From Consciousness to Computation: A spectrum of theories of consciousness and selected salient features germane to the development of thinking machines

by

Vivienne Gwen Wallace

B.Sc., University of Toronto, 2004

Thesis Submitted In Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY

Spring 2013

Approval

Name: **Vivienne Gwen Wallace**

Degree: **Master of Science**

Title of Thesis: ***From Consciousness to Computation: A spectrum of theories of consciousness and selected salient features germane to the development of thinking machines***

Examining Committee: **Chair:** Oliver Schulte
Associate Professor

Robert Hadley
Senior Supervisor
Professor

Richard Vaughan
Supervisor
Associate Professor

Fred Popowich
Internal Examiner
Professor
School of Computing Science

Date Defended: April 5, 2013

Partial Copyright Licence

Abstract

This study investigated the field of consciousness to isolate concepts that might be useful in producing thinking machines, potentially with full consciousness. Questions that informed the research were: Is it possible to identify "successful" theories of consciousness? Can there be a set of salient features that would be useful in the evaluation of theories of consciousness?

A literature survey identifies ways in which enduring problems in discussing intelligence, cognition and consciousness are addressed. The findings trace the development of Western theories of consciousness in the field of Cognitive Science.

The resulting Spectrum of Theories of Consciousness is a high-level organization schema, evaluating theories for their relative reliance on physically verifiable evidence. Four selected Salient Features are: The Turing Test, Qualia, Implementation and Systematicity.

The Spectrum and Salient Features provide a novel structure for comparison and evaluation of theories within the field of consciousness and the attempt to create thinking machines.

Keywords: Theories of Consciousness; Cognitive Science; Turing Test; Qualia; Systematicity; Computational Implementation

Dedication

For my mother, without whom I would never have been able to complete this thesis. But I'm still not going to be a Judge!

Acknowledgements

The act of researching and writing this thesis has been both challenging and satisfying, but the completion of this dissertation would not have been possible without the tireless efforts and support of a number of individuals.

First, to my senior supervisor, Dr. Bob Hadley: my never ending thanks for your support and belief, your insight, and your ability to give me a push exactly when I needed it. Without you, I would never have had either the opportunity or the courage to undertake this effort.

Dr. Richard Vaughan, for challenging me to constant improvement.

The staff and managers at the BC Film Commission, who were always supportive and flexible throughout the many ups and downs of this degree.

My family and friends for their patience with my many distractions and habit of using them as a thinking board.

Finally, my thanks to Dr. Elizabeth Wallace, for your guidance and clarity on even the dark days.

Table of Contents

List of Figures

1 Introduction

1.1 Preamble

Isaac Asimov's Three Laws of Robotics [Asimov, 1942] are familiar territory to many fans of science fiction, and the pervasiveness of this philosophy has spread through the field of speculative fiction and beyond. The spectre of SkyNet from James Cameron's "The Terminator" [Hurd & Cameron, 1984] and the rise of the machines from "The Matrix" series [Silver & The Wachowski Brothers, 1999] provide another, rather more frightening view of the possible future of thinking computers. The concept of artificial intelligence is one that has a long history in fiction, but with the advances in technology that started in the 20th century, thinking machines are moving from the realm of fiction and into the territory of real possibility. Although we may be far from the kind of technological advancement demonstrated by the movie "2001: A Space Odyssey" [Kubrick, 1968], it is fair to say that many of the current accomplishments in technology would seem like magic not too long ago. And yet some would argue the potential for robust artificial intelligence has never been realized. The search for a "thinking machine" has yet to find full success, in part perhaps because the reality of what human intelligence or consciousness actually encompasses and how it came about has yet to be answered to any general satisfaction.

Common wisdom seems to believe that someone building the right kind of program or providing enough data will somehow cause a computer to jump the gap to true intelligence and self-awareness. However, the field of cognitive science takes a rather different view. Cognitive science is a multi-discipline area that mixes philosophy, computer science, psychology, and sciences of brain physiology, neurology and chemistry. It is an interesting and quite modern approach to the problem of consciousness, which is one of the original questions of humanity and may well exacerbate the problem of building intelligent machines.

1.2 Purpose of the Study

This thesis is an attempt to provide a survey of many of the theories of consciousness with an aim towards providing a structure for critical analysis of theories that could potentially be expanded. The thesis developed out of a desire to try and find a "successful" theory of consciousness. However, this question became challenging to answer, since in order to crown a single theory as the most successful, the concept of "success" needs to be defined first.

The term "success" can be very arbitrary in definition and is often dependent upon the predisposition of the person who is defining success. This study is looking to find salient features that can be applied in order to consider the theory successful by its own measures. That is, there will be no requirements or salient properties that would be considered completely antithetical to the theory itself. This is not to say that some salient features may not attempt to push the boundaries of theories, however the purpose of the thesis to try to find particular properties which may provide new insights and enlightening connections between previous unrelated theories.

1.3 Structure of the Thesis

The first chapter will lay out the reasoning for and structure of the thesis.

The second chapter of this thesis will give an overview of a variety of theories on consciousness, starting will some of the relevant Classical and pre-Socratic philosophers and moving forward chronologically to reach the full pinnacle of modern theories. This chapter will also be introducing some of the main questions in the field, as well as common thought experiments, and challenges.

The third chapter will identify the salient features selected, the reason and purpose for their inclusion and how they apply to the theories in question. This chapter will provide a general overview of the spectrum concept, its purpose and development along with other details of the salient features. It will provide a high level explanation of the concepts, theories and salient features used.

The fourth chapter will do a further in-depth discussion of how the established theories of consciousness fit onto the spectrum organization structure and perform an evaluation of these theories based on the salient features established in the third chapter, including assessment of computational possibilities.

The final chapter will present conclusions and discuss limitations of the study, as well as direction for future study.

1.4 Rationale for and Significance of the Survey

This thesis is an attempt to survey, categorize, and explicate theories of consciousness. The two main additions or areas of unique work in this thesis are the spectrum as an organization structure and the emphasis on the importance of systematicity. The spectrum is a high level categorization tool for placing all theories on a single axis of comparison. Previously, comparison of theories was limited to a number of landmark points of divergence that indicated schisms at which new theories were developed. The spectrum represents a single and general concept with a multitude of possible interpretations that allows theories to relate not only to directly opposing theories, but also theories for which there were previously few opportunities for comparison.

The concept of systematicity, in contrast, is not entirely a novel concept. However outside of the classical/connectionist discussion there has been only modest emphasis on the need for systematicity with regard to abstract theories of consciousness.

1.5 Statement of the Problem and Research Questions

Is it possible to define a set of salient features that together define the area of deliberation which should be covered by all theories of consciousness to merit a distinction of some completeness? Successful definitions of consciousness may not be required to place heavy emphasis on all salient features but minimum discussion should exist for most of these properties. This central problem implies related questions

regarding the nature of success in terms of theories of consciousness, e.g. whether or not it is possible to isolate a single theory through comparison of disparate theories, and whether computational implementation is either a component, consequence or unrelated to a measure of success.

1.6 Methodology

The methodology for this thesis began with an in-depth literature survey. The focus of the survey was on attempting to cover the main viewpoints in the field as well as common objections and discussion points.

Throughout this literature survey, attempts were made to discover if certain qualities or lines of reasoning could be found as potential salient features to apply across broad categories of theories, concepts and systems. After a number of these qualities had been uncovered, a systematic approach was applied to ascertain the value for each of the potential salient features.

1.7 Limitations of the Survey

Realistically, scope will always be a problem with this thesis. There are a large number of variations upon the most popular of the theory families, each with their own emphasis and particulars. The work required in order to cover in appropriate detail each of these individual theories would be far beyond the expected scope for this thesis, therefore certain generalizations will have to be made.

1.8 Expected Outcomes

The expected outcome of the thesis is a framework for further discussion of the completeness of theories of consciousness, with a number of salient features identified as necessary benchmarks in the examination of the question of consciousness. These salient features are not meant to be sufficient for a complete or “successful” theory, but instead indicators of comprehensive coverage of a small number of key notions. The

expected outcome, does not however involve the discovery of a single theory which should be held as the main model of consciousness.

1.9 Chapter Summary

This chapter has laid out the framework for the motivation and questions for the area of study of the thesis, the framework for the following chapters and the expected outcome of the analysis to be preformed. The next chapter will begin the body of the thesis with a detailed discussion of the roots and evolution of the study of consciousness and the Mind-Body problem.

2 Literature Survey

2.1 Introduction

The purpose of this chapter is to survey the various theories in the fields of Psychology, Philosophy of Mind, Cognitive Science and consciousness that are applicable to this thesis. These theories cover a wide range of topics but the main focus will be on the concepts and discussions that have lead to the creation of the modern field of consciousness, and which are the influence for this study. In addition, the chapter will cover some of the main points of dialogue that have arisen through the various debates and developments of these ideas. These arguments or points of debate will be issues that still trouble the study of the field of consciousness today.

This survey is an attempt to ground the reader in the major terminology that will be used through the thesis.

2.2 Purpose of Study

One of the underlying goals of this study is to find theories of consciousness that have a good chance of either permitting successful computational implementations of the theory or explaining how and why implementations are not currently possible. From a strictly philosophical viewpoint, implementation is often difficult to define, however, for a computer scientist, implementation is the driving force for most work. Thus topics in the area of cognitive science and consciousness are often balanced between the two impulses. This thesis, although discussing theories strictly in the abstract sense, is meant to be a first step towards identifying new points of interest that could potentially lead to successful simulation or implementation of even limited consciousness-type processes.

This survey chapter, then, is meant to supply grounding in the issues and theories that discuss consciousness. Many modern attempts at Artificial Intelligence (AI) do not concern themselves with the question of consciousness, however as computational power and sophistication have increased it seems to have become clear that a more powerful computer is not the key to cracking the AI problem. But what is the solution? The long history of consciousness studies in the field of philosophy may provide important insights.

Much of the work of this thesis is to look at the corpus of works around consciousness in a new way and try to see if there are hidden relationships that have not been previously explored. When exploring the utility and effectiveness of the criteria of evaluation that will be introduced, discussion may refer to the various objections of established theories, and this survey will be laying out these theories in advance. Later chapters in this thesis will presuppose knowledge of these theories.

2.3 Structure of Survey

The survey has been conducted using a chronological timeline for the introduction and development of the various theories of consciousness. To begin with, some background on the classical positions of antiquity will be introduced both as a historical note and also to set the stage for the next leap in the work which occurred around the 17th century in Europe. Generally, most of the theories presented will be focused on the European traditions of mind, soul and consciousness as a means of defining the scope of discussion.

As to the choice of a chronological presentation, since many of these theories developed in direct response to previous work, either to counter perceived weakness or to build upon an established theory, it seems a logical and organized way of presenting the information. In further chapters, other orderings will be presented and theories will be arranged in a different manner to demonstrate interesting connections and relationships. However, as this chapter is meant as an overview, it seems prudent to present the history and main theories in a relatively straightforward manner so as to give the reader an idea of the area and scope of the survey that was performed.

2.4 Classical Thinkers

2.4.1 Introduction

Modern writings in cognitive sciences and consciousness studies often include historical remarks as to the origins of various theories. Names such as Plato and Aristotle are commonly invoked as the inspiration for theories, questions and arguments. Therefore when beginning a study of the various theories it seems only appropriate to begin with a section on the classical origins.

2.4.2 Classical Section Organization

This section will focus first on the classical origins of the Mind-Body question, which is often seen as a central tenet in philosophy of mind and consciousness studies. After introducing the main points of the Mind-Body problem, we will then turn to two of the most famous and long standing positions in answering this question: the early roots of the theory of Dualism and the early roots of the Monist position. Both of these views have been developed extensively by later philosophers, to the point that these classical views from ancient Greek philosophers seem to be but shadows of the later theories. However, it is important to note the various influences that caused theories to develop into the commonly held modern views.

2.4.3 The Mind-Body Problem

The Mind-Body problem or question, while identified and named by Rene Descartes in the modern world, was also a subject of debate in classical philosophy. This question is the problem of how the mind and body or soul relate to each other. What is the mind or soul? How does it attach to the body? This question is often answered by religion in the form of an immortal soul granted by some deity, and indeed many classical and modern philosophers wrestle with the distinction between the philosophy of mind and the religion of soul. However, it can be argued that in the classical period the idea of “soul” was perhaps used in a broader sense than is commonly considered in modern times. Indeed it sometimes seems as if the concepts of mind and soul were almost interchangeable. Despite issues of terminology,

semantics and conflation of mind and soul, there have been several notable positions on the Mind-Body problem that can be said to have provided strong influences on the development of important theories.

2.4.4 Classical Dualism

Dualism is the concept that the universe consists of two separate and distinct substances, and the central question of the character of these substances is often the defining point of any particular dualist theory. For Plato, Forms and physical matter are the two separate things. However, although Plato discussed his idea of the Forms throughout his work, it seems that there was no true discussion of the details of his theories in any of his works [Silverman, 2012]. The idea of forms is that although there may be many round things, there is only one Form for roundness. So the sun, a bowl, a wheel, etc. are all round items, but they all seem to share this characteristic or property of Round. It was these properties that Plato held to be a Form [Silverman, 2012].

Plato goes on further in the Phaedo to talk about the soul and "draws a contrast between unchanging Forms and changing material particulars" [Silverman, 2012]. Most common interpretations of this believe that the soul is some kind of Form when applied to humans, with each person having their own separate soul. Just as the form of Roundness is immaterial, universal and immortal, so too is the human soul.

> "The argument of the *Phaedo* begins from Plato's assertion that the soul seeks freedom from the body so that it may best grasp truth, because the body hinders and distracts it: the soul comes to be separate (*choris*) from the body, itself by itself ((*aute kath auten*) (64c5–8)). The senses furnish no truth; those senses about the body are neither accurate nor clear. The soul reckons best when it is itself by itself, i.e., not in contact with body (65a-65d3)." [Silverman, 2012]

It is this clear separation of body and soul that is a hallmark of what will later develop into Dualism: the theory that the mental and the physical are two different and separate substances. Although Plato's focus seems to be on the soul rather than the mind or mental, this separation of the body from the part of the human that performs thinking and feeling is why Plato's work on forms can be seen as a precursor to Dualism.

2.4.5 Classical Monism

Parmenides in the 5th century B.C. was one of the first philosophers to hold views that could be considered to be a form of Monism. That is, Parmenides believed that there existed only one thing in the universe [Silverman, 2012]. Monism being the theory that there is only one kind of substance in the world, or in terms of philosophy of mind, that both mental and physical are of the same substance.

Aristotle talked about the combination of form and matter [Beakley & Ludlow, 2006], which at first blush may seem similar to Plato`s use of Forms. "…Let us inquire about the parts of which *substance* consists. If then matter is one thing, form another, the compound of these a third, and both the matter and the form and the compound are substance" [Aristotle, 2006]. However, it does seem that Aristotle wants to say that the *form* of a person is the soul and the *matter* is the body. While this may seem to separate out the two in some kind of dualism, Aristotle seems to want to say that the two actually cannot exist without each other, at least in some kind of thinking person. "That is why we can wholly dismiss as unnecessary the question whether the soul and the body are one" [Aristotle, 2006]. And he believes that he has answered the question of what a soul is as "substance in the sense which corresponds to the definitive formula of a thing's essence" [Aristotle, 2006]. "From this it indubitably follows that the soul is inseparable from its body, or at any rate that certain parts of it are (if it has parts) – for the actuality of some of them is nothing but the actualities of their bodily parts" [Aristotle, 2006].

So Aristotle believes that the soul and the body are irrevocably bound, which contrasts strongly with the Dualism view of Plato that the soul is immortal and can separate from the body [Silverman, 2012]. While Aristotle's views are more sophisticated and detailed than a simplistic monism view, it is clear that he believes the physical and the mental to be parts of the whole. This can perhaps been seen as an early kind of a specific form of Monism commonly referred to as Functionalism [Beakley & Ludlow, 2006]. Indeed, some modern sources believe that Aristotle is making what can be called a "multiple instantiation argument", so called because a form can be instantiated or created by many different types of matter [Beakley & Ludlow, 2006], a point that will be revisited in future discussions of Functionalism.

2.4.6 Summary

The world of classical philosophy is a rich one, and there are many other classical philosophers who have contributed to the corpus of works on metaphysics and ideas on substance, mind, body and soul. However, this section was meant only as an introduction to cover some of the basic and most influential ideas that have a direct impact on the study of consciousness.

2.5 17th-19th Century Enlightenment

2.5.1 Introduction

The 17th century was an important time for the development of Philosophy of Mind. In 1641, Rene Descartes published the "*Meditations on First Philosophy"*, which would prove to be a landmark publication. In his Meditations, Descartes would present the basis for a theory that would later come to be called Cartesian Dualism. That publication and the subsequent works both by those who agree with Descartes as well as opponents to his theories would provide material for debate for the next several hundred years. Indeed, although Cartesian Dualism has fallen out of favour with modern theorists, even today the theories and debates first introduced during this period continue to be relevant and widely discussed.

2.5.2 Enlightenment Section Organization

This section begins by reviewing the work of Rene Descartes and the introduction of his modern incarnation of dualism, now widely called "Cartesian Dualism". The work of Descartes in his Meditations was a landmark point for the theory of dualism and arguably the entire field of philosophy of mind. Many of the seminal papers since Descartes have in fact been in response to the theory of Cartesian Dualism. Whether supporting or dissenting, it is difficult to underestimate the impact of Descartes. The fact that 20th century philosophers like Daniel Dennett still feel the need to address the idea of the "Cartesian Theatre" [Dennett, 1991], demonstrates the incredible influence it has on so many debates and underlying assumptions of current theories.

Late in this section, we will cover some of the prominent publications from this period that offer an opposing viewpoint to Descartes and Cartesian Dualism. These papers are the works of a large group of philosophers who may be called “The Empiricists”. Although The Empiricists often varied widely on the theories they supported, let alone specific details of theories, the common thread was the desire to base philosophy on The Empiricism Thesis: “We have no source of knowledge in S or for the concepts we use in S other than sense experience" [Markie, 2012]. That is, all knowledge in the universe that can be acquired is only knowledge that can be gained from the senses.

Some Empiricists, such as John Locke, supported a version of Dualism [Beakley & Ludlow, 2006], while others sought truth instead in various forms of Monism. It is useful to approach the papers and theories that opposed Descartes by presenting updated version of classical Monism. This leads to a presentation of some of the theories and arguments of empiricists who raised questions around the field of mind and consciousness, often inspired by Descartes’ Meditations, which remain noteworthy in the field of Philosophy of Mind.

2.5.3 Descartes and Dualism

If there is one author irrevocably associated with dualism, it is Rene Descartes. His specific brand of dualism, Cartesian Dualism, is a form of substance dualism. This view holds that the universe is made up of two entirely different and separate substances: the mental and the physical. Our physical bodies are created of an entirely physical substance and our mental selves are therefore entirely non-physical. No part of the mind actually arises or is created from our physical bodies, but is instead a separate object altogether that is somehow attached to the physical. This conclusion is aptly demonstrated by the thought experiment Descartes poses in his Meditations to try and imagine away all substance in the universe, and ends up with only the mental self, entirely divorced from all physical parts. Often referred to as the “brain in a vat” experiment, Descartes ultimately determines that “I rightly conclude that my essence consists solely in the fact that I am a thinking thing” [Descartes, 1911].

Another aspect that was highly controversial in Cartesian Dualism is the issue of mental causality. How does the purely mental substance of the mind and its thoughts affect the purely physical substance of the body? In Descartes' view, the mind was causally able to affect the physical, a view that is called "interactionism"[Robinson, 2011]. Descartes identified the central point of connection between the mind and the body as the pineal gland [Robinson, 2011]. However, although Descartes identifies a site for this connection, there is not much detail on how the actual connection occurs. Since the mind and body were entirely different substances, how is it possible that they interact and even affect each other? This shortcoming would inspire a number of questions and responses in the community.

2.5.4 The Empiricists

The Empiricists were philosophers deeply dedicated to the Empiricism Thesis and the idea that knowledge must be gained from sensory experience. The gaining of knowledge through the act of pure reason, such as Descartes` thought experiments to arrive at Dualism, was often deeply troubling to empirical thinkers. So, even if a particular philosopher had a certain amount of agreement with the content of the theory of Cartesian Dualism, Descartes methods were often called into question.

2.5.5 The Further Development of Monism

Monism is the idea that there is only one type of substance in the universe, building upon the ideas of philosophers such as Parmenides. Early Greek philosophers had many debates about what the actual substance of the universe was, but starting with this middle period, it seems that monist theories fall roughly into three different categories: those who believe that all substance is mental, those who believe it is physical and those who hold a more neutral position where substance is neither mental nor physical.

As a note, some of these philosophers at the time may not have placed their theories under the Monism banner, and in fact some of these authors have only had the Monist label applied in modern times [Stubenberg, 2010]. However, despite any labels, or lack thereof, the kinds of arguments, problems and theories raised during this time

period have been an influence on the current landscape of the Monist side of the Mind-Body issue.

2.5.6 Idealism: Substance as Mental

In 1710, George Berkeley published a work laying out the arguments for a theory that would later come to be called “Idealism“, which holds that the only substance in the universe is actually mental in character [Beakley & Ludlow, 2006]. His argument is based on the concept that for any physical object, the only thing that we can really know is what we perceive through our senses. So all our knowledge is sensory, and this leads Berkeley to the conclusion that any object perceived in this way, is actually just some group of sensory stuff. Furthermore, since sensations live only in our minds, then this group of sensory stuff also lives only in our mind and thus the physical object is actually only mental and exists only in our minds [Beakley & Ludlow, 2006]. Berkeley himself states: “that consequently so long as they are not actually perceived by me, or do not exist in my mind or that of any created spirit, they must either have no existence at all, or else subsist in the mind of some Eternal Spirit” [Berkeley, 1843]. There is also mention in his writings of the objection that perhaps instead of just an objection made of perception, there is in fact a real object with some kind of extension or physical matter. However, Berkeley eventually circles back to the idea that our ideas of this extension or matter exists only in our minds [Berkeley, 1843].

Another similar but distinct theory to Idealism is one that was introduced by Gottfried Wilhelm Leibniz in 1714, in his work “The Monadology“. Leibniz also believed that all substance must be of a mental character, however his work rested a concept he introduced called a “monad“. Leibniz held that monads are the component parts of all physical objects, however for monads their “whole only features are *perceptions* of the world” [Beakley & Ludlow, 2006]. Humans are aware of their perception, which is “apperception” [Beakley & Ludlow, 2006], but only at a macro-level and not at the level of individual monads [Beakley & Ludlow, 2006].

2.5.7 Materialism: Substance as Physical Matter

Materialism, the Monist concept of substance as physical, or matter, has a long history in metaphysics, as demonstrated by the earlier section on Parmenides. However, with the rise of scientific method during the Enlightenment and the 17th century emphasis on empirical methodology and knowledge, Materialism was on the rise.

Julien Offray de La Mettrie, a French physician, was a strong proponent of empirical data and materialism. De La Mettrie, actually used the term "materialism", in his treatise "Man a Machine", first published in 1748. In this work, de La Mettrie spends time discussing the connection between the mind and physical body. For example, the role that the mind plays in situations where a limb has been removed from the body, the particular mental character that occurs when the body is sleeping, and the effects that drugs such as opium have on the mental faculties as well as the physical body. In all of these examples there is a close tie between what is happening physically and what is happening mentally [De La Mettrie, 1912]. De La Mettrie wants to make a definitive statement in support of materialism: "Let us conclude boldly that man is a machine, and that in the whole universe there is but a single substance differently modified" [De La Mettrie, 1912]. De La Mettrie seems to believe that the soul, a subject of previous discussion even for early monist views such as Aristotle, is no longer a subject worth exploring. "Given the least principle of motion, animated bodies will have all that is necessary for moving, feeling, thinking, repenting, or in a word for conducting themselves in the physical realm" [De La Mettrie, 1912].

2.5.8 Neutral Monism

Neutral monism is a branch of monism, which means that mental and physical are considered to be one. However, in neutral monism, as opposed to other branches of monism, the ultimate stuff on the universe is neither physical nor mental. So perhaps it's better to say that there is only one type of stuff in the universe, but it is neutral, that is, a different kind of substance outside of the previous mental/physical debate.

Baruch Spinoza was a philosopher in the mid-17th century who did not declare himself as a neutral monist, but has been classified with this group by later philosophers. However, this label is debatable and it is perhaps more precise to name Spinoza as a

"Dual Aspect" theorist, which has close ties to neutral monism, but is subtly different [Stubenberg, 2010]. The main idea is that certain substances can have two "aspects": the mental and the physical. So although the substance is neither mental nor physical it can present as both or each under different circumstances [Stubenberg, 2010].

The Dual Aspect theory may perhaps been seen to be related to certain forms of dualism, for example Chalmers' view of the property dualism of Information [Chalmers, 1995]. However, the distinction would seem to be that dualism holds that there are two separate and distinct substances, while the Dual Aspect theory is that there is really only one substance, that this substance has dual characters, and can thus look different depending on the time/method of observation.

Hume is another philosopher that has been controversially claimed as a neutral monist in the years since his death [Stubenberg, 2010].

> "First, there is the idea of neutral entities: entities that are not intrinsically or essentially percepts or objects but can be counted as either, given the relevant context. Second, the idea that mind and body are reducible to/constructible from these neutral entities. Though this may not be the only plausible reading of these passages (see, e.g., Bricke 1980, Flagge 1982, 1991, Backhaus 1991) the case for counting Hume as an early neutral monist has considerable merit. While it may be controversial whether Hume really was a neutral monist, his enormous influence on the development of subsequent versions of neutral monism is beyond serious doubt" [Stubenberg, 2010].

A non-controversial point on Hume's views is that he was a strong Empiricist. Hume criticized the entire concept of substance in general due to the lack he felt of empirical content. Instead Hume proposed that the mind was simply a sort of "bundle", but lacked a convincing explanation on what exactly bound this bundle of mind together [Stubenberg, 2010].

2.5.9 Empirical Methodology Arguments Against Dualism

John Locke, an English philosopher and important writer in this Enlightenment period, while showing a preference for dualism over materialism, had serious reservations for the manner in which Cartesian Dualism was supported. "According to

Locke, our knowledge of the world is limited both by our *imperfect evidence* and by the *limited ideas* we use to understand this evidence" [Beakley & Ludlow, 2006]. Thus, a simple appeal to the theoretical concepts of mind as presented by Descartes in his *Meditations*, was lacking in empirical content and therefore not sufficient for Locke to decide between Monism and Dualism.

2.5.10 John Stuart Mill

John Stuart Mill was a mid-19th century philosopher who introduced a number of concepts that would have interesting implications in later work starting from the mid-20th century. These concepts are explored in the following two subsections.

2.5.10.1 The Other Minds Problem

The Other Minds Problem is a classical issue in the field of Philosophy of Mind, and one that can be applied to Cartesian Dualism. The exact origins of the problem are not entirely clear, but Mill is considered to have given a clear description when discussing the issues of causality between mental states and actual behaviour [Hyslop, 2010]. The basic concept of the other minds problem is that although we assume other humans have the same kinds of inner lives and mental objects, it is difficult to actually objectively prove this fact. The assumption that other humans are the same as us is hardly a drastic leap in logic, but nevertheless it is one that has little empirical proof.

Furthermore, this problem can be broken down into two separate but related questions. The first is a search for a way to prove that other minds do have the inner lives we assign them. The second question is regarding our ability to try and conceptualize mental notions that are not our own [Hyslop, 2010]. After all, our only knowledge of mental states is through our own experience. Therefore the ability to be able to understand the mental states of a mind that is not our own can only truly come through our own experience. This conceptualization problem seems to seek some kind of more objective solution, but an objective way to understand personal experience seems a challenging endeavour.

2.5.10.2 Emergence

Mill was one of the early proponents of the concept of Emergence. For Mill, when studying physical and chemical matters, it became clear that in some situations, specifically chemical reactions, the final result would in some way be an aggregate of the original component parts [O'Connor & Wong, 2010]. Mechanical properties, in Mill's view, could be analysed strictly by the components, so the results found by looking at the reaction as a whole would be identical to results found by studying and aggregating the components. The emergent properties of chemical reactions were in strict contrast to the mechanical view [O'Connor & Wong, 2010]. Simple study of the component parts was not sufficient to explain the results that would emerge from a chemical reaction as a whole. The sum of the whole was greater than that of its individual parts. This idea of emergence would evolve over the centuries, but the implications for consciousness theories would be the position that consciousness itself could be one of these emergent properties that appeared from actions of the mind or brain.

2.5.11 Summary

The Enlightenment period starting in the 17th century, was a period of high interest in metaphysics and ideas of self. It would be a time which saw the development of many foundational concepts in philosophy of mind that would lay the groundwork for later theories of consciousness. Beginning in the 20th century, much of the traditional Cartesian Dualism would become sidelined for newer ideas. The explosion of various forms of Monism during the Enlightenment would continue, often rooted in huge leaps in scientific knowledge. However, the appeal and influence of Cartesian Dualism would remain strong.

2.6 20th Century and Beyond

2.6.1 Introduction

The 20th century was a time of huge growth in terms of theories of philosophy of mind. Psychological theories were advancing as well, and often the science of psychology was starting to be incorporated or at least to affect philosophical theories.

The advances in sciences, understanding of the brain, and the burgeoning of computer science all lead to the advancement of the fields of Consciousness and Cognitive Studies.

2.6.2 Modern Section Organization

The section on 20th century positions will be structured in a loosely chronological order, beginning with theories that reached their prominence at the beginning of the time period and continuing through the various developments, often deeply affected by scientific breakthroughs. Although there is often overlap in chronology between disparate theories, this section will be grouping like-minded theories together in an effort to provide a cohesive overview of the often fragmented and shifting development of Philosophy of Mind.

To begin with, there will be an overview of Behaviourism, an area that was originally inspired by psychology. Despite the fact that traditional Behaviourism is no longer favoured in Cognitive Science studies, many of the basic principles of Behaviourism still retain a certain amount of appeal to various Monist groups.

Following the section on Behaviourism will be a summary of modern Materialism and two of the most prominent modern forms of Materialism: Eliminative and Reductive Materialism. Subsequently, Functionalism will be covered, a theory group that can be identified closely with Materialism although there are subtle difference. This section will include discussion of Physical Symbol Systems, the Computational Theory of Mind, Connectionism and several substantial opponents to Computationalism, including the seminal Chinese Room Argument by John Searle.

Towards the end of the section on 20th century theories and theorists, will be collected a loose grouping of theories under the title of “Non-Reductive Physicalism and Property Dualism”. This sub-section will cover the various theories that have arisen from attempts to ground theories in physicalism without the rejection of mental attributes that characterizes several forms of Materialism. Theories that rely on concepts such as emergence, supervenience and use mental properties will all be grouped under this heading. Finally, the section will finish with a brief discussion of some of the Quantum

theories of consciousness that have been proposed and refined based on recent knowledge of Quantum Physics.

2.6.3 Behaviourism

Behaviourism had its start in theories of psychology developed in the early 20th century. Philosophical Behaviourism is a theory of consciousness that actually seems to disregard the concept of consciousness in many ways. For behaviourists, the most important thing about consciousness is observing the behaviour of conscious beings. Searle believes that behaviourism entails that consciousness is *only* publicly observed behaviour [Searle, 2005].

Philosophical Behaviourism was first mentioned by Gilbert Ryle. It is widely considered that this concept of philosophical behaviourism as introduced by Ryle was one of the major final arguments against Cartesian dualism that caused such a huge decline in the general acceptance of dualism as a viable theory of consciousness [Tanney, 2009].

"A person therefore lives through two collateral histories, one consisting of what happens in and to his body, the other consisting of what happens in and to his mind. The first is public, the second private. The events in the first history are events in the physical world, those in the second are events in the mental world." [Ryle, 1949]

Ryle thinks that the problem with the traditional view of the mind, Cartesian Dualism at the time, is a problem of "Category-mistake" [Ryle, 1949]. That is, although the thinker may be dealing with different abstract concepts, the concepts may still be placed in the wrong kind of category [Ryle, 1949]. For his examples he uses one of a university, the idea being that there are structures in a university, such as different colleges, the registrar, etc., which all have physical centres and areas and thus are of the same category. But the university itself is of a different type of category and does not have the same kind of features as these other departments. All of these entities are abstract concepts in theory, but belong in different categories. For Ryle, the idea of mental as separate to physical, and the expectations therefore of the kinds of rules and requirements that would characterize a mental mind made of non-physical substance,

was a misunderstanding of what the mental was and therefore a case of false expectations. Ryle wants to have his theory be neither materialism nor idealism [Ryle, 1949].

2.6.4 Materialism

2.6.4.1 Materialism Overview

The materialism theory is a Monist theory, founded on the basic tenet that there is only one kind of substance in the universe, that is the physical, often called matter. The materialists reject any kind of dualism and therefore the idea that the mind and consciousness are some kind of special, non-physical things. Materialism is often referred to as physicalism as well. Although Materialism has been widely in use since before the Enlightenment, the term physicalism originated from the area of linguistics in the 1930s. However the two terms are mainly used interchangeably these days [Stoljar, 2009].

The idea of Materialism is a very broad umbrella term that encompasses several quite different and distinct versions of the basic Materialistic model. Reductive Materialism, Eliminative Materialism and Functionalism are three of the biggest sub-categories that fall under Materialism, and provide many of the most popular theories in the field of consciousness. All of these versions of Materialism tend to share a focus on scientific method and experimentation to discover the underlying causes of consciousness. The issue of falsifiability, for example, can be an important part of scientific method and experimentation, and thus also of Materialist discussion. The concept of falsifiability requires a thesis that concerns empirical verifiability, that is, a thesis that can be tested. For a certain hypothesis, an attempt to prove the truth of the statement would require an infinite number of tests; however a much smaller subset of tests, or even a single test, could prove that the thesis is false, and thus falsifiable. This strong emphasis on scientific method is one of the characteristics of the various forms of Materialism. The following sections will provide a brief review of each of these Materialist sub-categories.

2.6.4.2 Reductive Materialism

2.6.4.2.1 Introduction

This particular theory has many different names. One of the main names seems to be the Identity Theory of Mind. This name comes from the idea that an identity is giving an exact account of something. So if you have two supposedly different things or concepts and you can prove that they are actually identical, then you have given an identity. So the Identity Theory posits that mind processes and brain processes are actually identical, or the exact same thing. That is, brain processes are not correlated or somehow tied to mind processes, there is no difference. There is perhaps some form of misunderstanding when discussing mind versus brain processes, so discussion of brain processes may be intended to actually be discussion of mind processes and vice versa. "The identity theory of mind is to the effect that these experiences just *are* brain processes, not merely *correlated with* brain processes" [Smart, 2011].

Another main way of referring to the Identity Theory is using the term Reductive Materialism. The goal of reductionism is to be able to reduce anything down to its essential, physical component. So Reductive Materialism is reducing the concept of the mind and mental processes down to their physical base, which is the brain. Everything that is mind can be reduced down to being in fact brain. Although there are different terms used to label this theory, ultimately the meaning is very similar.

The main people credited with popularizing Reductive Materialism are three philosophers who were publishing in the 1950s. J.T. Place (1956), Herbert Feigl (1958) and J.J.C .Smart (1959) each authored a paper that would go on to become one of the main landmarks for the idea of Reductive Materialism.

2.6.4.2.2 Qualia

The term "qualia" is used to describe the subjective feel of experience. One of the most famous to discuss this issue was Thomas Nagel in his 1974 paper on "What is it Like to Be a Bat?" Although the concept of qualia or the personal character of experience had existed previously to this point [Lewis, 1929], Nagel's paper does an excellent job of summarizing the issue. That is, although it is possible *to try* and imagine what a bat experiences, it is entirely possible that humans are simply not equipped to be

able to understand the experiences that a bat will have. So the act of imagining will give only the experience of a human trying to understand a bat, not what or how the bat itself will actually experience the world [Nagel, 1974].

And taking this conclusion a step further, it is not really possible for anyone to truly understand the experiences of another, because of the deeply personal and subjective nature of each individual's experience. Although fellow humans have more in common with each other than a bat, there is still a wealth of factors that contribute to the deeply individual and personal nature of each person's experience. The question can be raised that there is perhaps no way truly to convey this subjective experience, or qualia. The apparent existence of qualia is an important stumbling block to the theory of Reductive Materialism, since if all processes of the mind are simply reduced to brain processes, there is still the challenge of qualia or qualitative experience. The works of Place, Feigl and Smart often take this challenge into account when laying out the basics of their theories.

2.6.4.2.3 J.T. Place

J.T. Place published a major Reductive Materialist paper in 1956, "Is Consciousness a Brain Process". In this paper, Place argues that modern physicalism is often simply behaviourism and sets about attempting to prove the possibility of consciousness as a brain process, with the caveat that the experiential quality of consciousness is part of an inner brain process, not separated out the way various forms of dualism would hold [Place, 1956].

2.6.4.2.4 Herbert Feigl

In 1958, Herbert Feigl published his own version on the Identity Thesis and the challenge of qualia. Feigl used the analogy of magnetic fields, an intangible substance described solely by physical laws, to explain how physical laws can be used to explain mental states [Feigl, 1958] Feigl also talked about creating new definitions of mental as subjective and phenomenal, while physical could be defined as that which is intersubjectively confirmable [Feigl, 1958]. For Feigl, emergence comes from the idea that subjective experience is not something that we can predict, since it emerges. For example, we understand the physical structure of perfume and how it interacts with the

nose physically, but we can't predict the "quality of the experienced perfume" [Feigl, 1958]. However, Feigl thinks that physicalists can finesse this problem by extrapolating from other previously known experiences [Feigl, 1958].

2.6.4.2.5 J.J.C. Smart

J.J.C. Smart published the third and latest of the significant Identity Theory papers in 1959, entitled: "Sensations and Brain Processes". Smart's arguments centre on the idea that when people are reporting on sensation, they are not simply reporting on physical behaviour but instead are actually reporting on brain processes, although the reporter themselves may not be aware of this [Smart, 1959]. So when a person reports on being in pain, despite not explicitly saying "I am having a brain processes that indicates pain", this is exactly what this report means. Smart replies to a number of traditional dualist objections to this type of argument as well as discusses the issues around sensation or qualia being over and above simple brain processes [Smart, 1959].

2.6.4.3 Eliminative Materialism

2.6.4.3.1 Introduction

Eliminative Materialism is a Monist theory from Philosophy of Mind that had its main roots in the works of mid-20th century philosophers, such as Wilfred Sellers, W.V.O Quine, Paul Feyerabend and Richard Rorty [Ramsey, 2012]. The premise of Eliminative Materialism or Eliminativism comes from the more general sense of Eliminativists being philosophers who are essentially denying something [Ramsey, 2012]. In this case, the concept that is being denied or challenged is that of mental states. These mental states have long been held by common sense to exist and be an integral part of the human mind. However, Eliminative Materialists hold that in fact these common sense ideas of mental states and processes, such as beliefs and desires, are in fact false and should be replaced with more accurate and truthful concepts.

2.6.4.3.2 Folk Psychology

Eliminative Materialists often use the term "Folk Psychology" when referring to the type of common sense ideas regarding concepts such as beliefs and desires as mental states. The concept of Folk Psychology is one that encompasses general terms about the types of states that make up the mind as well as abstract types of mental laws

or causal relationships [Ramsey, 2012]. These are the types of abstract ideas that would not seem at all out of place in a casual or everyday conversation regarding matters of the mind. However, since Eliminative Materialism is based on the hypothesis that our common understanding of the mind is actually entirely false, many Eliminativist arguments are specifically against the logic of folk psychology. However there are several different ways in which these arguments may be structured.

2.6.4.3.3 Types of Eliminativism

There seem to be two somewhat conflicting reasons for eliminating mental states. The first belief is that mental states don't exist. For example, Feyerabend argued that folk psychology mental states would actually be non-physical and that this view conflicts with any kind of true physicalism. Thus theories resting on the tenets of materialism would have to entirely deny all previous folk psychology mental states [Ramsey, 2012]. Thus mental states, like many previously held concepts on the nature of the universe and science, should simply be regarded as debunked and archaic theories.

Different approaches focus more on replacing the concept of mental states with more accurate, modern and scientific representations. For example, Quine wanted to work on replacing the concepts of folk psychology with a "more accurate physiological account" [Ramsey, 2012]. Often the requirement for a more accurate explanation takes the form of a neurophysical description for mental notions.

"Given these two different conceptions, early eliminativists would sometimes offer two different characterizations of their view: (a) *There are no mental states, just brain states* and, (b) *There really are mental states, but they are just brain states (and we will come to view them that way)*" [Ramsey, 2012].

2.6.4.3.4 Later Eliminativist Writers

Later proponents of the Eliminative Materialism viewpoint continue the search for answers in the field of neuroscience. Paul Churchland is a theorist who holds this position. In his paper "Eliminative Materialism and the Propositional Attitudes" [Churchland, 1981], Churchland discusses both the appeal of folk psychology as well as the weaknesses of such an ancient and relatively unchanged set of beliefs. Despite the

advancements in understanding of such areas as: learning, intelligence and altered states of consciousness such as sleep, the basic concepts of folk psychology have changed very little. Churchland also discusses the difficulties that various other theories of consciousness have with folk psychology. Ultimately he holds that while folk psychology could be viewed as a theory of behaviour, it should be held separate from the physical details [Churchland, 1981].

2.6.5 Functionalism

2.6.5.1 Introduction

Functionalism is a theory of mind that it is concerned not with the substance of the mind but the functions and relationships of the sub-processes within the mind. Thus for Functionalism, discussions on the substance of the mind are in some ways irrelevant. In this way Functionalism is actually compatible with forms of both materialism and dualism. Dualist theories that hold that mental processes can be caused by and effect the physical brain are in agreement with functionalism [Levin, 2010]. However, it seems that functionalism has possessed a particular appeal to many theorists who arrive from a more materialist viewpoint [Levin, 2010].

Since Functionalism is not concerned with discovering the substance that is responsible for mental states, the focus must remain firmly on the types of states and processes that create the overall conscious mind: "what makes something a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part" [Levin, 2010]. In fact, Functionalism can be seen as being less "chauvinistic", that is, making fewer assumptions about what is mental, than some of the previous views of consciousness, which claimed that only the human brain was capable of mental states [Levin, 2010]. With the focus strictly on the mental processes and states of the mind rather than the substance, the possibility of the different implementations arises. Similar to the multiple realization argument first seen by Aristotle, for Functionalists, there are multiple possible ways in which mental states can be caused or realized, which opens many exciting possibilities to potentially simulate or create a computational implementation that can

give rise to consciousness. It is easy to see why many materialists have found the concept of Functionalism so attractive.

An important consequence, however, of this freedom from substance, is the issue that it can be challenging to discuss specific details of implementation for functionalists. After all, due to the principle of multiple realizations, it is entirely possible that there could be multiple specific and unique implementations. Thus, Functionalists must focus on developing a strong and complete description of the identity and causal roles of mental states. So there must be a sense of causal relations between mental states and "sensory input". The example of pain is often used to illustrate this. Pain serves a function by having a causal relation. It is caused by external stimuli, such as damage to the body, and serves the function of letting someone know that the body has been damaged, which will thus cause a reaction to avoid further damage and protect the body of the conscious thing [Levin, 2010].

2.6.5.2 Types of Functionalism

According to the Stanford Encyclopedia of Philosophy, Functionalism can be broken down into three main families: '"machine functionalism", "psychofunctionalism" and "analytic functionalism"' [Levin, 2010]. However, most of the current incarnations of Functionalism are based in the computational tradition and are addressed in the Artificial Intelligence sub-section.

2.6.5.3 Turing and Machine Functionalism

Machine Functionalism is mostly associated with the philosopher Hilary Putnam, but stems from the works of Alan Turing. Turing, a mathematician and computer scientist, was responsible for the development of certain ideas that have been influential in many theories of artificial intelligence. The Turing Test, first explained in a 1950 paper [Turing, 1950], discussed the issues of thinking machines. In the Turing Test, an examiner would test both a human and a machine with no prior knowledge of which was which. If the examiner was never able to distinguish between the two, or incorrectly identified them, this could be taken as a positive indicator that the machine was indeed thinking [Turing, 1950].

The Turing Machine, first postulated by Turing in 1937, was a concept of a machine where the behaviour of the machine would be dictated by reading a series of symbols from a tape and acting on combination of the machine's current state, the symbols that are read and a defined set of rules, which allow the machine to move forwards and back across the tape [Barker-Plummer, 2012]. Putnam viewed the mind as similar to a Turing machine, where mental states and operations are related not only to inputs and outputs but also to each other [Putnam, 1967]. It was this view that would become strongly identified with the term Machine Functionalism.

Additional and important consequences that arose from Turing's work and the initial definition of The Turing Machine were the concepts of a Universal Turing Machine, Turing Equivalence and The Church-Turing Hypothesis. The Universal Turing Machine is a theoretical Turing Machine that can simulate the input, operations, and output of any given Turing Machine by reading both a set of data as well as the instructions for the Turing Machine. This Universal Turing Machine essentially provides an upper limit for computation since it has infinite storage and the ability to simulate any Turing Machine operation, which allows the Universal Turing Machine to theoretically handle any and all algorithms [Lanier, 1995]. Turing Equivalence means that any two computers that can fully simulate each other are equivalent. This engenders the Church-Turing Hypothesis, which stipulates that any algorithm that can be performed in a mechanical or computational fashion, can therefore be simulated by a Turing Machine, and thus is also Turing Equivalent [Copeland, 2008].

2.6.5.4 Psycho and Analytic Functionalism

In contrast to Machine Functionalism, "psycho-functionalists", based their theories around psychology and views that behaviour arises from a series of complex mental states and processes [Levin, 2010]. This form of Functionalism relies on the scientific characterization of processes along with the more commonsense ones, but only if there is some kind of scientific basis for the common knowledge [Levin, 2010].

Ned Block raises several objection to the new form of empirical functionalism that psycho-functionalism represents. Block argues that on one hand this form of functionalism may be too liberal in its labelling of what constitutes a mental state since there are organisms that can be said to have the state of pain without truly having the

kind of mental states that a human mind does. On the other hand, if functionalism avoids this problem of liberalism, it may often become too chauvinistic in its definitions, since mental states that have the same general character but differ in the fine details can be excluded [Block, 1980].

Analytic functionalists like: Smart, Armstrong, Shoemaker and Lewis, as cited in the Stanford Encyclopedia of Philosophy, want to deal not with knowledge that is like folk psychology but with knowledge that we have *a priori* about mental states, processes and the relations between them [Levin, 2010]. Analytic functionalists want to analyse our mental states by using functional descriptions in a manner somewhat similar to the identity theory. An argument against this identity, given by Max Black, was to the effect that we cannot identify physical states with irreducible mental properties [Block, 1980].

2.6.5.5 Role vs. Realizer Functionalism

In addition to the previous general types of functionalism, a cross cutting issue is the difference between role and realizer functionalism. A return to the familiar example of pain can be used to illustrate the differences. Role functionalists hold pain to be a high-level property beyond the merely physical, while realizer functionalists think that pain is actually identified with the low-level processes that are happening to cause pain [Levin, 2010].

2.6.5.6 Artificial Intelligence

Perhaps one of the most prominent and widely covered forms of Functionalism is the category which can be broadly described as Artificial Intelligence, or A.I. This group involves a number of different theories which revolve around the basic premise that intelligence is a computational system and can be re-created using computing machines. The term “consciousness” is not often associated with these particular theories and instead focus is regularly placed on the idea of “intelligence”. Frequently, the non-physical or emergent portions of theories of consciousness would be regarded as “magic” or “mystical dualism” [Lanier, 1995], while the empirical, physical and scientific methods of theorists such as Alan Turing often refer only to uncovering and replicating aspects of intelligence. This preference in terminology may arguably be seen as an attempt to emphasize the process-based nature of the systems and behaviour that are

being reproduced, and as a move away from some of the more theoretical and abstract concepts that are frequently the centre of more philosophical discussions on the nature of human thought and understanding.

2.6.5.6.1 Physical Symbol Systems

The concept of a Physical Symbol System draws much from the earlier works of Turing and his contemporaries. When describing a Physical Symbol System, there are several key points to keep in mind. The first is that the inclusion of the word "physical" is meant to indicate that this is a materialist-based theory where all component parts are intended to physically exist and obey the physical laws of the universe [Newell & Simon, 1976]. The second is that the word "symbol" is meant to denote a symbol that can be assigned arbitrarily to represent some kind of object or activity. However, a symbol can also be a composite symbol made of more simple symbols organized into some structure on the basis of existing rules of interaction between the objects they are assigned to represent. So the Physical Symbol System is meant to be a "machine" that can, in principle, generate and manipulate a theoretically infinite number of expressions based on symbols and composite symbol structures [Newell & Simon, 1976]. This system can then be used to obey and represent rules, and in effect acts like a Universal Turing Machine.

The Physical Symbol System Hypothesis states that "a physical symbol system has the necessary and sufficient means for general intelligent action " [Newell & Simon, 1976] and is one of the main, underlying, assumptions for not only the work of Newell, Simon and their collaborators, but for many forms of Computationalism. This hypothesis holds that a physical symbol system is not simply a component of intelligence, but is both necessary and sufficient for intelligent action.

There is, of course, the problem that the Physical Symbol System Hypothesis is concerned with intelligent action and does not use the term "consciousness". However, arguably the general intent of the term "intelligent action" would subsume the conception of consciousness that is commonly used by Materialists in general and Functionalists in particular. This subsumption could be considered credible due to the apparent fact that intelligent human behaviour would seem to require some form of consciousness of the agent.

Various researchers may contend that the Physical Symbol System Hypothesis is a strong claim regarding the ability of computational solutions to provide a satisfactory explanation of consciousness. However, Newell and Simon themselves discuss how there is no true logical link between intelligent systems and their hypothesis, and instead assert that evidence must be gathered to try and provide proof of this claim [Newell & Simon 1976]. Furthermore, they contend that this evidence is often of two different types: one directed towards the claim of necessity and the other towards the claim of sufficiency. So it is possible that there could be theories which would hold to, or be compatible with, weaker claims that might allow for *either* the necessity *or* the sufficiency of a computational answer without the much stronger claim of both.

2.6.5.6.2 The Computational Theory of Mind

The Computational Theory of Mind is one that developed from the early roots of Putnam and Machine Functionalism, and is based in large part on the idea that the mind acts like a digital computer [Horst, 2011]. This theory is one that is of deep interest to many computer scientists and is strongly associated with the field of Artificial Intelligence. One of the primary proponents of the Computationalism is Jerry Fodor. In one of the classical papers for defense of the theories of the Computational Theory of Mind, “Connectionism and Cognitive Architecture: A Critical Analysis” (1980), Fodor and Pylyshyn lay out key components of Classical Cognitive Science, which is closely associated with the Computational Theory of Mind [Fodor & Pylyshyn, 1988]. These components include the view on the nature of mental states as “representational”, that is, states with both semantic and syntactic content as well as discussion of the relationship between semantics and syntax, natural languages and learning capabilities [Fodor & Pylyshyn, 1988]. The two basic tenets of this Classical model are, firstly, that the symbol structures created by the system will correspond to actual structures in the brain and are therefore are sensitive to the rules and process of these mental structures. Secondly, the Classical model holds that the structures created must be combinatorial in nature [Fodor & Pylyshyn, 1988]. These issues are of a critical nature to computer and cognitive scientists who seek to create a computer with the power of cognition and consciousness.

2.6.5.6.3 Connectionism

Connectionism is an alternate view to Classical Artificial Intelligence that also falls within the realm of computational modelling. Although some forms of Connectionism seek to replicate the neurobiological structure of the brain using artificial Neural Networks, the more general approach is to draw inspiration from brain neurology, rather than to seek an exact replication of existing biology [Garson, 2010]. These Neural Networks very approximately simulate neuron connections in the brains using individual units that are linked together with adjustable weights on the connections. The weights can then be adjusted by the system to model the kind of feedback and activity that occurs in actual brain synapses. The success of Neural Networks in simulating learning and aspects of intelligent behaviour has attracted a wide array of proponents who were unsatisfied with traditional symbol systems and computationalism [Fodor & Pylyshyn, 1988].

Connectionism seeks to move away from traditional methods of symbol systems and manipulation and towards a method that works to model the kinds of activity that have been observed in the human brain. Connectionists look to ground their solutions in the current models of neuroscience and the biology of the brain in an attempt to produce intelligent behaviour, rather than seek to implement traditional symbol systems as do the Classical computationalist architectures. However, the extent to which Connectionism diverges from the underlying principles of Computationalism and whether or not it should simply be a different from of implementation of these principles has become a source of much debate [Fodor & Pylyshyn, 1988]. And there occasionally seems to be confusion as to whether or not the underlying neurology is the *cause* of consciousness or whether the activity itself simply *is* consciousness [Thagard & Aubie, 2008].

2.6.5.6.4 The Chinese Room Argument

The Chinese Room Argument is a famous argument raised by John Searle in his paper, “Minds, Brains and Programs” [Searle, 1980], specifically to challenge functionalist and computational views of consciousness and artificial intelligence. In this paper, Searle introduced the distinction between weak artificial intelligence, a tool to aid in studying the mind, and strong artificial intelligence, where the computer is a mind [Searle, 1980]. In the Chinese Room Argument, Searle sets up a thought experiment

with a person in a box who has no real understanding of the Chinese language, but who was given a specific program or set of instructions to manipulate Chinese. Searle argues that a program would have neither necessary, nor sufficient conditions for understanding, and thus any machine dependent on such a program would not have true understanding either. For Searle the simple equation of "mind is to brain as program is to hardware" [Searle, 1980] is not enough to capture the mental causal relations. Intentional states must be defined in terms of content, not form, and thus the strong AI model is not sufficient for true mental processes and consciousness.

In his initial paper Searle raises several common and famous objections to the problem of the Chinese Room Argument and provides what he believes are full replies to these arguments. For many functionalists, as well as theorist who hold to computational views of consciousness, Searle's argument has remained a constant challenge. More recent replies, such as Harnad (1993), have continued the discussion.

2.6.5.6.5 Harnad and The Total Turing Test

Stevan Harnad is a theorist who believes that Searle's Chinese Room Argument does not deliver an unrecoverable blow to the possibility of Computationalism. Harnad defines computationalism as the idea that "cognition is computation", but seems mostly interested in developing machines that have a human type intelligence or consciousness [Harnad, 1993]. Harnad returns to the "systems reply" to the Chinese Room Argument, which is that the system as a whole could have understanding even if the "person" in the room doesn't [Harnad, 1993]. Harnad's adaptation is that there is no way to prove that the system as a whole doesn't understand, due to an "other minds"-like problem. You have to be the system to know if it understands. As a further addition, Harnad proposes a "Total Turing Test" (TTT), which he believes is immune to Searle's Chinese Room Argument, by the addition of a robotic component which allows a system to manipulate and interact with the world. It is this component that will allow a system to move beyond simply accepting inputs and instead cause understanding and true cognition [Harnad, 1993].

Searle replied to Harnad in his 1993 paper: "The Failures of Computationalism", where he notes that the true issue for lack of understanding is that the brain has "internal causal powers" which will be lacking in any kind of computational system [Searle, 1993].

And without a system that accurately duplicates these powers, Harnad's modified Total Turing Test response is doomed to the same failures as the original systems Searle noted when first discussing the Chinese Room Argument. "If I in the Chinese Room don't have any way to get from the syntax to the semantics then neither does the whole room; and this is because the room hasn't got any additional way of duplicating the specific causal powers of the Chinese brain that I do not have. And what goes for the room goes for the robot" [Searle, 1993]. The problem is that input is only syntax and lacks semantic content, and without this semantic value there is no real proof of understanding or consciousness, or as Searle puts it: "Behavior plus syntax is not constitutive of cognition" [Searle, 1993].

Harnad later responds to Searle's criticisms by acknowledging that although the Total Turing Test is not a definite test for consciousness and that the other minds problem is one that is not necessarily solvable, there are certain "dead end signals" for understanding as well as positive signals that indicate the possibility of consciousness and thus encourage further research [Harnad 1993b]. Harnad believes that Searle is undervaluing the empirical value of these tests, and their worth in providing insight into the brain by attempting to create systems that mimic these brain processes [Harnad, 1993b].

2.6.5.6.6 Dennett's Multiple Drafts Theory

Daniel Dennett, an American Philosopher, wrote extensively on his functionalist theory of consciousness in his book "Consciousness Explained" first published in 1991. In this book, Dennett points out several of the main flaws in traditional views and worked on developing his own theory of consciousness that could be said to be modelled on previous works in functionalism and computationalism.

One of Dennett's specific points is the problem of the concept that he titles the "Cartesian Theater" or a lingering belief inspired by Descartes that there is one place in the mind where the mind and body meet and come together to create consciousness [Dennett, 1991]. Dennett believes that the Cartesian Theater is a fallacy and that there is no central locus of the mind. It is a common idea of both Ryle and Dennett that mental events do not require the kind of illumination of the Cartesian Theatre in order to

be seen [Akins, 2002]. Instead, Dennett introduces his own theory of “Multiple Drafts” [Dennett, 1991].

In the Multiple Drafts theory, Dennett proposes that we have many different processes that all act in parallel. Once an “edit” has been made by one stream, it is not passed on to a central processor as the Cartesian Theatre model would hold. Instead there are multiple drafts all being edited by different streams at different times [Dennett, 1991]. What Dennett seems to be saying is that consciousness is entirely subjective and based on the content that we create using our beliefs based on mental processes which are stimulated by the physical world. Thus he holds a view of consciousness as largely perceptual [Akins, 2002]. Dennett holds the following true, in direct conflict with the Cartesian model: “there are no fixed facts about the stream of consciousness independent of particular probes” [Dennett, 1991]. Dennett seems to be quite materialist in his views of the nature of the physical universe as being objective, therefore input from physical sources would be the same, and since human observers all have the same kind of mental processes, it would not at all be unusual that similar conclusions would be reached. However, a common complaint is that although Dennett spends most of his time discussing how these various mental processes combine, there is very little explanation on why or how consciousness developed from these disparate processes.

2.6.6 Non-Reductive Physicalism and Property Dualism

2.6.6.1 Introduction

Having discussed the rise of various forms of Materialism that started in the mid-20th century, it is now time to take stock of a set of theories that can be loosely grouped together under the title of Non-Reductive Physicalism. This group is so titled due to the theories’ similar viewpoints that consciousness, while rooted in the physical matter of the brain, cannot simply be reduced down to straightforward brain processes. Instead the answer will require more subtle and nuanced models for how the physical brain is able to produce the not-quite so physical conscious mind.

2.6.6.2 Donald Davidson

Donald Davidson was an American philosopher who had two important contributions to the field of Philosophy of Mind in the 1970s. The first one was the development of a personal theory of the origin of consciousness, which would be called Anomalous Monism. Davidson's second major contribution was to widely popularize the use of the concept of supervenience in consciousness studies.

2.6.6.2.1 Anomalous Monism

The theory of Anomalous Monism was first developed in Davidson's 1970 paper "Mental Events" [Davidson, 1970]. The theory introduced held that similar to other theories of Monism, for every mental event, there is a corresponding physical event. However, Davidson also argued that there could not only be no laws to explain the behaviour of mental events, so that not only are mental events irreducible to physical events, but it is this very lack of laws that ensures that mental and physical events are identical [Stubenberg, 2010].

2.6.6.2.2 Supervenience

The origin of the concept of supervenience is difficult to pin down precisely, as it had been used occasionally by mid-19th century writers as well as by Lloyd Morgan, a British Emergentist in the early 20th century [McLaughlin & Bennett, 2011]. However, the term did not gain popular use until used by Davidson in an argument to support his theory of Anomalous Monism [McLaughlin & Bennett, 2011]. The idea of supervenience is that certain non-physical properties are entirely dependent on the physical. So if you change the physical properties of an object, it will cause changes in the non-physical properties as well, or put another way, two objects that are identical in non-physical properties must also be identical in physical properties.

2.6.6.3 John R. Searle and Biological Naturalism

As concepts such as supervenience and emergence were popularized in the latter half of the 20th century, more philosophers were looking to include these concepts to adapt previously strictly physicalist ideas. One such theory was presented by John R. Searle, an American philosopher, who became known both for his critiques of strict physicalist theories, such as Computationalism, as well as his strong support for a theory

that was based in neurobiology and yet contained references to emergentism. Searle would come to refer to his position as Biological Naturalism [Searle, 2004].

The characteristics of Biological Naturalism are a belief that the physical brain is the seat of the mind coupled with the view that consciousness or mental phenomena are higher level or emergent properties of the brain. In Searle's words: "consciousness is a causally emergent property of the behaviour of neurons" [Searle, 1992]. Searle is neither a dualist, nor a materialist, though there are aspects of both theories involved in his position. Instead Searle seems to want to find a middle ground between the two, drawing from the study of processes and neurobiology of materialism, and yet with discussion of the mind, thoughts and the subjective quality of experience as central and integral parts of the discussion.

For Searle, when discussing consciousness, there are three main questions that need to be discussed: how the brain causes consciousness, what part of the brain is responsible for consciousness and the causal relationship between consciousness and our behaviour [Searle, 2011]. In order to provide a meaningful explanation of consciousness, Searle requires not only a plausible, neurobiological explanation for the types of brain processes involved in consciousness but discussion of such aspects as how and why the brain creates a "unified field of consciousness" [Searle, 2005].

2.6.6.4 Property Dualism

Property Dualism is a distinct type of dualism, separate from the Substance Dualism proposed by Descartes. In Property Dualism, theorists commonly hold that although there is only one type of substance, that is physical, there are in fact two kinds of properties: mental and physical [Robinson, 2011]. An influential contemporary philosopher who follows a rough kind of property dualism is David Chalmers.

2.6.6.5 David Chalmers and Naturalist Dualism

David Chalmers is an Australian philosopher who has written extensive critiques on the current state of explanations of consciousness. Chalmers holds that many of the most prevalent theories do not in fact actually answer the question of what consciousness is or how it arises. Chalmers has written papers that seek to more clearly define the actual problem to be solved as well as laying the groundwork for future

theories of consciousness. Chalmers himself seems to fall into the camp of philosophers who see consciousness as biologically based, but is deeply concerned with those aspects of consciousness for which a simple, materialistic viewpoint is not sufficient.

2.6.6.5.1 Hard Problem of Consciousness

In his landmark paper "Facing Up To The Problem of Consciousness", Chalmers discusses the differences between the "hard" and "easy" problems of consciousness [Chalmers, 1995]. Chalmers wants to differentiate between the so-called easy problem of finding a process to explain such conscious behaviour as learning, information processing and other seemingly sub-processes of consciousness, and the hard question of handling the subjective or phenomenal quality of experience. Chalmers believes that there have been many attempts to provide explanations for these easy problems, but that for the hard question, few attempts have been made [Chalmers, 1995]. In fact the existence of the hard problem of consciousness has even been disputed.

2.6.6.5.2 Naturalist Dualism

Chalmers, despite being grouped under the category of Property Dualism, entitles his particular theory as Natural Dualism. For Chalmers, materialist solutions cannot ever fully capture the essence of consciousness, nor its emergent character [Chalmers, 1996]. But despite his belief that consciousness is not truly reductive, Chalmers seeks to demonstrate that "consciousness is not logically supervenient on the physical" [Chalmers, 1996].

For Chalmers, there are two basic questions that must be answered with respect to consciousness: the existence of consciousness and the specific character [Chalmers, 1996]. In Chalmers' view, previous theories seem to frequently confuse the two problems or simply ignore one in favour of pursuing the other question. Behaviourism, Computationalism and Functionalist views all seem to ignore the phenomenal character of consciousness, though they often provide good accounts of causal and behavioural processes [Chalmers, 1996]. Thus Chalmers is looking for a theory which rejects these materialist views and diverges somewhat back towards the concepts in dualism.

2.6.6.5.3 Neural Correlate of Consciousness

In later works, Chalmers looks to contemporary work on attempts to map out the neural net of the human brain and find some kind of systematic correlation between mental states such as thoughts, emotions and desires with patterns of neural activity. In considering these studies, Chalmers seeks to adequately define exactly what a Neural Correlate of Consciousness or NCC should look like [Chalmers, 1998]. In endeavouring to find a clear, necessary and sufficient definition of an NCC, Chalmers is also looking to lay out guidelines for methodology and criteria regarding future work. Although Chalmers points out that a Neural Correlate of Consciousness does not necessarily give an explanation or understanding of consciousness, the ability to isolate some kind of correlation could provide vital insights in the relationship between the brain and consciousness. Chalmers eventually arrives at the following definition:

"An NCC is a minimal neural system N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient, under conditions C, for the corresponding state of consciousness" [Chalmers, 1998].

2.6.6.6 Epiphenomenalism and Problems of Causation

Epiphenomenalism is the idea that while physical events may cause mental events, mental events in turn have no causal effect on the physical [Robinson, 2012]. This is a serious problem for the vast majority of the theories that rely on some sense of emergence, supervenience of modified dualism. Jaegwon Kim notes the possibility of mental events being labelled epiphenomena in his book "Mind in a Physical World" [Kim, 1998]. After all, if mental events can have no true effect on the physical, then what is their purpose? Without some kind of causal chain, it seems that the importance of mental states and events is drastically reduced. The risk of mental events being branded as nothing more than epiphenomena is an important challenge to all modern non-Materialist views.

2.6.6.7 Summary

This section covered theories of consciousness that mostly originated in the latter half of the 20th century. These theories all had in common the idea that strict materialism could not answer the types of questions that had been raised with discussion of qualia

and the deeply subjective nature of experience and consciousness. Consciousness and the mind could not simply be reduced to a set of brain processes. However, this is not to say that physicalism should be entirely ignored. In fact, most non-reductive or modified dualism theories are also deeply concerned with the physical nature of the brain and seek inspiration and answers that can only be provided by scientific knowledge and insight.

2.6.7 Quantum Theories of Mind

Originally, quantum theory was adapted to try and add an element of randomness into discussions of the brain to keep them from seeming completely deterministic. One, somewhat controversial, reasoning was to try and introduce an element of free will into the mix [Atmanspacher, 2011]. However, there have been a number of different theories on where and how the quantum events will occur and be involved in creating consciousness, either as a whole or through specific processes. The theories of Beck and Eccles which discuss information processing at the synaptic cleft [Atmanspacher, 2011], exemplify this. Unfortunately, the relevance and utility of the quantum processes and their relationships to mental states remain fairly unclear [Atmanspacher, 2011].

The collaboration of Stuart Hameroff and Roger Penrose has provided another interesting discussion of the possibilities of quantum events being an integral part of consciousness. Hameroff, an anesthesiologist, had been studying the science and ability of anesthetics to cause a state of unconsciousness. His studies lead him to believe that the direct effect of anesthetics on microtubules in the brain would suggest that in some manner consciousness was seated in these microtubules.

Meanwhile Penrose, using Gödel's incompleteness theorem among other tools, had arrived at a conclusion that consciousness is in fact non-algorithmic, and thus non-computable. This lead to a development of a proposal "to relate elementary conscious acts to gravitation-induced reductions of quantum states" [Atmanspacher, 2011]. The collaboration of Hameroff and Penrose would thus hold that Hameroff's microtubules would create a protective environment in the brain that would allow for the kind of quantum events Penrose proposed as causing consciousness [Atmanspacher, 2011].

However, once again this theory relies on details of quantum mechanics and gravity that far exceed our current understanding of the topics. Hameroff and Penrose's work remains controversial and ultimately inconclusive until further advancements in quantum understanding allow for more definitive answers.

2.6.8 Summary

The modern era has been a time of exciting growth in the study of consciousness and the creation of the field of Cognitive Sciences. The number of potential theories has expanded rapidly, as have the number of controversial discussion points.

2.7 Summary

This survey chapter has covered ground in the field of Consciousness ranging from the early theories of the pre-Socratic philosophers to modern and still evolving work. The intention of the discussion of a multitude of positions was to provide a strong grounding in area and a framework for further discussion in later chapters of the thesis. The following chapter will introduce an alternate structure for arranging theories as well as a number of criteria to be used in future discussion.

3 Salient Features Overview

3.1 Introduction

In this Salient Features Overview chapter a set of salient features will be introduced to study, categorize and evaluate the various theories of consciousness that were previously introduced in the survey chapter. Although it can easily seen that various forms of criteria can be useful for a categorization kind of endeavour, the purpose of this thesis is to attempt to meet several specific goals which have informed the selection of this particular group of salient features.

The first, and general underlying desire for these salient features, is to study theories of consciousness with the aim of measuring the possibility of attempting to create a successful computer implementation or simulation of a theory. However, implicit in this study of consciousness and the various theories that attempt to define consciousness is the fact that it is entirely possible that there cannot be a successful implementation of consciousness, or at least not one that is within the limits of our current technology. This limitation implies that although a theory may capture the abstract insights of some parts of the explanation of consciousness, there may still be some kind of restriction that prevents the duplication of the factors and requirements involved. However, it cannot be denied that the possibility of creating an implementation or simulation of consciousness is a predominant motivation for this thesis.

Having stated this goal of implementation, the spectre of technological or understanding limitations still exists although it would seem somewhat rigid to hold that a theory of consciousness that precludes implementation cannot be considered to be a successful theory. So perhaps a better question to ask is: What exactly is a successful theory of consciousness? When we try to define success, we discover that it is a much more ambiguous and difficult-to-grasp concept than may seem on first consideration. For example, since many of the theories are fundamentally in disagreement, trying to

define “success” may seem to be picking one theory over the others and then judging all theories by their ability to try and duplicate the “successful” theory. However, since we don’t currently know which, if any, of the given theories provides an appropriate and accurate description of consciousness, it seems that the choice of the most successful theory could be somewhat arbitrary or biased.

Instead, this thesis will seek to define a number of salient features, inspired from the various areas of research around the topic of consciousness that might, if not directly identify the one and only most successful theory, then at least provide some pointers towards which theories have the most positive indicators of success. This could also perhaps be stated as an evaluation of which theories are the most internally cohesive and robust in the face of various challenges.

3.2 General Overview of the Salient Features

When considering potential salient features, there were several factors to keep in mind. The first issue is that the salient features selected needed to be acceptable to a wide range of theories. That means that salient features which were only applicable to one type of theory were not a good choice. The key to this concept of salient features is finding features that are applicable to as wide a range of theories as possible. This will allow us to make some useful comparisons as well as to provide enlightening and interesting connections between theories that previously seemed to have little to no connection.

Another key point when considering selection of salient features is that the salient features selected should attempt to in good faith not needlessly challenge the basic tenets of any particular theory. So a salient feature that is clearly designed to entirely eliminate a certain theory or group of theories would violate this precept. The aim of the chosen salient features is to try and examine theories on as much of their own merits as possible. Since there is no true knowledge of which theory, if any, is actually correct, choosing any salient feature that eliminates a particular theory seems entirely counterproductive to the endeavour at hand.

As a final point, the salient features selected are not meant to be a necessary and sufficient set of requirements for adequately defining consciousness. Instead these salient features are meant to be positive indicators of theories that have done an acceptable job of handling general questions raised in the field of consciousness as well as addressing questions and challenges that are targeted at the particulars of the theory itself.

To begin with, will be an introduction of a concept that I call a “spectrum”. This spectrum is a high-level way of organizing the various theories of consciousness. The spectrum allows theories that previously had few points of comparison to be oriented on a single axis of comparison, which in turn allows for further and deeper analysis. The spectrum is indeed the backbone of all work in the thesis for studying various theories and how they relate to each other.

Turning now to a consideration of the specific salient features which form the main focus of this chapter, let us begin with what may arguably be one of the most widely known conceptions, namely, the Turing Test, as well as a more recent variant, the Total Turing Test as introduced by Stevan Harnad [Harnad, 1993a]. The Turing Test has a long history of discussion in the field of Artificial Intelligence and thus seems to be a prospective salient feature.

Another salient feature pertains to qualia, or the private and experiential nature of consciousness. The general concept of qualia, or “why consciousness has a personal quality”, has been a long-held question for the study of the mind and consciousness. However since the latter half of the 20th century, the general acknowledgement of the importance of the question of qualia has grown significantly. Especially with the formulation of the “hard problem of consciousness” by David Chalmers in 1996, the question of qualia, or the experiential nature of consciousness can no longer be brushed aside. Therefore how a theory chooses to handle the issue of qualia, or how it chooses to dismiss qualia is an important factor in discovering how thorough any theory is in accommodating all aspects of the issue of consciousness.

Not surprisingly, given the overarching aim of implementation to this thesis, the question of implementability is itself one of the salient features. Although not all theories

have the potential for a computational implementation, it is undeniable that consciousness is somehow attached to human nature. This means that there is at least one form of implementation, namely the biological instantiation of humanity. While a goal of this paper is definitely to further the search for possible computer implementations or simulations, if this turns out to be impossible, then the given theories should at least address how the human instantiation is possible. It is a given that some of the ideas of human implementation will rely on sciences that are perhaps not yet fully understood, not simply biological sciences but also sciences that may embrace non-physical things, much like the magnetic field example given by Feigl [Feigl, 1958]. Nonetheless, the attempt of a theory to set forth possible avenues of logic or understanding is very important.

The last major salient feature to be discussed in this chapter is that of systematicity. The concept of systematicity is one that has been discussed in a number of different areas of the study of cognition, specifically in A.I. based attempts to implement computational theories of language processing and thought. This thesis will argue that the concept of systematicity has strong potential for application in non-computationalist theories as well and is in fact an integral part of human cognition and understanding.

3.3 Spectrum

3.3.1 Spectrum Overview and Origin

When contemplating the various different theories of consciousness, it is difficult to conduct any in depth comparison since the basis and beliefs of these theories vary so widely. Although there is discussion of traditional arguments and problems in the field of the theory of mind that must be addressed in many of these theories, there are few clearly defined arguments or commonalities that can be applied to all of the various theories in question. For example, Behaviourism's tight focus on external behaviour as manifested in the Turing Test, Functionalism's causal requirements, the non-physical character of Dualism and the strictly physical character of Materialism, all demonstrate the wide span of motivations in the theories. With the centre of attention focused on

such disparate matters, any kind of broad analysis seems to be difficult. This is perhaps a quality that comes from the development of the field of consciousness, as so many of the theories have developed as a direct attempt to address weaknesses identified in previous theories. While the history of theory refutation has created a timeline of the development of theories of consciousness and thus clearly defined temporal relationships between theories, it is a challenge to try and define other relationships within the field. If we are to attempt this feat it would seem that some kind of common ground would be useful. This would manifest as a rough approximation that demonstrates the orientation of the various theories of consciousness, not only to each other, but perhaps to some kind of objective and measureable factor.

The idea of a kind of "spectrum" developed originally as an intuition that at a high level of abstraction many theories seemed to almost "line up" along some axis. There seemed to be a rough ordering in the way that theories could be placed upon this axis by classifying theories based on a sort of intuitive similarity. For example, Behaviourism, Functionalism and the various incarnations of Physicalism seemed to have a certain degree of similarity, while Cartesian Dualism seemed to have very few commonalities with these physically-based theories. After placing the theories into a rough ordering, it would seem as if a spectrum of theories could be identified along one dimension. While this is spectrum idea seemed to be a very attractive outcome since it would create an empirical way to compare theories, the idea of the spectrum was still quite vague. One of the main challenges would be to give a precise definition for this dimension, or the empirical measurement used to categories theories. What was the spectrum actually quantifying and what was the axis upon which these theories were placed?

3.3.2 Development of Spectrum Metric

A first attempt at elucidating this spectrum idea was to look at the range of strictness when characterizing consciousness. That is, theories that fall upon the less strict end of the spectrum have less difficult to satisfy or fewer necessary conditions that are required in order to meet the theory's definition of consciousness. It follows of course that theories at the other end of the spectrum will have more stringent core characteristics for what it means to be conscious. So, when looking at each end of the spectrum, and viewed prima facie, Behaviourism would have very few requirements,

namely external behaviour and would be at the least strict end. In contrast, Dualism would have a larger number of relevant properties that would include physical and non-physical qualities and would therefore be at the opposite and most strict end of the range. Another, quantifiable way of stating this idea of strictness of definition could be the number of successful tests that needed to be passed in order for a creature or machine to be labelled as conscious.

However, when looking at a spectrum that classifies theories by measure of strictness of the conditions for consciousness, a number of possible problems arise. The first is that this seems to be almost too simple and coarse-grained of an ordering based on the traditional camps of materialism versus dualism. A spectrum based on number of requirements could also be vulnerable to issues with the granularity of requirements. Would a large number of very narrowly focused requirements have been a strict definition? Or would one loosely defined idea with deep consequences be stricter or looser than a very clear, high level requirement like a proper definition of non-physical consciousness? What about the idea that many of the same requirements could be potentially applied to many different theories? If several theories had the same number of requirements, therefore the same level of strictness, but each had a different sub-set of requirements, how would they be ordered?

So, perhaps simply using an arbitrary number of conditions, core characteristics or some other measure of strictness, which is assigned to a theory, allowing it to be placed in a single, linear line is not the best way to categorize theories. It seemed as if this method would result in a loss of the subtle differences between theories. Even as a first approximation, the lack of clarity was too great.

3.3.3 Final Spectrum Metric

A later and more successful attempt to define what a possible spectrum might measure was inspired when considering the importance of the “other minds” issue [see section 2.5.10.1] and how it was addressed in the various theories of consciousness. When discussing the other minds problem, the question involves the issue of externally, or publically observable indicators of consciousness in others. The degree to which

these externally accessible behaviours are relevant to a theory of consciousness could in fact be what the spectrum should be measuring.

Theories that are deeply concerned with the other minds problem and how to answer the challenge often involve discussion of factors that are not simply externally observable but instead have an internal or subjective quality. Non-physical theories of consciousness would fall under this characterization, and constitute one end of a spectrum that measures an internal vs. external split. At the extremely externally observable end of the spectrum would fall theories that have no interest in the internal character of consciousness or simply ignore or dismiss the other minds problem altogether. It seems that this proposal of a spectrum that runs from an entirely externally observable characterization of consciousness to one that is entirely private and inaccessible to external measurements, most accurately captures the broad range of theories in philosophy of mind and hopefully the kind of fundamental differences in opinion that define the area of debate.

3.3.4 Spectrum Organization

So now we have a working explanation of what the spectrum is measuring, but the question of how it orders specific theories still requires answering. For although the concept of the degree of relevance of externally observable salient features seems adequate for classifying theories of consciousness in a clear manner, there is still the issue of how to order theories that seem to fall at the same level in the internal vs. external spectrum. Flattening these conflicts into a single dimension would seem to once again create confusing and even misleading results. While, granted, the spectrum is supposed to be a high level approximation only and thus a certain lack of clarity seems a given, in order to be in a strict linear ordering, additional core characteristics would need to be identified to differentiate between theories that had identical placements on the spectrum. And the consequences of poor selection or evaluation could involve unintended implications regarding the nature of these theories and the subtle differences between them.

So instead of a single axis, perhaps other configurations might better capture the true relationship between theories of consciousness. One possible arrangement could

involve the use of concentric circles. This would allow several theories to sit in the same ring or region, and would also imply a kind of gradual build up or change between the various regions. However, concentric circles are often associated with set theory and could therefore indicate a much stronger sub set/super set relationship than actually exists between theories. Although theories such as dualism that have non-publically observable factors also may have publically observable factors, it is not correct to imply that Cartesian Dualism is in any way a super-set of the tenets of entirely physically based theories.

What is a useful take away from the concentric circle configuration is the idea that theories may fall into similar regions or planes of the spectrum, without requiring a strict and formal ordering. This region idea or partial ordering may create looser associations between theories, but it also allows for broader links that may encompass small groups of multiple theories. And discussion at a finer or increased level of granularity may serve to tease out more information on the connections between theories in a close grouping.

At this point, it may seem natural to discuss the full range of theories of consciousness and where they are positioned on the spectrum, however given that the spectrum is more of a structure for organization of theories in preparation for evaluation, it seems appropriate to first feature the reasoning for the various salient features of assessment before performing a detailed examination of all theories in the following chapter. The subsequent sections will therefore give an explanation for the selection of each individual salient features.

3.4 Salient Features

3.4.1 Turing Test and Total Turing Test

The Turing Test was selected for a salient feature with which to analyze theories of consciousness for several reasons. The first is the wide spread knowledge and entrenchment of the Turing Test. Although there have been multiple criticisms of the Turing Test, and revisions and objections such as the idea that the Turing Test only measures “human” understanding [Harnad, 1993a], the answer to these objections often

seems to involve a revision or update of the Turing Test rather than completely throwing it out all together.

The Turing Test was originally created by Alan Turing [Turing, 1950] as a way of gaining insight into whether a computer could exhibit characteristics of a thinking being. A human tester would sit down at a console and begin a conversation through the computer, not knowing if they were talking to a fellow human or in fact a computer program. If the human tester is, in principle, unable to discover that their conversation partner is in fact a computer, or if they believe their conversation partner to be another human when it is in fact a computer, this program can be said to have passed the Turing Test.

Although the Turing Test was created as a way of testing for artificial intelligence, it is arguable that the test is indeed applicable to any theory of consciousness. While Turing's concept of artificial intelligence may not be sufficient to be considered equivalent to the consciousness as espoused by many of these theories, the sentiment is quite similar. Turing was looking to test the kind of quick response, representation and understanding of concepts that are many of the same issues considered by consciousness theories. So, while many would argue that passing the Turing Test would not be sufficient for consciousness, it is nevertheless a positive indicator of the kinds of responses that we would expect from a conscious being. I believe that this is expressed very well by Harnad when he talks about the idea of positive guides and dead signal in his response to Searle's Chinese Room Argument [Harnad, 1993b].

"Harnad's Total Turing Test" is a variant of the original Turing Test, and has been proposed by Harnad as a response to Searle's Chinese Room Argument [Harnad, 1993a]. Searle himself claims that this variant is simply the rehashing of the systems objection that Searle had addressed in his paper "Minds, Brains and Programs", when initially introducing his Chinese Room Argument [Searle, 1980]. The value of this variant to the Turing Test depends perhaps on how convincing one finds the two differing positions. Should a theorist side with Searle and hold that the Total Turing Test makes the same kind of mistakes regarding understanding as the original Turing Test [Searle, 1993] then there is seemingly no reason to actually use Harnad's version when the original test would suffice. However, it can be contended that the Total Turing Test can

involve action, and not simple verbal exchanges which could imply an extra dimension of understanding in order to complete the interaction to a level of satisfaction for the examiner. At this point where behaviour and interaction are extremely complex, the line between syntax and semantic may perhaps become indistinct. In such a situation, Harnad's arguments may seem compelling enough that the Total Turing Test is perhaps the more appropriate gauge.

Many of the traditional arguments against the Turing Test would indeed still apply to the inclusion of this salient features as a measure of the completeness of a theory's treatment of the difficulties surrounding consciousness. For example, when a human is conversing they would expect human type responses [Harnad, 1993a] so the Turing Test is really only testing for human type consciousness. Similar arguments have been made that consciousness could exist in a being who lacks our sensory perceptions, or who has entirely different perceptions and representations [Thagard & Aubie, 2008]. These are credible objections that the Turing Test is not necessary for proving consciousness, but may in fact be entirely misleading. However, any use of the Turing Test is not intended to give definite proof of consciousness. Passing the test is in no way sufficient to establish consciousness in even a limited sense. Similarly, a failure of the test is not enough to entirely rule out the possibility of some form of consciousness. Instead the Turing Test constitutes a general indication for further investigation. And perhaps its utility is best limited to testing for human type consciousness. We may argue that should the applicant fail the Turing Test there may indeed be consciousness, but it is not a consciousness that our human minds have the possibility of grasping. This raises interesting questions. However, since the Turing Test would be but one of several salient features, it would effectively be a symptom of the possibility of a humanlike consciousness, rather than a definitive conclusion one way or the other.

Regardless, the Turing Test and Total Turing Test are interesting and useful criteria to be used when analysing theories of consciousness. How a theorist would respond to the need for their theory to meet the Turing or Total Turing Tests could potentially provide interesting information on how consciousness can be assessed and identified.

3.4.2 Qualia

The concept of qualia pertains to the experiential or phenomenal quality of experience. It is this non-reducible and very personal characterization of experience that authors such as Nagel [Nagel, 1974] and Chalmers [Chalmers, 1995] believe requires a deeper explanation than is currently given by many theories of consciousness.

It seems that the question of the personal, private, and experiential nature of experience has long standing roots, yet often theories choose not to approach the question at all. However, if we are to find a rigorous and correct theory for consciousness, it seems that any good candidate must form some kind of explanation. This explanation may be of the nature that qualia is an illusionary phenomenon, but if so, what are the processes that are responsible for the illusion? And where does the illusion exist? The issue of qualia may also perhaps be explained away as a by-product of consciousness instead of an integral part of experience, and the emphasis on this phenomenon is far too strong. Again, however, a thorough treatment of why the phenomenal quality is merely an interesting addition, and the ontological status of the by-product, should be explained in detail. Arguably, without any kind of discussion on this specific qualitative nature of consciousness, a theory could be considered incomplete.

3.4.3 Implementation

As previously mentioned, implementation is of deep interest to this thesis. When discussing implementation, the immediate inclination is to look to the history of Functionalism, specifically that of Computationalist theories and their strong emphasis on computational implementation. While the question of computational implementation is certainly valid, this thesis will additionally be considering the position of non-computational implementations. That is, if a particular theory has features that cannot be covered by a strictly computational implementation, it does not mean that the implementation salient feature cannot be adequately met by the theory in question. Instead, discussion of a kind of implementation that is not strictly limited to the computational could be sufficient.

3.4.3.1 Computational Implementation

Computationalism is a driving force in the field of cognitive sciences. It could even be said that the development of computationalist theories was one of the main foundations of current cognitive sciences and without the possibilities as presented by Newell and Simon the field would look entirely different. The idea that cognition is a form of computation is a fascinating idea that, if true, could provide numerous benefits. Despite a number of major obstacles that have been identified, gains have certainly been made in areas such as learning, reasoning, and planning, and the algorithms developed have provided novel and powerful solutions in a number of domains, such as: medicine, scientific study, social studies and many others. In fact, the ubiquity of computers in general life could perhaps be partially due to some of the advances in "machine intelligence" that have come out of the search for better and more robust intelligent computational programs.

When considering the challenge of general implementation as a salient feature, implementations of a computational format are certainly appropriate for several theories. Indeed, for Materialist theories, computational implementation seems to be, not only the most appropriate, but possibly a necessity. For other theories that involve postulated non-materialist components, a computational implementation could be a component of a larger implementation, or used as a simulation of the intended type of implementation.

One of the main objections to the use of computational implementations contends that a computational implementation cannot capture all of the factors involved in some theories of consciousness. For example, the non-physical factors in Dualist theories, by their very nature, are neither understood, nor potentially able to be reproduced in a physical system. There is also the idea that perhaps the physical substance with which a system is built may be responsible for the emergence of consciousness [Searle, 1992], and thus a computational system may simply not be the right substance or composition with which to create a conscious system.

While it is difficult to challenge the idea that a computational system is simply not built of the right type of stuff, an argument can be made that computational systems can be used to simulate the causal chain that *is* created by the right type of stuff. A full and complete simulation should be able to find ways to recreate the kind of interaction that is

required. Of course, this type of simulation will necessitate a full understanding of all causal factors involved and must be able to simulate these factors in a feasibly efficient manner. Thus the limitations of understanding, algorithm complexity and computing power may prohibit a full simulation; not to mention the issue of whether or not a simulation is capable of reaching true sufficiency. Also, the question arises: at what point does the line blur between a sufficiently high-level simulation and an implementation of system created on a different base? This particular problem is one that is rather complex and open-ended.

The matter of computational implementation seems very plausible with regard to Computationalist and even other Materialist theories. However, for theories that are not strictly materialist, the situation can be rather more complex. In these situations, a computational implementation may not strictly suffice, but may possibly be part of a larger answer.

3.4.3.2 Non-Computational Implementation

It often seems that discussions of implementation have been limited to Functionalist theories, and specifically Computationalist theories. So it may seem that implementability as a salient feature should strictly be limited to the computationalist section of the entire field of consciousness studies. However, despite decades of work in this area, there are still many unsolved problems when it comes to computationalist implementations for consciousness and conscious sub-processes such as language, learning, pattern recognition and others. This may suggest that the right approach to computationalist implementation has not yet been found, or that technological limitations are responsible for the lack of a single breakthrough. Conversely, an alternate view could be proposed that perhaps the question of implementation should be included as an area of discussion in a much wider manner that is not solely restricted to a computational focus. After all, an implementation that can reliably duplicate states of consciousness or even portions of the so-called abilities of consciousness would seem to be a useful tool in proving a theory's validity.

One possible objection to the inclusion of implementation as a salient feature is that it is simply a discussion of details that are far too low-level and indeed have little to do with the overall validity and robustness of a theory of consciousness. This criticism

has a certain amount of validity because any proposed implementation will naturally rely on the details of the theory itself. However, simply because the implementation will be a later development of any theory, does not mean that implementation is simply an afterthought. In fact, it could be argued that without some kind of effort to discuss details of implementation, for example, details of the biological instantiation of human consciousness, a theory is not fully developed. Since consciousness is considered not simply an abstract concept, but one which most theorists would agree is actually instantiated within the world, dismissal of implementation seems to be short-sighted.

Another potential objection is that the question of implementation will automatically exclude or be non-applicable to theories in which consciousness is considered to be of a non-physical, or mental character. In this way, the choice of implementation could be thought of a showing a bias towards Monist-type theories.

Again, when answering this challenge the discussion can come back to the fact that consciousness exists in humans. If consciousness is indeed non-physical, mental or possesses some other quality or substance beyond the physically observable operation of the human brain, then in some way it must attach or connect to the physical. This issue is often discussed when considering issues of mental causation or epiphenomena [see section 2.6.6.6]. So although the implementation for these theories might not be of a nature that is conducive to computation, perhaps a discussion of simulation or at least in depth discourse on why it will not be possible to "artificially" duplicate consciousness is appropriate. A theory does not have to present a computational program in order to provide a successful discussion of some kind of implementation. Alternatively, a comprehensive and logical argument as to why implementation is impossible or strictly limited could be important.

A further objection is that for some theories, such as Biological Naturalism, consciousness relies on some kind of quality that is potentially exclusive to the human mind, be this quality physical, non-physical, mental, some combination of factors or something entirely new.

If consciousness is indeed restricted to humans, several questions may be raised. Is it possible to simulate the human environment that created consciousness?

Although these simulations may not be sufficient to give rise to new consciousness, it may provide more information on the inner workings. Additionally, isolation of the specific human requirements that have allowed consciousness to arise would be a great breakthrough in understanding consciousness, even if duplication remains beyond our means. Again, the discussion of implementation is not required to produce a fully realized recipe for how to create consciousness in a laboratory, but should involve an examination of the details on which consciousness is built.

3.4.3.3 Implementation Summary

In summary, when discussing the issue of implementation, it is important to remember that the term “implementation” should not be limited to current, Computationalist work, but also encompass a more general, broad discussion of the actual details of recreating consciousness in some kind of medium, or at least a detailed discussion of low-level particulars for how consciousness exists and operates within the human form, such as Neural Correlates of Consciousness.

In passing, we should bear in mind that the question of the causality and how a strictly non-physical consciousness could have any effect on a physical, human body is one that has formed the basis for many objections to dualism [Kim, 2000]. Even if a theory holds that consciousness is essentially un-reproducible on a reliable basis, to attempt to truly understand consciousness will involve discussion on what physical, or other factors, create this situation of being un-reproducible and thus why all implementations will fail. Thus it seems that implementation should be a point upon which to appraise the completeness of any existing theory of consciousness.

3.4.4 Systematicity

The concept of systematicity can perhaps first be attributed to Fodor and Pylyshyn in their seminal 1988 critique of connectionism: “Connectionism and Cognitive Architecture: a Critical Analysis” [Fodor & Pylyshyn, 1988]. The Stanford Encyclopedia of Philosophy states that systematicity is thus defined: "The systematicity of language refers to the fact that the ability to produce/understand/think some sentences is intrinsically connected to the ability to produce/understand/think others of related structure” [Garson, 2010]. The basic idea seems to be that Fodor and Pylyshyn believe

that humans have the ability to break down sentences into component structures and then re-assemble the components in a novel way that is still understandable. A simple example given in Fodor and Pylyshyn to demonstrate this concept is that if something can understand "John loves Mary", it needs to also be able to understand "Mary loves John", in order to successfully satisfy the systematicity requirement. The question of systematicity is not limited to such simple sentences, but also components of larger and more complex sentences as well.

It is this ability, one that Fodor and Pylyshyn hold as a basic human ability [50], to easily create and understand novel sentences, that Fodor and Pylyshyn challenge connectionist architecture to replicate. For Fodor and Pylyshyn see systematicity as a basic human skill which must be successfully duplicated in any cognitive science implementation [Fodor & Pylyshyn, 1988].

Further debate on the idea of systematicity resulted in a view, expressed in several papers by Hadley [Hadley, 1994a], [Hadley, 1994b], that there are three different levels [Garson, 2010]. The weakest is that a system could recognize sentences that are not in the original training set. One of several more robust forms of systematicity is "strong systematicity"; an example of which would be if the system recognizes "Mary loves John" without "Mary" having ever been used previously as a subject. "Strong semantic systematicity" is where the system would have to demonstrate not just understanding of the grammar, but also of the semantic content of the novel sentence.

Chalmers had also entered the debate [Chalmers, 1993], and it may be that this debate inspired the discussion of systematicity which occurs in Chalmers' paper on the Neural Correlates of Consciousness or NCC. Although the concepts of systematicity in Chalmers and Fodor and Pylyshyn are related, there are differences between the two uses of the term. Chalmers states: "The systematicity in the correlation means that it can be extended to predict the presence or absence of phenomenal features that may not have been present in the initial empirical data set, for example" [Chalmers, 1998]. So it would seem that Chalmers' use of systematicity entails the existence of some means for extending a correlation to encompass features, such as emotions or mental states that had not actually been covered in the original, empirical set of data [Chalmers, 1998]. For example, if the original NCC was calibrated using a small sample of

emotions such as fear, pain and joy, a successful NCC could then be extended to show the mechanism for which additional emotions such as pride, an emotion that had never been tested in the original findings, could correlate to certain neural patterns or activity. This kind of logical extension seems familiar within the domain of scientific methodology. If a theory can be used to predict associations that were not included in the original test set, then confirmation of those predictions would seem to only strengthen the support for a theory. For Fodor and Pylyshyn, systematicity seems to be restricted to the domain of sentence creation, whereas Chalmers seems to have extended or further developed the concept to be a more broad and inclusive. The definition of systematicity has thus evolved beyond the limits of Fodor and Pylyshyn's original definition.

Although Chalmers was arguably correct in introducing the utility of systematicity in theories of consciousness, it could be said that he stopped short of stating the true importance of systematicity. In fact, for many of the theories that would seek to find an NCC, a lack of systematicity could prove fatal to this endeavour. There are several reasons why systematicity is such a crucial part of the NCC and consciousness puzzle. First, it may be beyond our current limits of technology to empirically track all states of consciousness, and thus without the ability to systematically extend an NCC to include novel mental states, this evidence would be incomplete or too weak to provide sufficient value. Secondly, for theories that hold that an NCC is possible, a natural next step would be to test these theories through implementation or simulation. However, without some kind of systematic correlation between physically observable neural activity and conscious or mental states, an implementation may also be beyond the realm of possibility, since a simulation or program that does not contain some kind of system may be deeply impractical to implement.

Finally, despite the fact that many mental processes are poorly understood today, it seems troublesome that there is not some kind of system to these processes. After all, human behaviour, while often confusing, has many predictable factors. The degree to which behaviour is predictable can be debatable, and may be limited by our understanding of mental processes, but it seems improbable that many theorists on the subject of consciousness would be satisfied with a theory that concluded that there are no mental processes that are reliable, repeatable or reproducible. Although it would seem that this last point may perhaps be falling into the realm of applying the principles

of “folk psychology” [see section 2.6.4.3.2], even the opponents of this concept would look to replace folk psychology with some kind of scientific set of mental processes [Ramsey, Stich & Garon, 1990] [Churchland, 1981]. In fact, Davidson’s Anomalous Monism is one theory that holds there are no laws for mental events, and yet this theory has few modern proponents [see section 2.6.6.2.1].

Chalmers, despite having identified the importance of systematicity in an NCC, seems to want to avoid positing it as a necessity. Or at least, Chalmers wants to allow for the possibility of consciousness without systematicity. “Of course we may hope that there will be more constrained neural systems whose content systematically matches the contents of some aspect of consciousness. But one might argue that it is not obvious that such a system *must* exist. It might be held, for example, that the contents of consciousness are an emergent product of the contents of various neural systems, which together suffice for conscious content in question, but none of which precisely mirrors the conscious content” [Chalmers, 1998]. It could be said that in attempting to allow for the correctness of theories that do not require systematicity, Chalmers is acknowledging the very real possibility that consciousness has a character very different from the more commonly accepted views. This is a very pragmatic view, however it could also be said that in attempting to permit the possibility of certain theories which may not allow for systematicity, Chalmers has ultimately undermined his own case for the importance of an NCC.

However, the separate cases for systematicity with regards to these various types of theories are perhaps best considered as separate arguments. For now, I will attempt to lay out some of the common objections to systematicity in systems that rely more on the physical for the basis of consciousness.

A possible objection to the importance of systematicity is the idea that it is only a construct for human understanding and that the mind itself does not require any kind of system to be able to create consciousness. However, a potential reply to this objection is that theories where consciousness is entirely physical are frequently grounded strongly in scientific method, which has strong ties to systematicity. After all, it is one of the hallmarks of scientific discovery that there is frequently some kind of underlying

system which needs to be discovered. After a system has been discovered, these principles can then be applied to novel data and new applications.

There is, of course, the type of debate that occurs between those who hold that these logical connections exist and merely need to be discovered and those who hold with Pragmatism [Hookway, 2010] and believe that in fact the law is being created, rather than discovered by the theorist. So perhaps those who hold this latter view, that systems and laws are an invention of humanity, might indeed believe that systematicity is nothing but a human construct. However, would this view then imply that systematicity is not at all required? After all, even if systematicity is created or imposed by humans, it is difficult to deny that much of our understanding of the world is built on scientific method and principles. Just because a system has been created does not mean that it holds no truth or pragmatic value. So despite the fact that systematicity might not be intrinsic to consciousness, it appears that the importance of systematicity for purely human understanding remains. It should be said, that although there is a connection between systematicity and scientific method, the connection is not always binding. There are always surprising or unexpected results, for example the field of quantum studies often defies previous patterns of explanation. Often times, these unforeseen consequences are due to some misunderstanding or incorrect assumption, but there are still mysteries in the realm that have yet to be explained and there remains the possibility of areas for which systematicity is in no way guaranteed.

The discussion of humanity and the drive for systematicity in scientific investigation leads to a further development from this point. That the challenge to systematicity may instead be that the systematicity that human understanding seeks is false or may not adequately capture the true relationships, and that consciousness itself has no true systematicity as we know it or as we are capable of understanding.

This argument arguably has merit, and it is true that the system itself may be difficult to recognize and understand. However, the basic concept of systematicity should not be seen as a mandate on what form a system or set of logical correlations take but instead constitutes a very high-level concept that there is some kind of quantifiable or causal chain between processes and/or states. As Chalmers notes, “Perhaps one could define a ‘systematic NCC’ as a neural correlate of a phenomenal

family such that states correlate with each other in some such systematic way. I will not try to give a general abstract definition here, as things are getting complex enough already, but I think one can see a glimmer of how it might go"[Chalmers, 1998]. If it is indeed the case that the underlying mechanisms of consciousness are beyond human understanding, then it is conceivable that we will never find an NCC, which is a troubling concept for those who seek definitive answers. If one is to subscribe to this belief on human limitations with regard to understanding consciousness, then arguably no human theory will ever be able to satisfy.

So, while it would seem that systematicity is virtually entailed by physically based theories, what about theories which are not quite as deeply rooted in the physical, and thus are not as strictly tied to scientific methods and conclusions? This would include theories that see consciousness as emergent, supervenient or reliant to various degrees on some kind of derivative, phenomenal or problematic character will can be loosely grouped under the term "mental".

Chalmers considers the possibility of theories where systematicity seems completely incompatible:

> In answering, I will assume that states of consciousness depend systematically in some way on overall states of the brain. If this assumption is false, as is held by some Cartesian dualists (e.g. Eccles 1994) and some phenomenal externalists (e.g. Dretske 1995), then there may be no NCC as defined here, as any given neural state might be instantiated without consciousness. (Even on these positions, an NCC *could* be possible, if it were held that brain states at least correlate with conscious states in ordinary cases). But if the assumption is true, then there will at least be some minimal correlation of neural states with consciousness [Chalmers, 1998].

Yet it seems that his ultimate conclusion is that there must be some kind of minimal systematicity.

When considering the question of systematicity as it relates to supervenient or emergence-based theories, it would seem that few conflicts would arise. After all, these theories hold that somehow the non-physical nature of consciousness is grounded in the physical properties of the mind [see section 2.6.6.2.2]. And this tie to the physical leads us back to the idea of systematicity being a large part of the physical world and laws.

However, a strong objection could be raised that although consciousness may rely on the physical, it emerges with its own property so there may not be a systematic connection between the physical and non-physical. That is, there may indeed be a causal connection between the physical and the non-physical consciousness, but this connection may not be grounded in any kind of systematic way.

While this means that systematicity is not necessary to these theories in the same way as physical theories, it does not mean that systematicity is entirely incompatible. It could in fact turn out that the non-physical has logical and causal connections that form a new type of systematicity. Indeed, without either the demonstrations of systematicity in emergent or supervenient theories or the explanation of why this systematicity is not possible, many questions regarding the validity of a theory would no doubt remain.

As a further point, despite the fact that much of our knowledge of consciousness is uncertain and highly debatable, few theories of consciousness indicate a belief that the causal relations between mental states do not show some kind of systematic character. There are two distinct but related points on which systematicity or consciousness can be demonstrated. The first is the way that consciousness seems to be reliable in humanity, e.g., the fact that consciousness exists in all human types and we expect a new human to be able to have it, or gain it as they grow. As systematicity was defined, it is the ability to extend to novel situations beyond the originally verified ones. We are able to extend our idea of consciousness beyond existing humans to encompass all new born infant humans that may be created. And we expect the conscious experience of these new humans to be relatable in some way to our existing experience of conscious humans. So although external behaviour in these novel situations may be different to what we have previously seen, it is rarely entirely incomprehensible. This point does however lead into the kind of territory that is often covered in the “other minds” debate [Hyslop, 2010].

The second point regarding systematicity and consciousness itself is the idea that there seems to be systematicity within conscious experience. It is a strong common sense presupposition that there is some kind of system, logic or chain of causation that underlies our mental states and processes. So for example, if one remembers the same

memory multiple times, one can assume that, barring new data, there will be the same kind of emotions associated with it each time. Therefore this idea that the same chain of mental events or associations will happen seems to indicate that there is some kind of system or systematicity in the purely mental character of consciousness, regardless of whether or not there is systematicity in the physical or even in the tie between the physical and the mental. In fact, many arguments regarding theories of consciousness rely upon this underlying belief about the intrinsic requirement of causation.

3.5 Summary

This chapter has provided an overview of the various salient features to be used in evaluating depth and completeness of theories of consciousness, a discussion of the development and reasoning behind each of the salient features, as well as notes on how the salient features may be applied to various theories at a high level. The next chapter will contain further in-depth analysis of specific theories using these salient features.

4 Details of the Spectrum and Salient Features

4.1 Introduction

In the previous chapter of this thesis, a high level introduction of a number of salient features for evaluating theories of consciousness was provided. That previous chapter was meant as an overview of the reasons for selecting each of the salient features, and an examination of the general validity and appropriateness of these salient features with regard to the types of objections that could possibly be raised. The current chapter is meant to provide a detailed analysis of a number of specific theories of consciousness using the previously introduced salient features. The reason for selecting these salient features was an attempt to try and provide several objective points upon which to debate the completeness of theories of consciousness. This chapter will be a further step towards performing this analysis, with the goal of perhaps finding some interesting points of discussion.

It is important at this point, to once again emphasize that the aim of evaluation by these salient features is not to identify a single theory as the most successful or to perform some kind of ranking on the various theories in the field. The scope of such an endeavour is far beyond this thesis. Instead, the goals of these salient features are to provide a foundation for future study, to suggest some kind of methodology for this study and perhaps to find some interesting ideas that the theories may have in common.

4.2 Chapter Organization

In this chapter, there will first be a detailed discussion on how the various theories will fit on the spectrum, starting from theories that focus upon what is publically observable and moving to theories that are almost entirely concerned with the private or non-publically observable. This work in defining the spectrum and arrangement of

theories of consciousness on the spectrum will help lay out a structure for comparing the theories.

After that, the following sections of this chapter will attempt to apply the specific salient features to each of the theories of consciousness, following the rough ordering as defined in the spectrum chapter. When discussing how a salient feature applies to a theory, the salient feature is intended to be an important measure of whether an agent is judged conscious by the particular theory. Each section for a specific salient feature will cover such topics as previous discussion by the various theorists, possible objections to the salient feature or suggestions on how such a discussion could potentially provide value to a theory.

4.3 Spectrum

The spectrum of theories of consciousness provides a high level method of placing all theories of consciousness along a single axis, to allow for comparison and to illustrate relationships between theories. In the previous chapter, discussion of the spectrum concept covered the development of the spectrum and how the spectrum is attempting to capture and place theories, from those that have a focus on the externally observable nature of consciousness to theories that have a focus on the internal and introspective nature of consciousness. This chapter will move on to a more in depth and detailed discussion on how various theories fit on the spectrum. It is important to remember, however, that the spectrum was originally intended, and is most effective, as a high level abstraction only. When discussing theories at a very fine level of granularity, the general ordering of the spectrum becomes far more complex and the planar ordering of various “clouds” or groupings of theories becomes a far more ambiguous subject.

In this section, the ordering of theories according to the spectrum from most externally ascertainable to least externally ascertainable will proceed roughly as follows: Behaviourism, Materialism, Quantum Theories, Non Reductive Physicalism, Cartesian Dualism and finally Anomalous Monism (see Fig. 1). For many of these general categories, there are cross cutting issues which will require more in depth discussion and may often have multiple interpretations. This section will sometimes present several

different arguments for the positioning of sub-theories within a grouping and even for the mixing of certain theories from different umbrella categories.

To begin with, let us start at the most external and arguably objectively observable end of the spectrum with Philosophical Behaviourism.

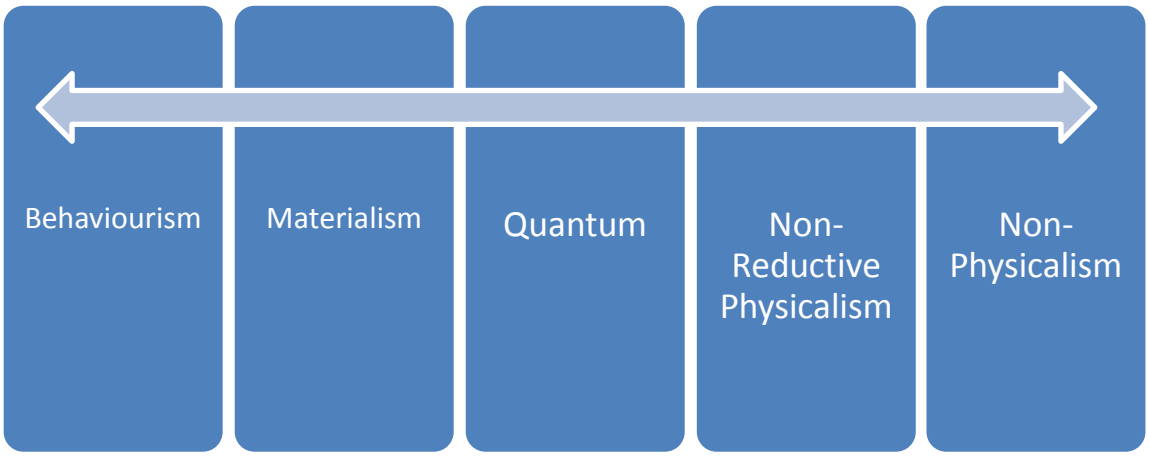

Figure1. High-level placement of theory groups on Spectrum

4.3.1 Behaviourism

At the most externally observable end of the spectrum sits Behaviourism; this is due to the basic tenet of philosophical Behaviourists that consciousness is merely a set of externally observable behaviours [see section 2.6.3]. For Behaviourists, there are very few further questions that need to be answered regarding the nature of consciousness. Since behaviour is, by definition, the external actions that can be viewed by others, and lacks any true sense of introspection or awareness of the internal nature of behaviour, it is clear that the rather straightforward characterization of consciousness by the Behaviourist theory is an ideal candidate to set a benchmark for one end of our spectrum.

4.3.2 Materialism

Moving a bit further along the spectrum towards theories that have a more internal and introspective account of consciousness, but are still very firmly grounded in what is externally observable, comes most of the various forms of Materialism. Since the general Materialist philosophy for consciousness holds that there is only the material or physical substance that makes up consciousness, it seems fairly clear that there is still a strong influence on externally measureable and observable factors. Indeed it may seem at first glance, that Materialism is equal to Behaviourism on the spectrum. It might be argued that behaviour is a relevant indicator for Materialists, and therefore an outward measure of behaviour should be sufficient for consciousness for Materialisms, which would place them roughly equal to Behaviourists.

The difference, however, lies in the fact that for most Materialist theories, there is some sense that the question of "mental" phenomena, that is, the internal kinds of qualities and processes of the mind, must be discussed and explained. If it could be said that Materialists would be satisfied simply by systems that mimic behaviour, then the theories would seem to be almost identical to Behaviourist theories. Of course, various theories differ on how these mental phenomena occur or exist, but this need to explain and identify internal processes is what differentiates Materialism from Behaviourism and is why the general grouping of Materialism is a step away from the strictly external end of the spectrum.

Since there are a number of cross-cutting issues and interpretations on the concept of mental within the rather large and complex grouping of Materialism, it is now perhaps a good idea to attempt to examine in more detail some of the various types of Materialism to see where they sit in relation to each other on the spectrum.

As a quick reminder, Reductive Materialism is closely tied with the Identity theory and holds that mental and physical states are actually the same. Eliminative Materialism is concerned with debunking the commonly held views of mental states, or folk psychology, and seeks more accurate and scientific explanations for the brain processes that underlie these incorrect views.

Because all of these versions of Materialism seem concerned with uncovering specific processes within the mind or brain that will then provide an explanation for the mechanisms of consciousness, at first there seems to be very little difference between them in terms of external versus internally observable factors. However, there are differences that when examined may provide subtle changes in overall positioning. Especially since there are potentially multiple different ways of viewing these differences as more or less internal.

To begin with, let us try placing Reductive Materialism at the most external end of the grouping, in part because Reductive Materialists seem deeply concerned with removing any idea of mind as separate from brain [see section 2.6.4.2], and indeed some Reductive Materialists such as Smart [Smart, 1959] hold that sensations are simply a form of brain process and not some correlated mind process. Reductive Materialists contend that through advances in science, and with the removal or collapsing of the concept of "mind", what would be left would be externally measurable and quantifiable brain states. There may, of course, be some work in explaining how consciousness has adapted to provide some kind of overlying interface, as in the reporting example of Smart [Smart, 1959], but is seems that Reductive Materialism would be almost entirely external.

Next could be the theories of Eliminative Materialists, who want to deny the existence of the common sense view of mental states, but seem to have varied positions on exactly what to replace folk psychology with [see section 2.6.4.3]. Although the common view seems to be a need for a more accurate, and scientific explanation for brain states, Eliminativists like Paul Churchland [Churchland, 1981], seem to think that there may be separate positions within psychology or behaviour that should never be confused with conscious brain states, but may still have a place in the discussion. Unlike Reductive Materialists, Eliminative Materialists may argue for a complete restructuring of our view of mental states, but there seems to be a sense that it may not be entirely possible to collapse all concepts currently classified as "mental" down to merely the physical. In this way it could be argued that there is the possibility for more discussion of internal and introspective issues than with Reductive Materialism. It certainly seems that the issues of experiential, sensory or qualia do not arise as

frequently as an objection to Eliminative Materialists as it does for Reductive Materialists.

However, reversing the positioning of Eliminativism and Reductionism on the spectrum seems plausible if the focus is shifted to different details of specific theories. For example, Feigl seems to want to distinguish between the mental as subjective and the physical as more objective or intersubjectively confirmable [Feigl, 1958]. This seems to imply that a discussion of the mental would have a separate and distinct position from that of the physical and gives the possibility of a more internal slant to his version of Reductive Materialism. In contrast, Eliminative Materialism could take on a more strict view by calling for the entire removal of any discussion of mental states.

It would seem that when considering how to place individual theories on the spectrum, the mixing of Reductive and Eliminative Materialism will be inevitable due to the subtle and complex debates which make up each individual theorist's position (see Fig.2).

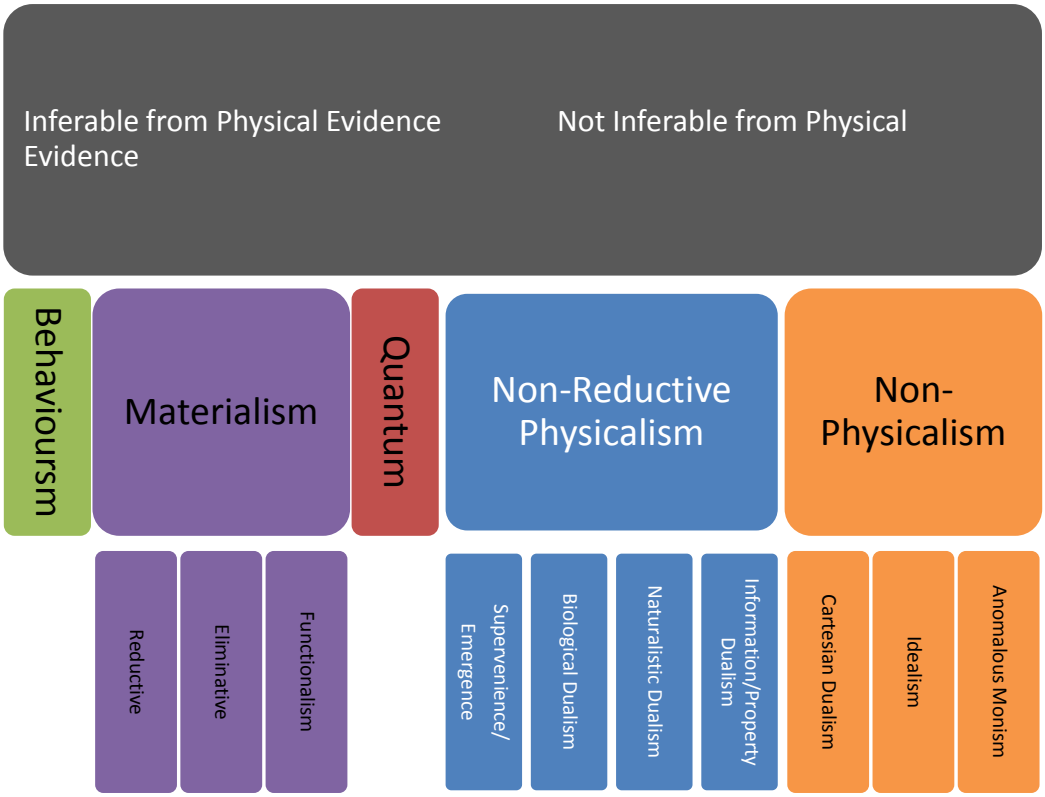

Figure 2. Possible detailed Spectrum organization

4.3.3 Functionalism

Let us now consider the Functionalists. Functionalism is concerned with discovering the underlying functions and causal processes of the mind and brain and due to the focus on function rather than nature, not all Functionalist theories will necessarily be Materialist theories as well. Although the majority of Functionalists may also be Materialist, there may indeed be some theories that are also compatible with consciousness as emergent or not entirely physical in nature. However, due to the large amount of overlap between the two theory groups, Functionalism can be closely arranged with the various Materialist theories.

The Functionalist group is perhaps the most sizeable amongst the Materialist-related sub-groups, as Functionalism has many forms, including the large number of Computationalists working on computational implementations or simulations of conscious processes [see section 2.6.5]. So, is Functionalism placed on the more externally verifiable end of the Materialist grouping or does its relative lack of chauvinism

regarding the anthro-centered nature of mental substance position it between some of the more strictly Materialist theories? Functionalism's emphasis is on roles, and causal relations between mind processes seem to provide a very liberal view and no true judgement of whether the mind is mental or physical. So in this way, it could be said to rely on less external factors because the mind could be entirely mental, which is not necessarily externally observable. However, the Functionalist drive to understand how the various functions causally interact seems to have a strong focus on externally observable factors, since it would seem that uncovering function would have a large component of external observation. And indeed it could be seen that those Functionalists who view the mind as computational would argue that there is little that is introspective or subjective to the mind that is not caused by some form or computation. So perhaps the discussion might come to demanding a more precise definition of exactly what is implied by the term "scientifically observable or ascertainable". Does a precise and comprehensive description of an internal, subjective and introspective process somehow make this process transparent and externally measureable? Nagel would perhaps disagree with this suggestion, since his entire discussion of this issue emphasized that while one might imagine another's experience it is not the same as actually experiencing it oneself [Nagel, 1974]. Explanation of how another's mental processes work does not exactly equate with externally, objectively measuring these processes.

Again, when examining the differences in position that define a general group of similar theories, the spectrum begins to grow dramatically in complexity and loses some of the straightforward clarity that it initially provided. Computational Theories of Mind would fall somewhat more towards the external end of the spectrum. However, something like Dennett's Multiple Drafts theory, with the discussion of the various forms of mental revision (Stalinesque or Orwellian) that occur and how these can cause interesting effects with time and our own perceptions, in fact requires much more discussion of internal, but still physical, effects [Dennett, 1991]. It would seem that Functionalist views are similarly tangled and intertwined with Reductionist and Eliminativist views (see Fig. 3).

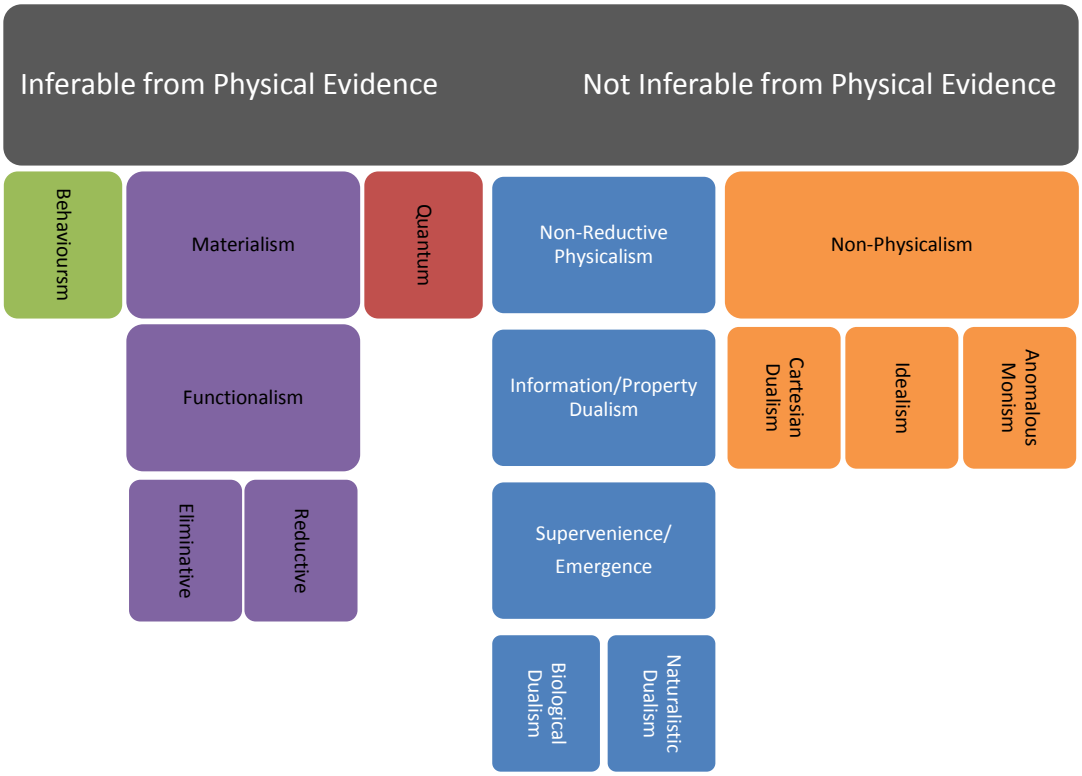

Figure 3. Alternate potential organization of Spectrum

4.3.4 Quantum Theories

Moving away from Materialism, we can now discuss Quantum theories of consciousness. Quantum theories are an interesting example of a set of theories that can move around the spectrum and take on differing placement depending on both viewpoint and the level of detail on which the theories are examined. On the one hand, Quantum theories are grounded in the laws of physics, meaning they are deeply concerned with physical and externally verifiable and measureable phenomena. However, the nature of Quantum physics, and our relative knowledge of it, means that the details of placement on the spectrum can be somewhat ambiguous. Certain Quantum theories seem to have a not entirely physical component to them, for example, those that use the unpredictable nature of Quantum events to introduce ideas of free will, or rely on the unusual properties of Quantum events to explain the emergence of consciousness. These theories, due to the seemingly intangible and introspective nature of the non-physical factor could be placed towards the internal or non-externally observable end of the spectrum.

On the other hand, it might actually be more appropriate to place the Quantum theories directly on par with general Materialist theories and then split placement of specific theories when considering the spectrum at a finer level of detail. There is also the issue that currently there is much that is not entirely known about Quantum events and how they could relate to consciousness. Of course, many physical-based theories rely on a hypothesis regarding the inner workings of the biological, chemical and neural aspects of the brain, but this problem seems especially acute for Quantum theories. For now, due to the fact that Quantum events have been added to theories of consciousness often specifically for their unusual and often unpredictable nature, it suggests that Quantum theories perhaps have room for internal and introspective elements and thus will not be grouped with strict Materialist theories.

4.3.5 Non-Reductive Physicalism

At this point in the spectrum, a switch occurs away from strictly Monist or Materialist theories into the realm of Non-Reductive Physicalism and towards entirely non-physical theories of consciousness. In terms of being externally versus non-externally observable, the non-reductive physicalist theories will divide strictly materialist from non-materialist theories. So, as we continue to travel along the spectrum from external to internally observable explanations for consciousness, let us now consider the non-reductive physicalist group. It is important to remember that the theories in the Non-Reductive Physicalism area have in no way entirely abandoned externally ascertainable factors; rather they place them on equal ground to discussion of the subjective, internal and introspective processes of the mind.

As discussed previously [see section 2.6.6], non-reductive physicalist theories can indeed be based upon or closely tied to the physical and most of the theories that will be discussed fall into this category. The question is more the degree to which consciousness arises from the physical and how much of what we term “consciousness” is actually non-physical. The Non-Reductive Physicalism category covers theories for which consciousness is emergent or supervenient on the physical, as well as certain modified forms of Dualism such as Property or Information Dualism.

When looking again to place theories starting from the most externally observable end of the spectrum, theories that are more strongly tied to the physical will most likely be placed at this end. Theories relying on supervenience, where the mental or supervenient character of consciousness is entirely reliant on the physical character could be classified as partially external due to the essential and intrinsic ties of the physical. The challenge is often to explain how these ties have come about and identify the specific brain components or correlations. Theories that look to find a specific Neural Correlate of Consciousness would ultimately need a full explanation of neural and biological processes, however such a theory would not be complete without a rigorous study of the supervenient, and correlated mental processes. Issues of how these mental states could have causal effects on the physical are just one of the questions that supervenient theories must take into account [Kim, 1998].

In a similar position, although perhaps not within an identical area of the spectrum, would rest something like Searle's Biological Naturalism [Searle, 2004]. Searle seems to not want to be tied to the specifics of supervenience, but instead offers a more generally emergent theory. Searle seems to be searching for specifics of a neuro-biological explanation that will allow consciousness to emerge as a higher level property.

Various forms of Dualism could also be placed in this general area of the spectrum, including Chalmers' Naturalistic Dualism [see section 2.6.6.5.2]. Chalmers' view of consciousness seems more concerned with issues of the phenomenal nature of consciousness or in Chalmers' own words: the hard problem of consciousness, although this emphasis may perhaps be due to a perceived deficiency in many of the current Materialist theories. However, Chalmers does not deny the importance or impact of the material and externally observable role the brain plays in forming consciousness.

Similar but distinct theories of Property and Information Dualism could be considered a further step towards the privacy oriented, non-observerable end of the spectrum simply due to the focus that is often applied to uncovering and explaining the mental side of the dualist equation. How these non-physical, mental, problematic or otherwise characterized properties are connected to the physical, and why "the physical" is not sufficient to explain these properties seems to be the main drive of positions.

However, an alternate view is that these modified forms of Dualism could in fact be considered to posses more externally verifiable aspects than other Non-Reductive Physicalist views due to the former's strict restriction of purely mental aspects to only a small portion of the overall system that forms consciousness. Therefore a more limited commitment to the non-physical would satisfy these forms of Dualism, and potentially a less radical definition of consciousness could suffice. The concern with internal factors could be said to be less pervasive or encompassing than some other emergence-based theories.

4.3.6 Non-Physical Theories

At the least physically-grounded level of the spectrum comes a variety of theories for which consciousness is entirely non-externally measureable. Cartesian Dualism falls into this region as it is the most extreme and demanding form of dualism due to the very strong separation of mind and body as defined by Descartes. Given this orientation, any theory of consciousness would have many hurdles to overcome to explain in sufficient detail the nature of the mental state of consciousness and the causal interactions of the mental upon the physical body.

Idealism is another theory that would fall into this category, as it holds that all substance is the world is entirely mental, with no physical nature. An argument could be made that Idealism in fact belongs much further towards the externally observable end of the spectrum, since even if the entire world is mental there must still be some kind of empirical way of measuring mental events due to existing logically sound observations of the world. This categorization would perhaps depend on how compelling the arguments of Idealism are held to be, and since in current times Idealism is a theory that has largely fallen out of favour, it is perhaps more appropriate to place it at the opposite end of the spectrum from Behaviourism as a kind of bookend.

4.3.7 Summary

This section has been an attempt to present a more detailed discussion of how the Spectrum of consciousness can give a big picture understanding of the complex and multifaceted field of Theories of Mind. Although there has been discussion throughout

this section on the limitations of the spectrum when utilized as a fine level of granularity, nonetheless, there does seem to be quite a lot of value for the spectrum as a tool of discussion and debate. Even when attempting to position two similar but distinct theories, interesting issues have arisen that allows for a new point of view to ongoing debate.

4.4 Salient Features

Having defined the spectrum as a structure for organizing theories, it is now time to turn our attention back to the salient features that were defined in the previous chapter. This section will delve into a more thorough analysis of theories of consciousness using these salient features. As a reminder, please bear in mind that the salient features are meant as a collection of indicators towards theories that have successfully tackled some of the major issues inherent in the field of consciousness, but no one individual salient feature is meant to be sufficient proof that a given theory is correct. Additionally, the set of salient features is not meant to definitively select a single theory but instead provide potential direction for further study.

4.4.1 Turing Test

The Turing section of this chapter will focus on the use of the Turing Test and Total Turing Test to evaluate potential implementations for typical theories. So, if some kind of implementation or simulation was created to test the general hypothesis and specific details of a particular theory of consciousness, would passing the Turing Test provide some kind of relevant indicator of consciousness, or at least the possibility of consciousness? If the Turing Test provides a test that is perhaps too weak, the Total Turing Test could be applied as an additional test to see if any extra information or context may be gleaned from the additional components of the Total Turing Test.

As a note, when specifying that the Tests should be applied to an implementation of the theory, it should be clarified that the implementation or simulation does not have to be a computational or machine based implementation. This may seem counter-intuitive since the Turing Test was originally created for computational implementations.

However, since the original test also specified that the tester and test subject would be separated [Turing, 1950], it seems plausible that the test subject is not necessarily required to be a computational machine. Although the Total Turing Test specifies the details of the transducers, again it seems possible that there could be a certain amount of leeway in how these transducers, and the system they are connected to, could be implemented. Therefore, although some theorists no doubt hold that any behaviour should be able to be simulated computationally, if for some reason a portion of any theory of consciousness should preclude a computational implementation, it does not mean that no implementation is possible. Thus when using the term "implementation", the usage will be meant in the broadest sense possible, as this thesis will remain agnostic as to the actual substance, details and character of any implementation.

4.4.1.1 Behaviourism

Behaviourism and the Turing Test seem to be quite a natural match. Behaviourism is entirely concerned with the outward reactions or behaviour of a subject [see section 2.6.3] and this indeed is what the Turing Test is specifically designed to measure. Arguably few Behaviourists could find fault with including this approach as an excellent measure of success for theories and implementations.

The question is now whether the addition of the Total Turing Test as applied to Behaviourism would provide any added value beyond the original Turing Test. The use of the Total Turing Test also seems quite natural for any implementation based on Behaviourism. Behaviourism, of course, is all about the behaviour, so adding in the idea of interaction with the environment in the form suggested by the Total Turing test seems not to provide any conflict. However, is there much to be gained by the addition of the extra level of testing? It's not entirely clear. Potentially there will more of a verification element in that things tested will not be simply symbolic or verbal. However, the strength of the positive indicator for consciousness may be slightly weakened, because interaction behaviours may be easier to simulate than more abstract behaviours, such as pattern recognition or analysis. Given either the Turing Test or the Total Turing Test, it seems quite straightforward that any Behaviouristic implementation should certainly be required to pass either form of the Turing Test before being considered a successful implementation of the theory's views on consciousness.

4.4.1.2 Materialism

In general, all Materialist theories, and most definitely Computationalist theories would seem compatible with the Turing Test with little argument. The weightiest arguments against using the Turing Test for Materialist theories would perhaps be the standard argument that the test is not sufficient to indicate consciousness or that the Turing Test only indicates consciousness of a specialized type, with the external behaviour to match. So the internal processes could indeed be creating consciousness but the kind of behaviour that is measured by the Turing Test does not entail the existence of this consciousness. Indeed, these types of argument against the use of the Turing Test and Total Turing Test as positive indicators of consciousness are applicable to most, if not all, of the theories that have some kind of internal or introspective component. After all, these Turing Tests are only evaluating outward behaviour and may not be rigorous enough to give true knowledge of any internal workings, nor solve the Other Minds problem. However as stated in the introduction to the salient features section, neither any individual salient feature nor the entire set of salient features together is intended to be either a necessary or sufficient component of consciousness. So the types of questions that are posed during any kind of Turing Test should be carefully gauged to test for the kinds of introspective understanding or processes that indicate a possibility of consciousness. Nevertheless, the objection can be taken further that answers given during the test may not indicate any kind of conscious process that is recognizable for humans, but instead a type of understanding that is impossible for the human mind to grasp. A reply could be made to this that truly, there could be many different kinds of consciousness in the universe, but it is perhaps only possible for humans to grasp and recognize human-type consciousness. It seems that the challenge of identifying non-human type consciousness is a question that is far beyond the scope of this thesis and perhaps of any of the types of theories that currently exist in the field of Theory of Mind.

As to the ultimate suitability of the Turing Test for Materialist theories, it may be possible that no definite answer can be given to satisfy skeptics, however it seems that the longevity in the use of the Turing Test as well as the strong Computationalist movement within the Materialist camp would allow for a strong argument as to its inclusion.

4.4.1.2.1 Reductive Materialism and Eliminative Materialism

When considering the Turing Test and Total Turing Test for Reductive Materialism theories, there seems to be little controversy. After all, according to Reductive Materialism, consciousness is simply brain processes, and so any test which probes for specific types of behaviour to indicate some kind of underlying physical processes should be fine as a non-sufficient, positive indicator. Interestingly, the Total Turing Test might potentially have a few more objections than the regular Turing Test when applied to Reductive Materialism. Since the Total Turing Test involves transducers to give interaction with the world, it could be postulated that unless the process of the transducer gives a very close simulation of the underlying processes in the human brain, certain behaviours or types of understanding may be different. However, the strength of that objection may rely on how tightly bound any Reductive theory is to the specifics of human brain processes being required to create consciousness.

Similarly, Eliminative Materialism would seem to be close enough to Reductive Materialism for the same kinds of arguments to apply. It seems that the Turing Test and Total Turing Test could potentially provide value to these theories as a general positive indicator of a human-like consciousness.

4.4.1.2.2 Functionalism

The Turing Test as applied to Functionalist theories and implementations would similarly seem to have few objections. Although the Turing Test does not directly measure any of the internal causal relations of a particular implementation, this does not seem to be a large impediment to the utility of the Turing Test, since one of the main claims of Functionalism is that each implementation would be unique to the material and properties of the original system [see section 2.6.5] and the Turing Test is similarly adaptable to all different implementations. It could be that when applying the Test, the types of questions could be "tuned" to try and prove the specifics of the underlying processes in the particular type of implementation. However, this seems to be a somewhat unnecessary adaptation since one of the main tenets of Functionalism is that it is the *functions* which are of importance. Attempting to ascertain and differentiate between different processes does seem to be a bit of a violation of the basics of

Functionalism. There is of course the Inverted Spectrum argument [Tye, 2009], which holds that if something is perceived different, i.e. a person sees the spectrum of colour as inverted, it could be an indicator of some kind of lower-level function that is actually quite different. However, the Inverted Spectrum argument is not a definitive conclusion and may depend on the specific function in question.

4.4.1.3 Quantum Theories

There would appear to be little to any of the Quantum theories of consciousness that would differ to the general issues of Materialism with regards to the Turing Test and Total Turing Test. Of course, any implementation of a Quantum Theory would need to be able to reproduce or simulate Quantum events; however that should really have no true effect on the utility of these Turing Tests or on how the testing would be carried out. It would seem that passing either the Turing or Total Turing Tests would be a useful indicator of a certain amount of success in creating consciousness through any implementation of a specific Quantum theory.

4.4.1.4 Non-Reductive Physicalism

Perhaps the Turing Test would be considered to be more controversial when applied to theories that move beyond strict physicalism, which are theories that can be generally grouped as Non-Reductive Physicalist theories. Arguable, it is possible that the Turing Test can in fact stay quite neutral with regard to these theories. Passing the Turing Test is not a claim that consciousness actually will be present in the given subject, simply that the behaviour is a good indicator that something similar to consciousness is currently (or perhaps has in the past) occurred.

What of Searle's Biological Naturalism? It would seem logical that much of Searle's potential objection to the Turing Test is in fact raised in his famous Chinese Room Argument [Searle, 1980]. However, the strength of that argument lay in the idea that simply getting the correct reply gives no guarantee of understanding. By using the Turing Test as but one salient feature amongst a group of salient features, this thesis is in fact agreeing with Searle. There is no guarantee of true understanding or consciousness by passing the Turing Test, but there is a possibility.

For other emergent theories and for the restricted variants of Dualism, it could be claimed that the Turing Test is indeed still applicable but that perhaps the guidelines for a conversation that passes the test should become more stringent. So the discourse that occurs during the test could additionally involve the discussion of topics that include that qualitative nature of experience or ones that test the boundaries of topics that would require strongly abstract abilities. On the contrary, the Turing Test was originally designed for the questioner to ask a rather exhaustive set of questions, so perhaps a better response would be to not require the discussion of qualitative experience at all but instead limit the strength of the results of the Turing Test. So a positive Turing Test result for an emergent theory would not carry the same weight that it would for a more physicalist one, which does not have the additional requirements of somehow trying to prove a non-physical, mental or emergent character.

4.4.1.5 Non-Physical Theories

Finally, when considering theories which hold that consciousness is entirely mental, or non-physical in nature, it would perhaps be quite challenging to create some form of implementation. However, the difficulty of using the Turing Test does not seem to in anyway impede the theoretical usefulness of the test, which is simply that a successful passing of the test could indicate that there is the possibility of consciousness in the subject. In this case, the Total Turing Test might seem to be a bit of an impediment or problem for a theory like Cartesian Dualism. If consciousness is entirely mental, albeit somehow attached to the body, Harnad's suggestion that understanding, and thus some portion of consciousness may be seated in the physical transducers [Harnad, 1993] could be problematic. A response could be that the transducer could replace the pineal gland as the seat of connection between the mind and the body. However it seems that Harnad's point was that a transducer would be intimately involved in the actual act of understanding would seem to violate the essence of Cartesian Dualism. In this case, perhaps the original Turing Test would be more appropriate.

4.4.1.6 Summary

Throughout this section, the goal has been to cover in detail how and why the Turing Test could be useful to various theories of consciousness. Where appropriate, there has also been discussion of the merits of the Total Turing Test, either as a

replacement or additional test. There appears to be little in the way of conflict between the application of the Turing Test to a potential implementation and the foundations of the majority of these theories. This is somewhat to be expected considering the nature of the Turing Test is of a somewhat limited test to give an indicator of a certain level of understanding.

4.4.2 Qualia

In this section of the chapter, there will be a discussion of how specific theories line up with the qualia salient feature. "Qualia" concerns the subjective, sensory and qualitative nature of experience. In some cases for specific theories the issue of qualia was simply not one that was considered in depth by the original theorist. However, discussion starting from the end of the 20th century has put more and stronger emphasis on the question of qualia. Therefore it may be interesting to return to some earlier theories and attempt to determine how compatible with modern ideas of qualia, the theories are.

As in earlier sections, theories will grouped into the rough ordering that was determined for the spectrum of theories of consciousness, starting from externally ascertainable to non-externally ascertainable.

4.4.2.1 Behaviourism

When considering Behaviourism, as the theory that is classified as most externally oriented by the spectrum, it can difficult to see how or why the salient feature of qualia should be applied. After all, Behaviourism is almost entirely concerned with the issue of external behaviour, and qualia contains the internal and subjective character of experience. These two seem to have little in common. Perhaps the question should be instead, should the salient feature of qualia even be applied to Behaviourism?

However, it is not enough for Behaviourists to merely dismiss the issue of qualia without further discussion. At a minimum, Behaviourists should address the issue of why qualia is not actually an integral part of consciousness. Although the goal of the various salient features is to study and investigate theories of consciousness in hopes of finding new interpretations in a spirit of good faith, there is also the goal of pushing

boundaries of understanding and finding areas that have need of further examination. If the theory of consciousness for Behaviourism in no way involves subjective qualia, then what is responsible for the qualitative nature of experience and what precludes it from being either directly involved in creating consciousness. Alternatively, why is it created as a by-product of consciousness?

4.4.2.2 Materialism

When considering Materialism in general and the issue of qualia, it seems that there should be some form of discussion of qualia in most varieties of Materialism. If the basis of most Materialist theories is that consciousness is entirely of the physical brain, then the subjective nature of conscious experience should also be based in the physical. There is of course the possibility that qualia is some kind of reporting error [Smart, 1959], but again, a discussion of why this is an illusion would seem to provide no threat to the internal cohesion or basic tenets of the theory. In fact, it would seem that the challenge of qualia is well-merited and could provide opportunities for Materialists to possibly make inroads in proving the entirely material nature of this aspect of consciousness. So it seems that all complete materially-based theories of consciousness should provide discussion of qualia at some level. We can now proceed to evaluate some Materialist theories to see how the issue of qualia is handled by each.

It would seem that Eliminative Materialism might have some of the same problems with qualia as faced by the Reductive Materialists [see section 2.6.4.2.2]. Namely, if consciousness is only brain processes, then how do we explain the nature of subjective experience that seems to go beyond a simple brain process? However, it seems that the focus for Eliminativist writers is often on the issue of folk psychology, which may be considered to encompass the issue of qualia. Since folk psychology is about the common sense view of how our minds work, it could be that qualia or the subjective and experiential nature of our mental states falls within this common sense way of viewing the world. Eliminativists seek a more accurate explanation of the how the brain works, and it seems that this search may yield a better, more accurate, and physically oriented explanation for the phenomenon of qualia. While there has been discussion of qualia by such Eliminativists as Churchland [Churchland, 1989], it would seem that qualia provides interesting discussion for Eliminative Materialists and further

analysis of how it fits into the Eliminative worldview could be useful in separating from some of the other Materialist theories.

4.4.2.3 Functionalism

Functionalism and Qualia, somewhat surprisingly in comparison to other forms of Materialism, can have a certain amount of compatibility. One Functionalist view is that certain qualia in fact have a specific function, for example, pain [Tye, 2009]. That is, pain made be used as a signal of damage, or danger. So although different physical states may cause pain, they all have the same function. An objection to this is the Inverted Spectrum, the supposition that there may in fact be people who view colours entirely in an "inverted" fashion. For a Functionalist, it actually doesn't matter if we see colours different, as long as the functionality is the same. The Inverted Spectrum objection is that there may be a case where the fact that colours are viewed in a different way may actually be an indication of different functionality at a low level. Therefore different physical states that cause different pain, or different experiences of specific qualia, may indicate entirely different functions. The result of this objection could be a call for further investigation into the function of qualia and a demand for a more specific and a more comprehensive treatment of qualia in Functionalism. However, it does not seem that the Inverted Spectrum poses a significant threat to a high level compatibility between Functionalism and Qualia. Indeed the discussion of the functions and purposes behind Qualia may prove to provide interesting results.

4.4.2.4 Quantum Theories

As for Quantum theories, it seems there has been little examination of the problem of Qualia, so it is difficult to say whether or not there could be serious conflicts between the two. It is possible that the unusual nature of Quantum events could account for some of the subjective and private nature of experience, but more effort to investigate a possible connection would certainly be required.

4.4.2.5 Non-Reductive Physicalism

Non-Reductive Physicalist theories of consciousness and Qualia seem an almost perfect match. If a particular theory in this group has not discussed Qualia, it may largely be that the major development of the theory occurred before the question of

Qualia had become a wide-spread and well accepted phenomenon. However, it would seem that for any of these theories, examination of the private and subjective nature of the experience should be a considerable portion of either the justification or the inspiration for the particular theory of consciousness.

4.4.2.6 Non-Physical Theories

Similarly to the Non-Reductive Physical group of theories, Non-Physical Theories such as Cartesian Dualism or Idealism are almost entirely focused on the private, mental nature of consciousness. There is actually nothing to consciousness except this internal, subjective and introspective character. Again, if there is a lack of comment on the specifics of Qualia, it can only be that the concept of Qualia had not been isolated at the time of the development of the theories to the extent that later works would have embraced the concept.

4.4.3 Implementation

The present section will cover in detail how an implementation requirement could be used as a salient feature for success for theories of consciousness. As a point of order, implementations are not strictly restricted to computational implementations. For some theories, a computational implementation may be outside the boundaries of possibility. This does not mean that such a theory would automatically fail to provide adequate information for an implementation. Instead, a discussion of why implementation is impossible, how future developments in technology could be required, or even a discussion of the details of how consciousness is specific and limited only to a human implementation would suffice.

4.4.3.1 Behaviourism

For Behaviourism, it seems entirely plausible that implementations or simulations could be created to demonstrate the foundations of this theory. Arguably, many of the current methods of Artificial Intelligence are in fact implementing a Behaviourist approach, even if they are not actually labelled as such. Behaviourism is wholly concerned with analyzing and replicating patterns of behaviour, and it could be said that

particular A.I. systems, where the emphasis is dedicated to reproducing these effects, could be considered strongly Behaviourist.

4.4.3.1.1 Computational Implementation Assessment

As mentioned, many computational implementations that fall under the label of Artificial Intelligence could be said to be Behaviourist. The concept of codifying a set of rules corresponding to conventions of external human behaviour, as a Behaviourist implementation would require, seems very close to the kind of behaviour specified by a Turing Machine. Therefore the types of challenges of creating a Behaviourist computational implementation would be analogous to those of creating a Turing Machine. For example, a theoretical computational ceiling could be specified by the Universal Turing Machine and the Church-Turing Hypothesis provides equivalence between implementations. It would seem that the remaining challenges of creating a Behaviourist computational implementation would be familiar issues such as defining the set of behaviours and the rules and relationships between behaviours.

4.4.3.2 Materialism

Since both Reductive and Eliminative Materialism hold that the only true components of consciousness are brain processes, theoretically there should few obstacles to some form of implementation of these theories. This is not to say that the concrete realization of an implementation should be simple and straightforward. There is much that is still not understood about the operation of the brain, and the processes that create consciousness are far from fully identified. However, the task of creating an implementation for these forms of Materialism seems impeded only by our current lack of in-depth knowledge of the brain.

4.4.3.2.1 Computational Implementation Assessment

It is also possible that there may be a computational implementation for these theories; or at least a computational simulation for the biological and chemical reactions that help form the required brain processes. Additionally, since Materialist theories are clear that there are only material substance and physical processes, then the conditions for Turing Equivalence could perhaps be met if this simulation of biological and chemical reactions can duplicate all consequences of the physical reactions. Thus questions of

computational implementation could be collapsed into comparable questions regarding computational, Turing Machine or Artificial Intelligence implementations in general.

4.4.3.3 Functionalism

Functionalism seems custom made for implementation. Since the functions and causal relations of the various parts of the system are the core of Functionalism, it seems inevitable that implementation should be the final goal of any Functionalist theory. In fact, many researchers in the field of Artificial Intelligence seek to create a thinking machine using the tenets of Functionalism.

For most forms of Functionalism, the question seems to be not whether an implementation is possible but how successful the various approaches are for creating an implementation. This does not make the question of defining consciousness any clearer, but it does focus the search into a scientific methodology for testing and trials.

4.4.3.3.1 Computational Implementation Assessment

When discussing implementation for Functionalism, the default form of implementation would appear to be Computational. This is due in large part to the prominence of Computationalism as a leading form of Functionalism. Considering the large amount of literature and research devoted to the topic of various forms of Artificial Intelligence, any reservations regarding the viability of computational implementation of Functionalism appear to be hasty.

4.4.3.4 Quantum Theories

The question of implementation with Quantum theories is very interesting. Since the focal point of any Quantum theory is clearly the quantum events that drive consciousness, any implementation would either have to reproduce quantum events or in some way simulate their effects. In terms of traditional, non-quantum computational implementations, simulation is perhaps the best that can be hoped for. Of course, the limitations on current understanding of quantum events and effects mean that any examination of possible computational simulations will also be limited. However, it seems that a number of Quantum consciousness theorists, such as Hameroff and Penrose [Atmanspacher, 2011] have put a great deal of effort towards speculation on

what specific events could be involved. In terms of implementation, it is this kind of discussion that attempts to work with specific constraints of a theory that can provide a theoretical model without actually creating a working implementation.

4.4.3.4.1 Computational Implementation Assessment

Quantum computation is a field designed to handle the nature of quantum phenomena. Although there may be challenges in representing the full nature of quantum events it is arguable that there has been success in representing these phenomena via computation. The extent to which quantum computation has accurately captured quantum phenomena is beyond the scope of this discussion, but it is clear that there is at least some basis for computational implementation.

4.4.3.5 Non-Reductive Physicalism

Theories that fall under the general heading of Non-Reductive Physicalism have certain rather large challenges that need to be met when discussing the issue of implementation. For theories that hold with emergent or supervenient views, an implementation needs to create the correct kind of physical structure which would allow consciousness to emerge or supervene upon it. This may or may not amount to simply having the right kind of underlying physical micro-processes and might require additional environmental, biological or chemical requirements.

For Property and Information Dualists, creating any kind of implementation has a set of similar difficulties. How can any kind of physical implementation be certain to capture the mental properties that are required? Without these non-physical attributes any implementation would be entirely fallacious in capturing the essence of the theory. Would a simulation be sufficient? The question arises, at what point would a simulation of a mental property be in effect entirely separate from an actual mental property? It seems there are too many unanswered questions and very little of a starting point from which to attack the problem.

4.4.3.5.1 Computational Implementation Assessment

Many of the Non-Reductive Physical theories rely on a strong attachment to the physical. That is, there is a physical component in which the non-physical portion of

consciousness is grounded. Therefore, it would seem conceivable that the physically grounded portion of many Non-Reductive Physicalist theories could be a computational system. However, some theories may hold that the specific nature of the human brain is what causes consciousness to emerge or supervene, in which case the computational system would need to accurately simulate the conditions of the brain as well as the brain operations and processes. However, it is contentious whether a simulation can replace the original system that is being simulated [Searle, 1980], and thus there may be a limitation on the ability of a computational system to move beyond simulation into actually creating the correct base for consciousness. Despite these unresolved issues, arguably, a computational solution could be used as at least part of a non-reductive physical implementation.

4.4.3.6 Non-Physical Theories

Implementation concerns are only magnified when it comes to theories for which consciousness is entirely non-physical, i.e. not grounded in the physical at all. Without completely solving the problem of consciousness and holding a strong and complete understanding of the mental realm, creating any kind of implementation seems impossible for Cartesian Dualism.

4.4.3.6.1 Computational Implementation Assessment

It would seem that a non-physical system would have zero ability to have a computational implementation. The only possibility is that of a simulation; however this would require quite in-depth knowledge of the non-physical factors and causal relations. Without such detailed knowledge on which to base a simulation, it would appear that computation is unrealistic as a form of implementation for these theories.

4.4.3.7 Summary

In the examination of the salient feature of implementation, it seems clear that there is a rather close tie between physical grounding and implementation. However, whether or not this tie is necessary or merely the result of less-physically grounded theories lack of attention to the problem of implementation is unclear. Although the challenges of attempting to implement factors that are non-strictly physical are very evident, the proof of a human realization of consciousness is fairly undeniable. Perhaps

more examination of the factors that limit consciousness to the human brain would be helpful in these situations.

4.4.4 Systematicity

The systematicity section will cover in detail the application of the systematicity salient feature across the spectrum of theories of consciousness. As discussed in the Salient features Overview chapter, systematicity is a concept that should most likely apply to all theories in this field.

4.4.4.1 Behaviourism

When considering Behaviourism and systematicity, it seems that few Behaviourists would take serious issue with a requirement for systematicity, as there appears to be little that would violate any firmly held precepts of the theory. One possible objection concerns the question of whether or not sets of behaviour should have systematicity. It seems plausible that after observing behaviour of a conscious being, future behaviour could be extrapolated from these observations. However, there could be a theorist who holds that conscious behaviour should have an element of unpredictability to it, and therefore we should expect novel and unexpected behaviours from any conscious being. However, it does not seem that this objection would provide a major flaw in the ideal of systematicity as applying to behaviour as a whole. Although it should be the case that a large amount of behaviour could be predicted through systematicity, it seems entirely possible that there would occasionally be surprising results. This may be partly due to changes in motivation or understanding, although these internal factors may actually be precluded from the Behaviourist view due to its focus on external matters only. But it does not seem that occasional novel and unexpected results would cause fundamental damage to the general requirements for systematicity in the general theory of Behaviourism.

4.4.4.2 Materialism

Materialist theories generally depend on the physical brain to be the seat and the cause of consciousness. Uncovering a clear and scientific picture of the functions and processes of the brain is both the methodology and goal of these theories. A

fundamental assumption for this kind of scientific investigation is that there exists some kind of system in the physical domain. Therefore when investigating the physical brain as the seat of consciousness, it would seem nonsensical to hold that there is no form of systematicity in the brain and consequently in consciousness that arises from the working of this physical brain. After all, if there is no underlying system, any kind of scientific explanation would perhaps not even be possible. Thus there should be some form of systematicity in both the Materialist search for consciousness as well as the perception of what a final Materialist answer to the question of consciousness would look like.

4.4.4.3 Quantum Theories

The discussion of Quantum theories and systematicity could be challenging when taking into account the unpredictable and not clearly understood nature of quantum events. Is systematicity possible? It seems that there is scientific method and understanding of quantum issues, as demonstrated by the quantum laws identified and quantum computers created. And when considering the rebuttal to some of the work of Hameroff and Penrose, it does seem that there could be some kind of systematicity in the types of quantum events that occur. So perhaps the form of systematicity contained within Quantum theories could be a qualified form of systematicity that handles the general nature of the quantum events, rather than a strong and detailed version.

4.4.4.4 Non-Reductive Physicalism

Theories which rely strongly on underlying supervenience will probably have the easiest time conforming to the needs of systematicity due to their strong reliance on the physical. In fact, it could be said that there is a sense of systematicity built implicitly into the concept of supervenience. Since the idea of supervenience is that any changes to the physical will result in changes to the non-physical then clearly there is some kind of systematicity. It seems fairly unproblematic that specific physical changes will result in specific non-physical changes.

Searle's vision of Biological Naturalism and its robust ties to human biology would similarly seem to sit happily with systematicity. Although it is true that Searle in no way thinks consciousness can be reduced solely to neurobiological process, he does

hold that brain processes are responsible for consciousness and believes there must be "neuronal correlates of consciousness" [Searle, 2004]. Systematicity is almost implicit in this idea of a correlate.

General emergence could be challenging with regards to systematicity, depending on how consciousness is said to emerge. If emergence is reliable then it implies that there is some form of systematicity, as some combination of physical factors will result in a combination of non-physical properties that creates or causes consciousness, similar to the concept of supervenience. However, isolating these factors and discovering the underlying systematicity remains a troubling question.

As for Naturalist Dualism, since Chalmers discusses the desirability of systematicity in his work on finding a Neural Correlate of Consciousness [Chalmers, 1998], it would seem that Chalmers believes systematicity has a place in his own theory of consciousness.

When considering Property and Information Dualism, it seems that when considering systematicity a large part of the discussion would have to be whether or not there is a system to the mental or non-physical properties. Although a discussion of underling rules and systems would no doubt involve a rigorous discussion of the specific details of the theory, it is not clear that systematicity would actually be required. Much of that argument would likely depend on how the details of the non-physical properties are realized.

There does however seem to be a strong argument with regards to systematicity and emergent or non-physically reductive theories when we consider issues of mental causation. How and why mental states could have an effect on physical states is a question that should be answered if the problem of epiphenomenalism [see section 2.6.6.6] is to be avoided. While there is no guarantee that if this causation does in fact exist, there will be some kind of systematicity to the causation, it does seem as if a solid argument for the possibility of systematicity could indeed be raised.

4.4.4.5 Non-Physical Theories

Theories in which consciousness has no dependence on physical matter, such as Cartesian Dualism, are challenging to discuss with regards to systematicity. There can be no *a priori* assumptions made simply because any understanding of the mental or non-physical realm is incomplete, and quite separate from physical and scientific investigations of the physical world. This is not to say that Cartesian Dualism and systematicity are necessarily incompatible, but in order to explain systematicity within the mental realm, a certain amount of groundwork and fundamental rules must be established.

4.4.4.6 Summary

For many theories of consciousness, the question of systematicity is one that is rarely raised in any detail. However, it seems as if work to develop the concept of systematicity could in fact be a useful tool for future dialogue.

4.4.5 Summary

This chapter provided a detailed explanation of the placement of existing theories on the spectrum of consciousness, including an examination of several alternative theory arrangements and their merits. The second half of the chapter replied upon the main ordering of theories of consciousness presented in the spectrum section to work through the set of salient features isolated in the previous chapter as potential measure of completeness.

5 Conclusion

5.1 Introduction

When first starting research on the area of consciousness, the problems seemed quite abstract and theoretical, similar to a metaphysical thought experiment in philosophy. Does it matter what the true nature of consciousness is if we were able to find or build programs that could simulate the behaviour of the conscious mind? It seemed clear that a better algorithm or a more powerful processor would be the key to unlocking the problem of intelligent machines. In fact, if a philosophical question could be of importance in the field of artificial intelligence it would seem more pressing to handle ethical questions of the morality of creating consciousness and the rights of the entities thus created. Self-aware computers almost seemed to be the manifest destiny of computer science, only held back by current limitations on technology, not understanding.

However, as clearly demonstrated by the numerous pages of this thesis and indeed the numerous pages, spread across decades, that have been devoted to the problem of consciousness, the conclusions are not quite so straightforward. Without being able to fully understand what consciousness is, it is not the technology that will prohibit the creation of artificial consciousness but understanding. While it can be argued that certain classical and modern problems to do with human consciousness, for example the Other Minds problem, may be strictly abstract and not constructive, it is these problems that point towards the missing pieces in our understanding. With the Other Minds problem, it is not simply an abstract and philosophical curiosity, but a true problem to be overcome. Is there an objective way to prove consciousness in another entity? Without this ability, can we ever truly know if consciousness is created? The act of creating such an objective test for consciousness requires understanding of consciousness.

But again, even disregarding the Other Minds problem, is it perhaps possible to stumble upon creating consciousness without truly understanding it or being able to prove it? Possibly, but this indicates some kind of accidental creation that would then perhaps be impossible to duplicate. Engineering a machine or simulation that duplicates the existence of consciousness without truly calculating and understanding the factors involved would not only seem to rely entirely upon luck but could potentially move into a huge number of ethical problems, not to mention the kind of potential consequences seen only in horror movies.

It could be argued that the centuries of study of the mind-body problem have yielded no true results and that the field has in many ways become only more confused.

5.2 Conclusions

This thesis has resulted in the spectrum, a novel organizational structure for producing an excellent high-level overview of the field of consciousness. The spectrum is based on the idea of placement of theories along a single axis that moves between two poles, one where theories are concerned with external and scientifically observable phenomena associated with consciousness, and the other pole, for which the internal, introspective and subjective aspects of consciousness are the main focus. Although the spectrum gives a clear picture of the relationship of the major theories as a coarse level of granularity, when considering the specific details of theories, the picture becomes more complex and the simple and straightforward ordering of theories becomes increasingly ambiguous. Thus, the spectrum would seem to have the most value when used as rough approximation of a loose ordering of theories.

This thesis has also concluded that there can be a number of salient features used to evaluate the completeness of theories and determine whether or not they provide a rigorous explanation of the challenges of consciousness. The salient features isolated for this thesis are: The Turing Test, Qualia, Implementation and Systematicity. The Turing Test is a well known test in the field of artificial intelligence but would seem to have benefits when applied to the greater field of cognitive sciences in general. The question of the nature of qualia has existed in one form or the other for many years, but

has recently received a fair amount of attention with the works of David Chalmers [Chalmers, 1995]. Implementation has long been the ultimate goal of Materialists, and Functionalists. It would seem that implementation could provide decisive proof of the truth of a particular theory of consciousness, but despite the world-transforming potential of a conscious machine, the aspiration may prove to be far beyond the limitations of even the most advanced understanding. Finally, systematicity has been debated at much length in the world of Artificial Intelligence [Fodor & Pylyshyn, 1988], [Hadley, 1994a], [Hadley, 1994b], but proposition of systematicity as implicit and necessary to abstract theories does not seem to have been so widely debated. This thesis has advocated for the advancement of the important of systematicity when discussing consciousness.

5.3 Recommendations

- future researchers in this field take serious account of systematicity when creating a blueprint for an NCC or model of consciousness
- discussion of qualia should be included even in theories that seem to have a strong focus on externally verifiable characteristics
- spectrum like organizational structures be used for analysis across theories of consciousness
- development of implementations of thinking machines take into account the wide range of disparate theories and concerns in the field of consciousness and cognitive sciences

5.4 Recommendations for Future Research

In order to manage the scope of this thesis, several more salient properties were considered and put aside. These salient features were eliminated from the thesis due to the lack of time and space to fully explore the implications of these features. Therefore, future work could perhaps involve the attempt to delve further into the possible necessity of these salient features and evaluate how existing theories handle these new features.

This section will suggest several potential future salient properties. The first issue that could benefit from further investigation is the issue of what is, and what is not, a mental system. This was raised by Searle when discussing the possibility of strong

A.I. [Searle, 1980]; however it does seem an important point that for theories which involve a physical and mental component, the parameters of each of these components should be strictly defined.

Another potential salient feature or property is the question of why consciousness exists. This could include questions of how it has evolved and what was the original purpose of consciousness. Clearly the human species has gained great benefit from having consciousness, but as the human mind and society has developed it seems rather that the initial benefits of consciousness may have been lost in all of the modern advances of humanity.

Mental causation is another topic of concern that has been raised over the years. Kim [Kim, 1998], for example has dedicated a fair amount of work to the problem as well as the various solutions suggested by philosophers. For any theory that involves a non-physical factor the problem of mental causation is a not so distant spectre. A treatise about how mental causation on the physical occurs should arguably be a non-trivial portion of these theories.

5.4.1 Related Works

Machine Consciousness is a relatively new term that apparently arises from the tradition of cognitive science. The concept of machine consciousness concerns artificially created consciousness, but machine consciousness differs from traditional Artificial Intelligence in that the focus is strongly on the philosophical questions of consciousness that perhaps are not as strongly addressed by traditional Artificial Intelligence. Additionally, the field of machine consciousness seems less tied to a specific theory of consciousness, and more to the search for an artificial solution that can answer many of the challenges presented by the question of consciousness. Yet, it should be noted that researchers do not presume to guarantee either the existence of consciousness in machines, nor the ability to identify this consciousness should it exist.

Goutam Paul discusses several ways in which Artificial Intelligence can be defined and opines that most of the current work is focused on the search for a “rational

agent" [Paul, 2004, p. 2]. Paul holds that it may not be possible to create consciousness artificially but believes that the systems created in the search can still have value.

David Gamez presents an alternate definition of machine consciousness by breaking the search into the following four basic types:

> MC1. Machines with the external behaviour associated with consciousness.
>
> MC2. Machines with the cognitive characteristics associated with consciousness.
>
> MC3. Machines with an architecture that is claimed to be a cause or correlate of human consciousness.
>
> MC4. Phenomenally conscious machines. [Gamez, 2008, p. 3]

Gamez believes that the first three types of machine consciousness address the easy problem of consciousness as identified by Chalmers [Chalmers, 1995], but maintains that MC4 potentially address Chalmers' Hard Problem of Consciousness. Gamez also spends some time discussing some other serious challenges to machine consciousness such as Searle's Chinese Room Argument [Searle, 1980]. These and other similar issues have been examined by this thesis in the section on the Other Minds Problem [2.5.10.1], and various discussions on The Turing Test.

A different approach to machine consciousness is present by Sidharta Chatterjee, who sets out to discuss the history and challenges of considering consciousness and attempts to define a blueprint or framework for a model of a thinking machine [Chatterjee, 2012].

A notable issue that is discussed in the field of machine consciousness is the ethics of creating an artificial consciousness. The first side of this ethical quandary is the spectre of creating a conscious, thinking machine that is capable of independent thought and action, yet lacks any sense of ethics or morality [Gamez, 2008]. A second issue is the problem that in attempting to create a conscious machine, are we thus creating

slaves without free will [Chatterjee, 2012]. This ethical quandary was briefly mentioned in section 5.1.

Another issue that seems of concern to theorists working in the field of machine consciousness is the ability to recognize or identify consciousness in machines [Gamez, 2008]. Would this consciousness be of a similar or different quality to that of human consciousness? Gamez mentions the possibility that the presence of consciousness in machines may in fact be "indeterminable" [Gamez, 2008, p. 12].

The field of machine consciousness seems to have much to offer in terms of the type of work and direction already examined by this thesis. Future work should consider the area of machine consciousness in more depth.

5.5 Computational Summary

Due to the strong emphasis of philosophical positions in this thesis, it seems an important point to summarize the areas of the thesis that have been devoted to considering computational positions and concerns.

The survey chapter of the thesis contained information on the origins and background of many of the most widely held views on consciousness in the computer science community. Section 2.6.4.1 presented an overview of the Materialist viewpoint, and emphasized the appeal of materialism to falsifiability and scientific method. Arguably, most modern computer scientists would fall under the umbrella grouping of Computationalism, and thus time has been devoted in this thesis to laying out the history of Turing in section 2.6.5.3, an overview of Artificial Intelligence in section 2.6.5.6, with Physical Symbol Systems (2.6.5.6.1) and further discussion of how the Computational Theory of Mind (2.6.5.6.2) developed as one of the most prominent forms of Computationalist theory. A brief overview of Connectionism in section 2.6.5.6.3 provides an alternate, but still computational, viewpoint.

In section 2.6.5.6.4 The Chinese Room Argument is introduced as one of the strongest and most lasting arguments against Artificial Intelligence as a reliable re-creation of human consciousness. Harnad's Total Turing Test in 2.6.5.6.5 demonstrates

an attempt to mitigate the challenge of the Chinese Room Argument through an evolution of the classic Turing Test. Finally, a discussion of Dennett and his Multiple Draft theory in section 2.6.5.6.6 shows a prominent Functionalist viewpoint that seeks to map brain processes used to create consciousness, and is amenable to computational implementation.

The third chapter of this thesis covers questions surrounding the overarching issue of computational implementation in section 3.4.3.1, and how this form of the more general implementation question could be important component of modern theories of consciousness.

Finally, the details of computational implementation and an assessment of the possibilities, advantages and disadvantages are provided for each of: Behaviourism (4.4.3.1.1), Materialism (4.4.3.2.1), Functionalism (4.4.3.3.1), Quantum (4.4.3.4.1), Non-Reductive Physicalism (4.4.3.5.1) and Non-Physicalist theories (4.4.3.6.1).

5.6 Thesis Summary

This thesis presented a survey of many of the most prevalent and widely discussed theories of consciousness, introduced a spectrum for arrangement and classification of these theories, and provided a group of salient features for considering how theories approached some of the major and central issues surrounding the puzzle of consciousness.

References

Akins, K. (2002). A Question of Content. *Daniel Dennett*, 206.

Aristotle (2006). From Metaphysics, Book 7, and On the Soul, Book 2. (W.D. Ross, Trans.). In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Asimov, I. (1942). Runaround. *Astounding Science Fiction, 29,* 94-103.

Atmanspacher, H. (2011, Summer). Quantum Approaches to Consciousness. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/sum2011/entries/qt-consciousness/.

Barker-Plummer, D. (2012, Fall). Turing Machines. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/fall2012/entries/turing-machine/.

Beakley, B., & Ludlow, P. (Eds.). (2006). The *Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Berkeley, G. (1843). *From The Principles of Human Knowledge.* In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Block, N. J. (1980). Troubles with Functionalism (revised). In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Chalmers, D. (1993) Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation. *Philosophical Psychology*, 6(3), 305–319.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, *2*(3), 200-219.

Chalmers, D. J. (1996). *The Conscious Mind*. New York, NY: Oxford University Press.

Chalmers, D. J. (2000). What is a neural correlate of consciousness. *Neural correlates of consciousness: Empirical and conceptual questions*, 17-39.

Chatterjee, S. (2012). Machine Minds: The Blueprint for Artificial Consciousness. *Available at SSRN.*

Churchland, P. M. (1989). Knowing Qualia: A Reply to Jackson. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, 67-76.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Churchland, P. M. (1985). Reduction, Qualia and the Direct Introspection of Brain States. *The Journal of Philosophy*, Vol. 82, No. 1 (Jan., 1985), 8-28.

Copeland, B. J. (2008, Fall). The Church-Turing Thesis. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/fall2008/entries/church-turing/.

Davidson, D. (1970). Mental Events. In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

De La Mettrie, J. O. (1912). From Man a Machine. In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

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

Descartes, R. (1911). From Meditations on First Philosophy II and VI and Replies to Objections II. (E. Haldane and G. R. T. Ross, Trans.). In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Feigl, H. (1958). From "The 'Mental' and the 'Physical'". In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Fodor, J.A. and Pylyshyn, Z.W. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, 28, 3-71.

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

Garson, J. (2010, Winter). Connectionism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/win2010/entries/connectionism/.

Hadley, R. (1994a). Systematicity in Connectionist Language Learning. *Mind and Language*, 9, 247–271.

Hadley, R. (1994b). Systematicity Revisited. *Mind and Language*, 9, 431–444.

Harnad, S. (1993a). Grounding Symbols in the Analog World with Neural Nets. *Think*, 2(1). Retrieved from http://www.archipel.uqam.ca/144/2/index.htm.

Harnad, S. (1993b). The Failures of Computationalism. *Think*, 2(1). Retrieved from http://www.archipel.uqam.ca/144/2/index.htm.

Hookway, C. (2010, Spring). Pragmatism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/spr2010/entries/pragmatism/.

Horst, S. (2011, Spring). The Computational Theory of Mind. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/spr2011/entries/computational-mind/.

Hurd, G. A. (Producer), & Cameron, J. (Director). (1984). *The Terminator* [Motion picture]. USA: Orion Pictures Corporation.

Hyslop, A. (2010, Fall). Other Minds. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/fall2010/entries/other-minds/.

Kim, J. (1998). Mind *in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kubrick, S. (Producer), & Kubrick, S. (Director). (1968). *2001: A Space Odyssey* [Motion picture]. USA: Metro-Goldwyn-Mayer (MGM).

Lanier, J. (1995). You Can't Argue With a Zombie. *Journal of Consciousness Studies*, 2(4), 333-344.

Levin, J. (2010, Summer). Functionalism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/sum2010/entries/functionalism/.

Lewis, C. I. (1991). *Mind and the world order: Outline of a theory of knowledge*. Dover Publications.

Markie, P. (2012, Summer). Rationalism vs. Empiricism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/sum2012/entries/rationalism-empiricism/.

McLaughlin, B., & Bennett, K. (2011, Winter). Supervenience. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/win2011/entries/supervenience/.

Nagel, T. (1974). What Is It Like To Be a Bat? In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113-126.

O'Connor, T., & Wong, H. Y. (2012, Spring). Emergent Properties. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/spr2012/entries/properties-emergent/.

Paul, G. (2004, April). Artificial Intelligence and Consciousness. In *2nd-Human-E-Tech Conference.*

Place, U.T. (1956). Is Consciousness a Brain Process? In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Putnam, H. (1967). The Nature of Mental States. In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Ramsey, W. (2012, Fall). Eliminative Materialism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/fall2012/entries/materialism-eliminative/.

Ramsey, W., Stich, S., & Garon, J. (1990). Connectionism, Eliminativism and the Future of Folk Psychology. *Philosophical Perspectives*, 4, 499-533.

Robinson, H. (2011, Winter). Dualism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/win2011/entries/dualism/.

Robinson, W. (2012, Summer), Epiphenomenalism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/sum2012/entries/epiphenomenalism/.

Ryle, G. (1949). From The Concept of Mind. In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Searle, J. R. (1980). Minds, Brains and Programs. In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

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

Searle, J. R. (1993). The Failures of Computationalism. *Think*, 2(1). Retrieved from http://www.archipel.uqam.ca/144/2/index.htm.

Searle, J. (2007). Biological Naturalism. *The Blackwell Companion to Consciousness*, 325-334.

Searle, J. R. (2005, January 13). Consciousness: what we still don't know. *The New York Review of Books.* Retrieved from http://www.nybooks.com/.

Searle, J. R. (2011, June 9). The Mystery of Consciousness Continues. *The New York Review of Books.* Retrieved from http://www.nybooks.com/.

Silver, J. (Producer), & The Wachowski Brothers (Director). (1999). *The Matrix* [Motion picture]. USA: Warner Bros.

Silverman, A. (2012, Summer). Plato's Middle Period Metaphysics and Epistemology. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/sum2012/entries/plato-metaphysics/.

Smart, J. J. C. (1959). Sensations and Brain Processes. In Beakley, B., & Ludlow, P. (Eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues* (2nd ed.). Cambridge, MA: MIT Press.

Smart, J. J. C. (2011, Fall). The Mind/Brain Identity Theory. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/fall2011/entries/mind-identity/.

Stoljar, D. (2009, Fall). Physicalism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* *http*://plato.stanford.edu/archives/fall2009/entries/physicalism/.

Stubenberg, L. (2010, Spring). Neutral Monism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/spr2010/entries/neutral-monism/.

Tanney, J. (2009, Winter). Gilbert Ryle. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/win2009/entries/ryle/.

Thagard, P., & Aubie, B. (2008). Emotional consciousness: A neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. *Consciousness and cognition*, 17(3), 811.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, New Series, 59(236), 433-460.

Tye, M. (2009, Summer). Qualia. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy. Retrieved from* http://plato.stanford.edu/archives/sum2009/entries/qualia/.