interaction is not expressible by a finite initial input string. Interaction machines extend the Chomsky hierarchy, are modelled by interaction grammars, and precisely capture fuzzy concepts like open systems and empirical computer science." (Wegner, 1998; p. 316).

Conclusion

Dealing with the ambiguity is a long-term goal of machine learning, and there exists a number of methods by which systems can deal with reasoning in natural language (Xiong, et. al., 2016) (Hermann, et. al., 2015). These deductive systems are able to produce and accurate answer to question where chaining facts are not always required information for the agent. What is proposed here is that recent deep learning methods allow a machine to successfully answer questions from a script by reading, and also to reason when the information is not made explicit. Solutions to Question Answer (QA) systems, especially those containing incomplete knowledge, it is often a requirement for highly expressive - or declarative - language to be used so that incompleteness can be represented and reasoned with. Connectionist and Developmental Networks currently present the most accurate repose and refutation of brains as Turing machines of classical logic, and towards symbol processing by dealing with the deeper levels of networking. By limiting the systems to a specific set of deterministic and closed algorithms, it is near impossible to model the brain on such a system.

Bibliography

[1] Abelson, R. P. (1980) “Searle’s argument is just a set of Chinese symbols". Open-Peer commentary in: Searle, J. R. (1980) “Minds, brains, and programs". *The Behavioral and Brain Sciences* (1980) 3, 417-457

[2] Agudelo, J. C., Carnielli, W. (2008). “Paraconsistent Machines and their Relation to Quantum Computing" At: *arXiv:0802.0150v2*

[3] Alexander, Samuel A. (2013). “A machine that knows its own code". At: *arXiv:1305.6080*

[4] Anand, B. S. (2006). “Why we shouldn’t fault Lucas and Penrose". *arXiv:math/0607333v1*

[5] Antony, Michael V. (2001). “Concepts of Consciousness, Kinds of Consciousness, Meanings of ’Consciousness’." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Vol. 109, No. 1 (May, 2002), pp. 1-16

[6] Aron, J. (2010)“Quantum links let computers understand language." *The New Scientist* 2082790,10?11 (2010)

[7] Aaronson, S. (2011). “Why Philosophers Should Care About Computational Complexity". AT: *arXiv:1108.1791*

[8] Behme, C. (2014). “Is the ontology of biolinguistics coherent?" *Language Sciences* 47 (2015) 32-42.

[9] Benacerraf, P. (1967). “God, the Devil, and Gödel." *The Monist*, Vol. 51, No. 1, The Present Situation in Philosophic Logic (Jan,1967), pp. 9-32

[10] Bringsjord, S. (2012). “The Brain is Obviously at Least a Turing Machine". At: *http://kryten.mm.rpi.edu/SELPAP/BRAINTM/SBringsjord_ BrainAtLeastTM_ 041515NY.pdf*

[11] Bringsjord, S., Xiao, H. (2000). “A Refutation of Penrose’s Gödelian Case Against Artificial Intelligence." *Journal of Experimental & Theoretical Artificial Intelligence*. 12(2000) pp. 307 - 329.

[12] Brooks, Rodney, A. (1990) “Elephants Don’t Play Chess". *Robotics and Autonomous Systems* 6. 3 - 15

[13] Bostrom, N. (2013). “Superintelligence: Paths, Dangers, Strategies." *Oxford University Press*. 2012

[14] Boden, M. A., editor. (1990). “The Philosophy of Artificial Intelligence". *Oxford University Press*, 1990.

[15] Boolos, G., et al. (1990). “An Open Peer Commentary on The Emperor’s New Mind". *Behavioral and Brain Sciences* 13 (4) 655

[16] Carlson, T. J. (1999). “Knowledge, machines, and the consistency of ReinhardtâĂŹs strong mechanistic thesis". *Annals of Pure and Applied Logic* 105 (2000) 51âĂŞ82

[17] Carnap, R. (1950) “Empiricism, Semantics, and Ontology" *Revue Internationale de Philosophie,* Vol. 4, No. 11 (Janvier 1950), pp. 20-40

[18] Chalmers, D, J. (1992). “Subsymbolic Computation and the Chinese Room". In “The Symbolic and Connectionist Paradigms: Closing the Gap" (J Dinsmore, ed.). *Psychology Press*; 1 edition (1992)

[19] Chalmers, D, J. (1995). “Minds, Machines, And Mathematics: A Review of Shadows of the Mind by Roger Penrose". *PSYCHE, 2(9),* June, 1995. url: http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html

[20] Chalmers, D, J. (1995). “Facing Up to the Problem of Consciousness" *Journal of Consciousness Studies*, 2(3):200-19.

[21] Chalmer, D. (1996). “The Conscious Mind: In Search of a Fundamental Theory", *Oxford University Press*; New Ed edition (1997)

[22] Chen, Y., Saffidine, A. Schwering, C. (2018). “The Complexity of Limited Belief Reasoning – The Quantifier-Free Case". At *arXiv:1805.02912*

[23] Chomsky, N. (1956). “Three models for the description of language". *Information Theory, IRE Transactions* on. 2. 113 - 124. 10.1109/TIT.1956.1056813.

[24] Chomsky, N. (1957). “Syntactic Structures" *Martino Fine Books* (2015)

[25] Clark, A. (2013). “Whatever next? Predictive brains, situated agents, and the future of cognitive science". *Behavioral and Brain Sciences. Cambridge University Press*, 36(3), 181âĂŞ204.

[26] Clark, A. (2016). “Surfing Uncertainty: Prediction, Action, and the Embodied Mind". *Oxford University Press* (2016).

[27] Copeland, B, J. (1998). “Turing’s O-Machines, Searle, Penrose and the Brain’. *Analysis*, Vol. 58, No. 2 (Apr., 1998), pp. 128-138

[28] Copeland, B, J. (2011) “Turing and the Physics of the Mind" At: http://www.mathcomp.leeds.ac.uk/turing2012/Images/copeland.pdf (2011)

[29] Copeland, B, J. (ed.), 2017.“The Turing Guide". *Oxford University Press*, (2017)

[30] Descartes, R. (1641) “Meditations on First Philosophy", in “The Philosophical Writings of RenÃľ Descartes", trans. by J. Cottingham, R. Stoothoff and D. Murdoch, *Cambridge University Press*, 1984, vol. 2, 1-62.

[31] Dennett, D, C. (1980) “The Milk of Human Intentionality". Open-Peer commentary in: Searle, J. R. (1980) “Minds, brains, and programs". *The Behavioral and Brain Sciences* (1980) 3, 417-457

[32] Dennett, D, C. (1991)“Consciousness Explained" *Back Bay Books*, 1992.

[33] Edelman, S. (2008) “On the Nature of Minds, or: Truth and Consequences" At: *http://kybele.psych.cornell.edu/âĹijedelman (2008)*

[34] Edelman, G.M., Gally, J.A., Baars, J. (2011) “Biology of consciousness", *Frontiers in Psychology* 2, 2011, 4.

[35] Freidman, A. (2002). “The Fundamental Distinction Between Brains and Turing Machines", *Berkeley Scientific Journal*, Vol. 6, Issue 1, 2002, 28-33

[36] Gödel, K. (1995). “Kurt Gödel, Collected Works: Vol. 3. Unpublished Essays and Lectures" Feferman, S., Dawson, J. W., Goldfarb., Parsons, C., Solovay R. N. (ed.). *Oxford University Press.* (1995)

[37] George, F. H. (1962). “Minds, Machines, and Gödel: Another Reply to Mr. Lucas" in *Philosophy,* Vol. 37, no. 139 (Jan., 1962) 62 - 63

[38] Gershenson, C. (2011). “Are Minds Computable?" At: *arXiv:1110.3002v1 [cs.AI]*

[39] Glymour, C. 2015. “Thinking Things Through: An Introduction to Philosophical Issues and Achievements". *MIT press*

[40] Goddard, C. (2010).“The natural semantic metalanguage approach". *The Oxford Handbook of Linguistic Analysis*: 459âĂŞ484.

[41] Gödel, K. (1931). “On Formally Undecidable Propositions of the Principia Mathematics and Related Systems:. Trans. Meltzer, B. *Dover Publications Inc.; New edition edition*(1 April 1992)

[42] Gödel, K. (1995). “Collected Works of Kurt Gödel: Unpublished Essays and Lectures, Volume 3:. Edited by: Feferman, S. Dawson Jr, J. W., Goldfarb, W. Parson, C., Solovay, R. N. *Oxford University Press USA; New Ed edition* (1 Jan. 1995)

[43] Gödel, K. (1995). “Some basic theorems on the foundations of mathematics and their implication". In: Feferman, S. Dawson Jr, J. W., Goldfarb, W. Parson, C., Solovay, R. N. (ed.). Collected Works (Volume III): Unpublished Essays and Lectures: Unpublished Essays and Lectures Vol 3. New York: *Oxford University Press USA; New Ed edition* (1 Jan. 1995). pp. 304 - 323.

[44] Hameroff, S. Penrose, R. (2013). “Consciousness in the universe A review of the ‘Orch-OR’ theory". *Physics of Life Reviews* Volume 11, Issue 1, March 2014, Pages 39-78.

[45] Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay,W., Suleyman, M., Blunsom, P. 2015. “Teaching Machines to Read and Comprehend" At: *arXiv:1506.03340v3*

[46] Hofstadter, D. R. (1979). “Gödel, Escher, Bach: An Eternal Golden Braid". *Basic Books*; New Ed edition (1999)

[47] Hopcroft, J. E.; Ullman, J. D. (1967). “Nonerasing Stack Automata." *J. Computer and System Sciences.* 1 (2): 166âĂŞ186.

[48] Hopcroft, J. E.; Ullman, J. D., Motwani, R. (2007). “Introduction to Automata Theory, Languages, and Computation". *Reading/MA: Addison-Wesley*

[49] Jahren, N. (1990) “Can Semantics Be Syntactic?" *Synthese*, Vol. 82, No. 3, Epistemology and Cognition, Part II (Mar., 1990), pp. 309-328

[50] Jay, B., Vergara, J. (2014). “Confusion in the Church-Turing Thesis" DRAFT. At: *arXiv:1410.7103v2*

[51] Kauffman, S. (2012). “Answering Descartes: Beyond Turing". Reproduced from B. S. Cooper and Hodges, A. (editors): “The Once and Future

Turing: Computing the World" *Cambridge University Press*. Available at: *https://pdfs.semanticscholar.org/1d97/3000573811cc6403c2ee274b9b594523be7e.pdf*

[52] Kurzweil, Ray (2005), "The Singularity is Near", New York: *Viking Press* (2015)

[53] LaForte, G., Hayes, P. J., Ford, K. M. (1998). "Why Godel's Theorem Cannot Refute Computationalism. *Artificial Intelligence*. Vol. 104, Issues 1âĂŞ2, September 1998, Pages 265-286

[54] Leibniz, G. W. (1684). "Discourse on Metaphysics." In: Loemker, L. E. "Philosophical Papers and Letters: A Selection" (*Synthese Historical Library 2): Volume 2. Dordrecht: D. Reidel/Springer; Second edition.* (31 Dec. 1976),pp.303 - 331

[55] Loemker, L., (1975), (ed. and trans.), "G. W. Leibniz: Philosophical Papers and Letters: A Selection". (*Synthese Historical Library 2): Volume 2. Dordrecht: D. Reidel/Springer; Second edition.* (31 Dec. 1976)

[56] Libet, B.; Gleason, C. A.; Wright, E. W.; Pearl, D. K. (1983). "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity. (Readiness-Potential)". *Brain*. 106 (3): 623âĂŞ42.

[57] Libet, B. (1985). "Unconscious cerebral initiative and the role of conscious will in voluntary action". *Behavioral and Brain Sciences*. 8 (4): pp. 529 - 566.

[58] Lidz, J., Waxmanb, S., Freedman, J. (2003) "What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months" *Cognition* 89 (2003) B65âĂŞB73

[59] Lindström, P. (2001). "Penrose's New Argument". *Journal of Philosophic Logic,* Vol. 30, No. 3 (Jun., 2001), 241-250

[60] Lindström, P. (2006). 'Remarks on Penrose's "New Argument"". *Journal of Philosophic Logic*, Vol. 35, No. 3 (Jun., 2006), pp. 237-237

[61] Lorenz, E. N. (1963) “Deterministic nonperiodic Flow." *Journal of Atmospheric Sciences*. Vol. 20 : 130-141

[62] Lucas, J. R. (1961). “Minds, Machines and Goödel". *Philosophy*, Vol. 36, No. 137 (Apr. - Jul., 1961), 112-127

[63] Lucas, J. R. (2002). “Conceptual Roots of Mathematics." *Routledge*; 1 edition (2011).

[64] Ma, Shuming. Cui, Lei. Wei,Furu. Sun,Xu. (2018) “Unsupervised Machine Commenting with Neural Variational Topic Model". *ArXiv* arXiv:1809.04960v1.

[65] Manson, N. C. (2011). “Why “Consciousness" Means What it Does. *Metaphilosophy*, Vol. 42, No. 1/2 (January 2011), 98-117.

[66] McCarthy, J. (1988). Mathematical Logic in Artificial Intelligence." *Daedalus*, Vol. 117, No. 1, Artificial Intelligence (Winter, 1988), 297-311

[67] McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E. (1955) “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." 1955. Available at: *http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf* August, 1955

[68] Minsky, M. (1967). “Computation: Finite and Infinite Machines." *Prentice-Hall, Inc.*

[69] Minsky, M. (1982). “Why People Think Machines Can’t." *AI Magazine* Volume 3 Number 4 (1982)

[70] Minsky, M. (1988). “Society Of Mind". *Pocket Books*; Pages Bent edition (1988)

[71] Michie, D., & Camacho, R. (1994). “Building symbolic representations of intuitive real-time skill from performance data." In K. Furukawa, D. Michie, & S. Muggleton (Eds.), “Machine Intelligence 13". (pp. 385-418). Oxford: The Clarendon Press, OUP.

[72] Mitchell, M., & Hofstadter, D. R. (1990). “The emergence of understanding in a computer model of concepts and analogy-making." *Physica D*, 42, 322âĂŞ334.

[73] McCall, S. (1999). “Can a Turing Machine Know that the Gödel Sentence is True?" *The Journal of Philosophy*, vol. 10, 525-232

[74] McGinn, C, (2000). “The Mysterious Flame: Conscious Minds In A Material World", *Basic Books*; New Ed edition (200)

[75] MeGill, J. L. (2004). “Are we Paraconsistent? On the Lucas-Penrose argument and the computational theory of mind". *Auslegung* 27 (1): 23-30.

[76] MeGill, J. L. (2012) in “Internet Encyclopaedia of Philosophy". edited by J. Fieser and B. Dowden (2012), At: *www.iep.utm.edu/lp-argue*.

[77] Nasuto S. J., Bishop, M. J. Roesch, E. B., Spencer, M. C. (2015). “Zombie Mouse in a Chinese Room". *Philos. Technol.* (2015) 28: 209.

[78] Newell, A., Simon, H. A. (1976), "Computer Science as Empirical Inquiry: Symbols and Search", *Communications of the ACM,* 19 (3): pp. 113 - 126,

[79] Obermeier, Klaus, K. (1983). “Wittgenstein on Language and Artificial Intelligence: The Chinese-Room Thought Experiment Revisited". *Synthese*, Vol. 56, No. 3

[80] Penrose, R. (1989). “The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics." *Oxford University Press*; Revised edition (2016)

[81] Penrose, R. (1994) “Shadows of the Mind: A Search fro the Missing Science of Consciousness". *Oxford University Press* (1994)

[82] Płonka, B. (2015). “Neurobiology of Consciousness: Current Research and Perspectives." *Studia Humana*, Vol. 4:4 (2015), 23 - 38. doi: 10.1515/sh-2015-0023

[83] Putnam, H. (1960) “Minds and Machines." *Journal of Symbolic Logic*. New York University Press. pp. 57-80 (1960)

[84] Raatikainen, P. (2005)“Truth and Provability: A Comment on Redhead." *The British Journal for the Philosophy of Science*, Vol. 56, No. 3 (Sept., 2005). 611 - 613

[85] Redhead, M. (2004). “Mathematics and the Mind." *British Journal for the Philosophy of Science*, vol. 55, No. 4 (Dec., 2004), 731 -737.

[86] Reinhardt, W. (1980). “Necessity Predicates and Operators." *Journal of Philosophical Logic*, Vol. 9, No. 4 (Nov., 1980), 437-450

[87] Robinson, William S. (1992) “Penrose and Mathematical Ability". *Analysis*, Vol. 52, No. 2 (Apr., 1992), 80-87. Oxford University Press on behalf of The Analysis Committee.

[88] Rogers, H. (1987). “Theory of Recursive Functions and Effective Computability". *MIT Press* (1 January 1987)

[89] Rosser, J. B. (1936). “Extensions of Some Theorems of Gödel and Church." *The Journal of Symbolic Logic* Vol. 1, No. 3 (Sep., 1936), 87 - 91

[90] Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). “Parallel distributed processing." *MIT Press.*Cambridge, MA: 1986

[91] Russell, S. J.; Norvig, P.. (2003). “Artificial Intelligence: A Modern Approach (2nd ed.)", *Pearson*; 3 edition (2016)

[92] Ryle, G. (1949). “The Concept of Mind". *Penguin Classics*; New Ed edition (2000)

[93] Shapiro, S. (2003). “Mechanism, Truth, and Penrose’s New Argument." *Journal of Philosophical Logic*, Vol. 32, No. 1 (Feb., 2003), 19 - 42.

[94] Schmidhuber, J. (2006). “Gödel Machines: Self-Referential Universal Problem Solvers Making Provably Optimal Self-Improvements." At: *arXiv:cs.LO/0309048v5*

[95] Schwering, Christoph. (2017). “A Reasoning System for a First-Order Logic of Limited Belief". In *Proc. IJCAU,* 2017. To appear. At: *arXiv:1705.01817v1*

[96] Searle, J. R. (1980). “Minds, Brains and Programs." *Behavioral and Brain Sciences*, (1980) 3, 417 - 457

[97] Searle, J. R. (1984). “Minds, Brains and Science". *Cambridge, MA: Harvard University Press.* (1986)

[98] Searle, J. R. (1989). “Artificial Intelligence and the Chinese Room: An Exchange." *New York Review of Books*, 36: 2 (February 16, 1989).

[99] Searle, J. R. (1990a). “Is the Brain’s Mind a Computer Program?" *Scientific American*, Vol. 262, No. 1 (JANUARY 1990), pp. 25-31

[100] Searle, J. R. (1990b). “Is the Brain a Digital Computer?" In *Proceedings and Addresses of the American Philosophical Association*, Vol. 64, No.3 (Nov., 1990), pp. 21-37

[101] Searle, J. R. (1992). “The Rediscovery of the Mind." *MIT Press*; First Paperback Edition edition (1992)

[102] Searle, J. R. (2002). “Consciousness and Language". *Cambridge University Press,* 2002

[103] Siegelmann, H. T. (1995). “Computation Beyond the Turing Limit. Computation beyond the Turing limit." *Science*, 268(5210), 545 - 548.

[104] Silver, D. Huang,. A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). “Mastering the game of go with deep neural networks and tree search". Nature, 529(7587): 484 - 489, (2016).

[105] Sipser, M. (1997). “Introduction to the Theory of Computation." *Course Technology*; 3 edition (14 Nov. 2012)

[106] Smolensky, P. (1988). “On the proper treatment of connectionism." *Behavioral and Brain Sciences, 11,* p. 1 - 74.

[107] Somer, J. (2013) “The Man Who Would Teach Machines to Think. *The Atlantic.* Avaible at: *https://www.theatlantic.com/magazine/archive/2013/11/the-man-who-would-teach-machines-to-think/309529/*

[108] Solomonoff, R, J. (1964). “A Formal Theory of Inductive Inference. Part 1." In *Information and Control 7*, p.1 - 22.

[109] Sporns, O., Tononi, G., Kötter, R. (2005).“The human connectome: A structural description of the human brain." *PLoS Comput. Biol.* 1(4): e42.

[110] Sprevak, M. D. (2007). “Chinese Rooms and Program Portability". *The British Journal for the Philosophy of Science,* Vol. 58, No. 4 (Dec., 2007), pp. 755-776.

[111] Squire, Larry R. (2008). “Fundamental neuroscience" (3rd ed.). *Academic Press*, (2012)

[112] Tarski, Alfred. (1931). “The Concept of Truth in Formalized Languages" In Logic, Semantics and Metamathematics" 1983. Corcoran, J. (ed.) *Hackett Publishing Co, Inc.*; 2nd Edition (1983)

[113] Tarski, A. (1983) “Logic, Semantics and Metamathematics" . Corcoran, J. (ed.) *Hackett Publishing Co, Inc.*; 2nd Edition (1983)

[114] Thagard, P. (1986). “The Emergence of Meaning: How to Escape Searle’s Chinese Room." *Behaviorism,* Vol. 14, No. 2 (Fall, 1986), pp. 139-146

[115] Turing, A. M. (1936), “On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society, 2* (published 1937), 42 (1), pp. 230 - 265.

[116] Turing, A. M. (1950). “Computing Machinery and Intelligence". In: Boden, M. A (ed.) The Philosophy of Artificial Intelligence (1990). New York: *Oxford University Press, U.S.A.; New Edition edition.* p.40 - 67

[117] Touretzky, D. S.,. Pomerleau, D. A. (1994) “Reconstructing Physical Symbol Systems". *Cognitive Science* 18(2):345-353, 1994.

[118] Watumull, J. (2012). “A Turing Program for Linguistic Theory." *Biolinguistics* 6.2: 222 - 245, 2012.

[119] Wierzbicka, A. (1972). “Semantic Primitives." *Athenaum Verlag* (1972).

[120] Weng, J. (2015). “Brains as Naturally Emerging Turing Machines." At: *http://www.cse.msu.edu/ weng/research/IJCNN15-807.pdf*

[121] Weston, J., Chopra, S., Bordes, A. (2015) “Memory Networks." *arXiv:1410.3916v11.*

[122] Weston, J., Bordes, A., Chopra, S., Rush A. M., MerriÃńnboer, B. v., Joulin A., Mikolov T. (2015). “Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks". At: *arXiv:1502.05698v10*

[123] Wittgenstein, L. (1953). “Philosophical Investigations". *Wiley-Blackwell*; 4th edition (2009)

[124] Yee, R. (1993) “Turing Machines and Semantic Symbol Processing: Why Real Computers Don’t Mind Chinese Emperors in *Lyceum, 5* (1).

[125] Xiong, C., Merity, S., Socher, R. (2016). “Dynamic Memory Networks for Visual and Textual Question Answering" *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48.

[126] Zizzi, P. (2012). “The non-algorithmic side of the mind". At: *arXiv:1205.1820v1*