

# **A Critical Study of the Philosophical Critique of Artificial Intelligence**

*Dissertation submitted to Jawaharlal Nehru University  
in partial fulfilment of the requirements  
for the award of the degree of*

**MASTER OF PHILOSOPHY**

**BULLO KANO**



**CENTRE FOR PHILOSOPHY  
SCHOOL OF SOCIAL SCIENCES  
JAWAHARLAL NEHRU UNIVERSITY**

**NEW DELHI-110067**

**INDIA**

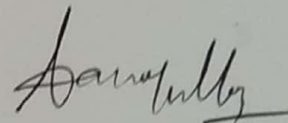
**2018**

## DECLARATION

I, **Bullo Kano**, do hereby declare that the dissertation entitled *A Critical Study of the Philosophical Critique of Artificial Intelligence* in partial fulfilment of the requirements for the award of the degree of **Master of Philosophy** of Jawaharlal Nehru University is my original research work. The dissertation has not been submitted in part or in full to any other university or elsewhere to obtain any other degree.

Date: 23/07/2018

Place: J.N.U., Delhi



**Bullo Kano**

Centre for Philosophy  
Jawaharlal Nehru  
University  
New Delhi - 110067



जवाहरलाल नेहरु विश्वविद्यालय JAWAHARLAL NEHRU UNIVERSITY

सामाजिक विज्ञान संस्थान SCHOOL OF SOCIAL SCIENCES

दर्शन शास्त्र केंद्र CENTRE FOR PHILOSOPHY

नई दिल्ली - ११००६७ NEW DELHI - 110067

Date: 23/07/2018

### CERTIFICATE

This is to certify that the dissertation entitled *A Critical Study of the Philosophical Critique of Artificial Intelligence* submitted by **Bullo Kano**, in partial fulfillment of the requirements for the award of the degree of **Master of Philosophy** of Jawaharlal Nehru University, New Delhi is his original work. It is further certified that the dissertation has not been submitted in part or in full to any other university or elsewhere to obtain any other degree.

The dissertation may be placed before the Examiners for evaluation.

*Smita Sirker*  
23/07/2018

**Dr. Smita Sirker**

(Supervisor)  
Supervisor  
Centre for Philosophy  
School of Social Sciences  
Jawaharlal Nehru University  
New Delhi - 110067, India

*Rendu Kuei*

(Chairperson)

23/7/2018  
Chairperson  
Centre for Philosophy  
School of Social Sciences  
Jawaharlal Nehru University  
New Delhi - 110067, India.

# Contents

	<b>Page No.</b>
<b>Acknowledgement</b>	<b>i</b>
<b>Introduction</b>	<b>1</b>
<b>Chapter 1: Historical Antecedents of AI</b>	<b>5</b>
1. Early History of AI	7
2. Middle Ages of AI Development	14
3. The Divide	19
4. Dark Ages of AI Research	21
5. AI Renaissance	21
6. Knowledge Representation	26
7. The Present and Future	28
<b>Chapter 2: Philosophical Discourse on the Nature of Mind</b>	<b>31</b>
1. Ancient Notion of Mind/Soul in Western Philosophy	31
2. Dualist Approaches to Mind	34
3. Varieties of Materialism (From Identity Theory to Functionalism)	40
4. Computationalist Theory of Mind	53
<b>Chapter 3: Philosophical Debates on AI</b>	<b>60</b>
1. Searle's Thesis and Critique of AI	61
2. Criticism of Searle's Thesis	66
3. Penrose's Thesis and Critique of AI	74
4. Warwick's Thesis	86
<b>Chapter 4: Analysis of the Relevant Debates and Concepts in AI</b>	<b>94</b>

1. Understanding Computers and Programs	95
2. What is the notion of “Understanding” in Humans and Machines	104
3. Understanding Intelligence as a Notion	108
4. Intelligence in AI	121
<b>Conclusion</b>	<b>126</b>

## **Bibliography**

## **Acknowledgement**

This dissertation has been completed under the supervision of Dr. Smita Sirker, Associate Professor, Centre for Philosophy, School of Social Sciences, JNU, New Delhi. I would like to express my gratitude to Dr. Sirker for her invaluable advices, suggestions, corrections, bountiful patience, relentless and much needed encouragement, and all that she has done as my supervisor. Working under her supervision has been a unique experience for me.

I would also like to acknowledge all that I have learned under various professors of Centre for Philosophy. I have truly leaned a lot during the two years of my MPhil course and for that I am grateful to the entire faculty members. Special thanks to the staff of the centre, especially Mr. Yogender and Mr. Praveen, for all the service and help they have provided to us students over the years. I would also like to give credit to the JNU library for providing books required for my dissertation. In general, I would like to express my gratitude to Jawaharlal Nehru University (JNU), New Delhi.

This acknowledgement would not be complete without thanking my friends and classmates for all the help, support and assistance they have provided. The discussions with each one of them have been enlightening. Special thanks to Ronald Lallienthang for his help in finding books and articles online.

Lastly, my heart-felt gratitude to my family for the care and support they have provided me throughout my life. I would always be in debt to them for all that they have done for me.

**Bullo Kano**

## Introduction

Can a machine think? This question was first proposed by Alan M. Turing in 1950<sup>1</sup>. Today, a variant of this same question may be proposed – is Artificial Intelligence (AI) possible? One may answer “yes” to these questions or one may say “no”. Based on the response to these questions, there may be two camps – one supporting the possibility of AI and the other opposing any such possibility. Either of the camps, whether those who supports the possibility of AI or those who oppose it, have to provide some grounds on the basis of which they assert or deny their theses. AI discussions, in so far as there are claims regarding intelligence, must also talk of systems, already in nature, that are intelligent. An obvious candidate for such a system is the human being. A discussion of such a system, or for any system, requires a conceptual framework that provides for the discussion itself. The various concepts and theories in philosophy do provide such a framework as already much have been discussed on human nature and more specifically the human mind. Thus, it becomes important to have a philosophical study or analysis of such discussions. In fact there have been quite a lot of philosophical debates centred on the earlier questions – “Can machine think?” or “Is AI possible?” Some thinkers such as John Searle and Roger Penrose have argued against it, while others such as Kevin Warwick have argued in support of it. The philosophical debates then provide for an interesting area of study to understand the conceptual framework underlying these debates. Also claims in AI, in turn, have provided a fresh perspective on the age-old mind-body problem in philosophy. The problem in a general sense refers to the kind of relation between mind and body, the mental and the physical. But more specifically it refers to the causal relation between the two seemingly distinct kinds of substances/states/properties. While philosophers have often argued that one of them may be reduced to the other but such explanations may come at an expense of eliminating one for the other. While explanations that reduce the physical to the mental have gained less popularity, the explanations reducing the mental to the physical have gained acceptance among many. This is largely owing to the acceptance of a physicalist picture of the world. AI may perhaps be termed as the product of the progress in science and more specifically mathematics and computer science. With the development in AI research achieving more and more in terms of what have

---

<sup>1</sup> Turing, “Computing Machinery and Intelligence”, *Mind* 49, 1950, pp. 433-60.

been considered, till recently, only unique to humans, it presents for a subject matter for all traditional forms of studies on human nature, philosophy being one of them.

The goal of this thesis is then to pick out some key concepts and arguments from the philosophical debates on AI and critically analyse them. An important issue that this thesis aims to deal with is the question – What is intelligence? As such we will also analyse various concepts associated with the concept of intelligence, such as, symbolic systems, representations, semantics, “understanding” and so on, all in terms of the relevant philosophical debates on AI. A broader goal is to understand the field of AI and draw out its relation and relevance to philosophy and vice-versa. Having outlined the aims, the following is a brief outline of the different chapters.

The first two chapters set the background for the philosophical debate on AI. As such Chapter 1 is titled “Historical Antecedents of AI” and Chapter 2 is titled “Philosophical Discourse on the Nature of Mind”. Chapter 1 deals with the beginnings of AI. It traces out the historical progress of AI and also discusses some of the key concepts in AI. The chapter discusses the mathematical models of neurons as described by McCulloch and Pitts and some key concepts such as Turing machine, algorithm, computability, physical symbol system and knowledge representation. It also describes the Turing’s test as given by Alan Turing and its implications on the study of artificial intelligence. Further discussions deal with various developments in the field of AI such as expert systems, artificial life (A-life), artificial neural network and some other technical progress. Chapter 2 deals with various philosophical discourses on mind or soul and its relation to body. This chapter starts with a discussion on the different ancient notions of mind/soul in Western philosophy. It goes on to discuss the theories of dualism, which includes Platonic dualism and Cartesian dualism. There are discussions of Aristotle’s notion of soul and parallelisms of Spinoza and Leibniz. Further, theories like various forms of identity theory, various forms of behaviourism, functionalism and computationalism are also discussed. Chapter 2 ends with a brief comment on the inter-relation of AI and philosophy of mind and how computationalism serves as the philosophical basis for AI. Together, chapter 1 and 2 points out, among many other things, antecedents of some modern concepts of both AI and philosophy as found in the history of philosophy. These chapters will give us some idea about what forms the framework within which the dissertation is to be limited.



Chapter 3, “Philosophical Debates on AI” takes up some of the important philosophical debates on AI. It discusses specific arguments provided by John Searle, Margaret Boden, Roger Penrose and Kevin Warwick. The philosophical debates on AI, thus, revolve around the arguments given by them. Searle’s famous Chinese Room argument and its implications are discussed followed by Boden’s criticism of Searle’s arguments. One of the implications of Searle’s argument is that “computational theories ... cannot possibly help to understand mental processes”<sup>2</sup>. Boden in “Escaping from the Chinese Room”<sup>3</sup> replies to this implication of Searle. Her reply provides for arguments against the Chinese Room argument. Penrose in asserting his own position argues against those who support the possibility of AI by attacking computationalism. However, he too criticizes Searle’s position. Some important implications are drawn from Penrose’s discussion of “understanding”, awareness and arguments against AI. Lastly, Warwick’s arguments in support of AI are discussed. He criticizes both Searle’s and Penrose’s respective viewpoints. He appeals and argues for a more inclusive attitude towards all kinds of intelligence. This chapter aims to develop a sense of the philosophical discussion centred around AI. The chapter mostly deals with the arguments and criticism confined within Searle, Boden, Penrose and Warwick. Searle’s position may be termed as weak AI. Weak AI is the position that computers can merely simulate human mental features. Boden’s position may be termed as strong AI as she not only argues for a computer which can instantiate features of mind such as understanding, but also argues that AI provides for important insights into the nature of human cognitive processes. Strong AI is the position that machines can be as intelligent as humans and may even surpass humans someday in the future. Penrose argues against the computationalists’ claim that mind is a function and is thereby computable. Penrose is against the views that claim AI to be capable of understanding or claim AI to be capable of simulating “understanding”. For him, “understanding” is non-computational and there by the mind cannot properly be simulated as “understanding” is an important feature of the mental. Warwick talks of machines as if it can be seen as another kind of species, and just as other species (humans or animals) have capabilities that are unique to them, machines are then to be considered as capable in their own unique way too. Warwick feels that the subjective differences between various

---

<sup>2</sup> Boden, Margaret A., “Escaping from the Chinese Room”, *The Philosophy of Artificial Intelligence*, ed. Margaret A. Boden, Walton Street, Oxford: Oxford University Press, 1990, p. 89.

<sup>3</sup> *Ibid.*, p. 89-103.

species should be taken into account when considering them as intelligent or not. The main opponents against whom Warwick provides his arguments are Searle and Penrose.

In chapter 4 “Analysis of the Relevant Debates and Concepts in AI”, I have tried to put forward my own arguments regarding the discussions and debates that were discussed in the preceding chapter. It also contains some of my own criticisms of those arguments. Further, analysis of the definition of a computer/program and analysis of concepts such as representation, “understanding” and intelligence have been taken up and based on the understanding of these concepts some of my arguments are developed. The concept of representation plays a central role in the discussion. “Understanding” is analysed in terms of “understanding” as having semantics and “understanding” without any semantic content. Based on this, the case is made for whether computers can be said to have understanding as Warwick asserts. A main argument here is that intelligence is simple just like Moore’s notion of good as simple. Arguments are provided in favour of it by establishing an analogy between the concept of intelligence and the concept of good. This chapter ends with a discussion on the question of – what is intelligence. The answer to this also, it has been argued, can be located in the above analogy. The analogy though borrowed from ethics, is not to be taken as making any claims about ethics. However, it does help us to bring out a comparison for the discussion on intelligence with particular notions in ethics just as Searle explains mind-brain relation by drawing analogy with notions in physics.

In the conclusion, I have provided a brief summary of the study conducted in this dissertation. The contents of each chapter and their purpose are briefly laid out. Furthermore, the limitations of the study carried out in this dissertation are briefly discussed followed by suggestions and prospects for future research.

# Chapter 1

## Historical Antecedents of AI

This thesis is, in a general sense, a discourse on philosophical critiques of Artificial Intelligence (henceforth, AI), and strives to be a part of the discourse, of philosophical nature specifically, on AI. Embarking on such a discourse requires that one is able to grasp the concepts, employed and developed within the discourse. As such the initial two chapters are dedicated mostly to the task of gleaning out relevant concepts. This may require us to start with the discussion of concepts and theories, perhaps, from the philosophical domain owing to the philosophical nature of the discussions, but a more immediate requirement, is perhaps, discussing the concept of the main subject matter of the thesis itself, so that one doesn't have to wonder as to what it is exactly about or get dejected when the discussion doesn't address the reader's assumptions. So the first question that any philosophical critique of AI asks is – What is Artificial Intelligence? This question doesn't have to be taken in the sense that the question may further be divided into questions of simpler concepts – what is intelligence; what is artificial; and then what is Artificial Intelligence? Questions of this kind will be taken up later when the concepts of intelligence, artificial and so on are discussed. For now the question, “what is Artificial Intelligence?” asks a broader question about Artificial Intelligence as a branch of study or a branch of science. What is one studying when one studies artificial intelligence? The answer or answers to this question has been satisfactorily dealt with by Margaret Boden in the introduction section of *The Philosophy of Artificial Intelligence*<sup>4</sup>. Artificial Intelligence is defined in various ways. For instance, according to some, it is the “study of how to build and/or program computers to enable them to do the sorts of things that minds can do”<sup>5</sup>. Here the things<sup>6</sup> that minds can do refer to those things which require intelligence and thus computers are taken as things or systems which can perform at least some of those intelligent things depending on their program. Artificial intelligence then deals only with such programming and its goal is to

---

<sup>4</sup> Boden, Margaret A., “Introduction”, *The Philosophy of Artificial Intelligence*, ed. Margaret A. Boden, Walton Street, Oxford: Oxford University Press, 1990, pp. 1-20.

<sup>5</sup> Boden, 1990, p. 1.

<sup>6</sup> Boden classifies these things (that mind can do) into those that “are commonly regarded as requiring intelligence” and those that “can be done by all normal adults... (and sometimes by non-human animals too). The second kind “typically involve no conscious control: seeing things in sunlight... finding a path through cluttered terrain,... using one's common sense”. As such AI may seem to also include non-human behaviours. See Boden, 1990, p 1.

replicate intelligent behaviours in computers or machines. A more controversial (yet fitting, in my opinion) definition provided by Boden is, AI is seen as “the science of intelligence in general”. Its goal then is to arrive at systematic theories that provide for explanations for the various mental or psychological capacities<sup>7</sup>. It then encompasses the entire range of possible minds<sup>8</sup>. Based on the explanations AI provides, perhaps, one can then be able to say whether intelligence is possible only in some particular systems (humans, animals) or it is possible for intelligence to be replicated in systems that are radically different in nature in terms of their composition from those particular systems. Such a field then requires interdisciplinary approach from various fields such as psychology, neuroscience and cognitive science, of which AI is an important domain. The philosophical significance in the discourse on AI lies in the concepts pre-assumed by those in the field of AI, understood as either of the two definitions. As Dreyfus puts it – “what underlying assumptions lead workers in AI.. Can these assumptions be justified?”<sup>9</sup> But what are these assumptions? To situate the discourse chapter 1 will discuss the historical antecedents and chapter 2 will discuss the theories on the nature of mind.

Warwick in his book *Artificial Intelligence: The Basics* classifies and provides for the history of AI. He observes, “There are strong links between the development of computers and emergence of AI”<sup>10</sup>. The beginning of AI, in the present form, perhaps could be traced back to the 1900s. In 1940s and 1950s, the focus of AI development was on getting computers to do things that if a human did them would be regarded as intelligent. Subsequently the discussion centered on how close to a human brain could a computer be. These periods in the AI development is referred to as Classical AI. Computers and thereby AI were limited in its potential. The 1980s and 90s saw efforts directed toward building artificial brain to bring about artificial intelligence. The focus was to develop AI in its own way and not just restricted to merely copying human intelligence. AI could not only mimic human intelligence but also had

---

<sup>7</sup> Boden, 1990, p. 1.

<sup>8</sup> The range of mind may start from the simplest of minds as may be found in simple organisms to the complex ones found in human beings. It also leaves open the possibility of mind more complex than human. Boden says AI “ must encompass not only the psychology of terrestrial creatures.... It must tell us whether intelligence can be embodied only in systems whose basic architecture is brainlike... or whether it can be implemented in some other manner.” See, Boden, 1990, p. 1.

<sup>9</sup> Dreyfus, Hubert L. “Artificial Intelligence.” *The Annals of the American Academy of Political and Social Science*, vol. 412, 1974, pp. 21–33, p. 22.

<sup>10</sup> Warwick, Kevin, *Artificial Intelligence: The Basics*, Abingdon, Oxon: Routledge, 2012, p. 2.

potential to out-perform it. Today AIs are being given<sup>11</sup> their own bodies to perceive the world in their own way and to move and modify it as they see fit. With this AI are given the ability to learn, adapt and carry out their tasks. As mentioned above the emergence of AI is often linked with the development of computers. AI is thought of in terms of computers (mostly digital computers, but may also be mechanical). However, Warwick points out: "... seeds of AI were sown long before the development of modern computers"<sup>12</sup>. This section will also deal with the present and future prospects of AI development as pointed by Warwick.

## 1. Early History of AI

The term Artificial Intelligence was first coined by John McCarthy in 1956. But artificial beings (claims Warwick<sup>13</sup>) can be traced back early in history to stories of Prague Golem and other Greek myths. There have also been machines designed for calculations, starting from Gottfried Leibniz to Charles Babbage<sup>14</sup>. However, it may be claimed that AI, or at least a part of it, was foreseen in the 1840s by Lady Ada Lovelace<sup>15</sup>. She said that a machine might compose music of great complexity or even express facts of this world which would contribute greatly to the sciences. The machine she was referring to was the Analytical Engine designed by Charles Babbage, who was her friend, in 1834<sup>16</sup>. What she had realized was the potential generality of the engine. Of course she had no clue of today's notion of neural networks or any of the AI forms. She was focused on symbols and logic. For her, the machine had the ability to process symbols representing basically every subject in the world. For her, AI was possible but how to go about it was still obscure. In case of the machine composing music, she said it is possible but she never mentioned how the machine would do so. This obscurity was later overcome by Alan Turing.

---

<sup>11</sup> AI computers are put into bodies of robots.

<sup>12</sup> Warwick, 2012, p. 2.

<sup>13</sup> Warwick, 2012, p. 2.

<sup>14</sup> Reus, Bernhard, *Limits of Computation: From a Programming Perspective*, Switzerland: Springer Nature, 2016, p. 14.

<sup>15</sup> Lovelace, A. A. (1843), "Notes by the Translator." Reprinted in *Science and Reform: Selected Works of Charles Babbage*, ed. R.A. Hyman, 1989, pp. 267-311; for a discussion on it, see Boden, Margaret A., *AI: Its Nature and Future*, Great Clarendon Street, Oxford: Oxford University Press, 2016, p. 7.

<sup>16</sup> Boden, *AI: Its Nature and Future*, 2016, p. 8.

But before that, according to Warwick, “the strongest immediate roots probably date back to the work of McCulloch and Pitts”<sup>17</sup>. In 1943, they described mathematical models of neurons in the brain, called *perceptrons*<sup>18</sup>, based on a detailed analysis of the biological originals. They indicated not only how neurons operate in a switching binary fashion but also how neurons could learn and hence change their action with respect to time. The fundamental assumption in neurophysiology is that the nervous system is a net of neurons<sup>19</sup>. Each neuron has a soma and an axon. A soma is the body of a neuron. Axon is a special cellular extension that arises from the cell body and travels for a distance. The point of connectivity between the axon of one neuron and the soma of another is called synapse or a synaptic connection (see Fig. 1).

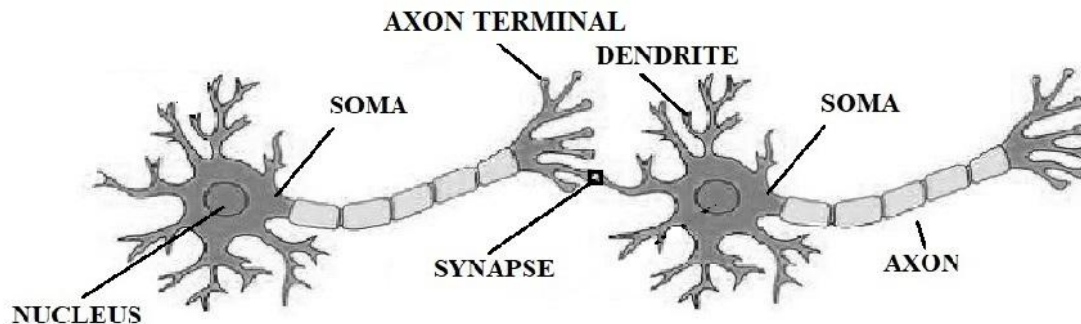


Fig. 1

McCulloch and Pitts explain, “... a neurone has some threshold, which excitation must exceed to initiate an impulse. This, except for the fact and the time of its occurrence, is determined by the neuron, not by the excitation. From the point of excitation the impulse is propagated to all parts of the neurone. The velocity along the axon varies directly with its diameter.... Excitation across the synapses occurs predominantly from axonal termination to somata”<sup>20</sup>. The axon of one neuron is not physically attached to the soma of another neuron at the synaptic junction. The

---

<sup>17</sup> Warwick, 2012, p. 2.

<sup>18</sup> The term “perceptron” was coined at a later date for such a kind of neural models. It was developed by Frank Rosenblatt (1928 – 1971).

<sup>19</sup> McCulloch and Pitts, ‘A Logical Calculus of The Ideas Immanent in Nervous Activity’, reprinted in *The Philosophy of Artificial Intelligence*, ed. Margaret A. Boden, Walton Street, Oxford: Oxford University Press, 1990, p. 22.

<sup>20</sup> *Ibid.*, p. 22

further explain, “Between the arrival of impulses upon a neurone and its own propagated impulse there is a synaptic delay... During the first part of the nervous impulse the neurone is absolutely refractory to any stimulation. Thereafter, its excitability returns rapidly, in some cases reaching a value above normal from which it sinks again to a sub-normal value, whence it returns slowly to normal... Such specificity as is possessed by nervous impulses depends solely upon their time and place and not on any other specificity of nervous energies”<sup>21</sup>. Only inhibition as opposed to excitation is said to have an effect on these operations. Inhibition is explained by them as, “the termination or prevention of the activity of one group of neurones by concurrent or antecedent activity of a second group”<sup>22</sup>. McCulloch and Pitts, conceived the idea of “the response of any neurone as factually equivalent to a proposition which proposed its adequate stimulus”<sup>23</sup>. So if a neuron fires, it could be taken as the proposition that asserts the stimulus which is adequate to elicit that behavior from the neuron. For instance, if Sheila says, “It will rain”, then this may be taken as equivalent to the proposition “Sheila believes that the statement “It will rain” is true”. Similarly, if a neuron fires, then this may be taken as equivalent to the proposition that “there is an adequate stimulus for the neuron to fire” is true. Just as the utterance of the statement by Sheila can be taken as representing her belief which is the cause or stimulus of her utterance, similarly, the firing of a neuron may be taken as representing the stimulus that evoked that behavior of the neuron. The fact that a neuron either fires or does not fire was termed by McCulloch and Pitts as the ‘all-or-none’ law. They explained, “The ‘all-or-none’ law of nervous activity is sufficient to insure that the activity of any neuron may be represented as a proposition. Physiological relations existing among nervous activities correspond... to relations among the propositions.”<sup>24</sup>

It is the case that our neurons and thereby neural networks do undergo changes which may perhaps not be easily translatable to propositions or formal nets. McCulloch and Pitts proposed: “To each reaction of any neurone there is a corresponding assertion of a simple proposition.”<sup>25</sup> As such, corresponding rules of logic such as conjunction, disjunction, negation and so on equally applies according to the configuration of synapses and the threshold of a

---

<sup>21</sup> Ibid., p. 23.

<sup>22</sup> Ibid., p. 23

<sup>23</sup> Ibid., p. 23.

<sup>24</sup> Ibid., p. 23-24.

<sup>25</sup> Ibid., p. 24.

particular neuron. However, McCulloch and Pitts saw two difficulties in considering neuronal activity as propositions, as propositions remain unaltered whereas neuronal behavior may be altered. The first difficulty is with temporary changes, facilitation and extinction. An antecedent activity may temporarily change the responsiveness of the net to subsequent stimuli. The changes will be in terms of facilitation or extinction, facilitation when the changes are in favor of excitation of impulses and extinction when the net is prohibited to be excited or an already excited net is inhibited so that no more impulses are generated. The second difficulty is that of learning. In this case the changes are permanent. The changes are due to activities concurrent with some previous activities of the net. For instance, some previous activity might have affected changes in the net so that a stimulus which was adequate earlier is no longer adequate. The duo suggests that these two difficulties can be overcome as fictitious non-altered nets, in terms of its connections and thresholds, could be conceived. The alteration, of both kinds, then doesn't affect the conclusion of the formal nets which are formal equivalence of neural nets. Thus, McCulloch and Pitts makes a few physical assumptions for their mathematical model of neurons –

1. “The activity of the neuron is an ‘all-or-none’ process.
2. A certain fixed number of synapses must be excited... in order to excite a neurone at any time, and this number is independent of any previous activity and position of the neurone.
3. The only significant delay within the nervous system is synaptic delay.
4. The activity of any inhibitory synapse absolutely prevents excitation of the neurone at that time.
5. The structure of the net does not change with time.”<sup>26</sup>

Based on these assumptions the duo constructs a formal model of neural network.

Turing in 1950 wrote a paper on the question ‘Can a machine think?’ He came up with a test, known as Turing Test to answer such questions. Turing starts with the question “can machines think?”<sup>27</sup>. He replaces this question with another. For this, he proposes a game. The game is played between A (a man), B (a woman) and C (a third person who is the interrogator).

---

<sup>26</sup> Ibid., p. 25.

<sup>27</sup> Turing, Alan M., “Computing Machinery and Intelligence”, reprinted in *The Philosophy of Artificial Intelligence*, ed. Margaret A. Boden, Walton Street, Oxford: Oxford University Press, 1990, pp. 42-64, p. 42.



They are each placed in a separate room with only a teleprinter as the mode of communication between the interrogator and the other two participants. The interrogator may ask questions to the other two and A and B will respond. The object of the game then is for the interrogator to determine which one of the two is a man and which is the woman based on their responses. At the end of the game C says either “A is man and B is woman” or “A is woman and B is man”. Also it is supposed that A’s objective is to make C give a wrong answer to this test and B’s objective is to help the interrogator. The best strategy for A is to give the same answers and respond to C’s questions in the same way as B would. A has to imitate B to make C believe that he is a woman. Thus, the game is called “The Imitation Game”. The question then is asked “What will happen when a machine takes the part of the man in this game?” Will the interrogator be able to tell if A is a computer and B a woman (or man or just a human being)? These questions then replace the original question, “Can machines think?” The implication of this game is not that a machine can think but that if the machine can successfully make one believe (in the above case, the interrogator) that it is a human being, then the machine can be understood to have a capacity as good as thinking in human. This implication is partly due to the assumption that thinking is the essence of human and partly due to the consideration that human mind may be accounted for in mechanical terms<sup>28</sup>. To re-phrase, the implication of this game is not that a machine thinks like humans or is thinking at all, but that it is as intelligent as a human. Thus, a machine to be considered intelligent has to take part in this game and succeed to make the interrogator believe it’s a human. This is then called “Turing’s Test”.

Whether a machine can pass this test or not is still an issue as none of the AI developed yet has actually passed it. What is then Turing’s thesis? First, a computer is anything that computes. A computer can be a human computer (which basically means a human who is computing), or an alien computer (which basically means an alien that is computing, that is, if there are aliens that have capabilities same as or higher than the human mind). What is of importance in the concept of computer is that whatever it is, it computes. Turing’s thesis is then the thesis that asserts or claims that a machine can compute what a computer can compute. The machine is called Turing machine. The computer here is basically a human who computes, a

---

<sup>28</sup> Antecedents of this may be found in the section dealing with history of the concept of mind, such as the discourse of Aristotle and Descartes, but especially in Locke’s and Hume’s account.

human computer<sup>29</sup>. The human computer sits in his desk and by using a pen and paper performs his task of computing numbers. Turing claims that what this human computer computes can also be computed by a machine<sup>30</sup>. Thus, computing which is a capacity (some may consider it as intelligent) of human beings in general, and human computers in particular, can be performed by a machine. Of course, like before, a machine does not need to compute in the same way as a computer<sup>31</sup> (human) to succeed at its task, just like a machine does not need to think like a human to be intelligent. Turing's thesis then is the claim that a machine, that is a Turing machine, can compute all that a human computer can compute. What is this Turing machine then? Turing proposes such a machine and its operations in his 1936 article "On Computable Numbers, with an Application to the Entscheidungsproblem". He says, "It is possible to invent a single machine which can be used to compute any computable sequence"<sup>32</sup>. This machine is not to be thought as a physical machine. Rather, what Turing is proposing in his article, is an abstract, theoretical or mathematical model of a possible machine. An illustration of this machine is given by Tim Crane in his book *The Mechanical Mind*. A Turing machine consists of: a tape divided into squares capable of bearing a symbol and a device that can read (scan) and write symbols on the tape. The machine also has certain 'internal states'<sup>33</sup>, what Turing called configuration or condition of the machine and therefore machine configurations or *m*-configurations<sup>34</sup>. The scanned symbol is the only one of which the machine is directly aware of. However, by altering its *m*-configuration the machine can remember some of the symbols which it has scanned previously. The machine can only read one square at a time. Suppose there are two symbols '0' and '1'. There are four possible behaviours. First, it can move the tape one square at a time from left to right or from right to left. Second, it can read a symbol on the tape. Third, it can write a symbol on the tape, either by writing onto a blank square or by overwriting another symbol. Lastly, it can change its 'internal state'. The possible operations of the machine can be represented by the machine's machine table. The machine table is a set of instructions. For

---

<sup>29</sup> Turing makes this comparison between a man in the process of computing and a machine which is capable of a finite number of conditions called *m-configurations*.

<sup>30</sup> Turing, Alan M., "On Computable Numbers, With An Application To The Entscheidungsproblem", *Proceedings of the London Mathematical Society*, Series 2, Vol. 42 (1936-37) pp. 230-65, pp. 249-52.

<sup>31</sup> Turing in fact uses the term initially to refer to a human computer.

<sup>32</sup> Turing (1936-37), p. 241.

<sup>33</sup> Crane, Tim, *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representations*, first published in 1995, second edition. London: Routledge, 2003, p. 93.

<sup>34</sup> Turing, 1936-37, p. 231.

instance it may say, if in state  $X$  and reads a symbol  $S$ , then print either  $O$  or  $1$  or leave blank, move the tape either to left or right and then change to state  $Y$ . The machine table is the machine's program. It tells the machine what to do.

How are all these comparable to what humans do? Turing does compare such a machine to a human computer. Suppose a human computer has a task assigned to it. The task is to find whether a given number is computable. "Computable" numbers are described by Turing as the real numbers whose expressions as a decimal are calculable by finite means<sup>35</sup>. So the computer's task is to find out the decimals of a given real number by finite means. The computer is given a pencil and a one-dimension tape consisting of squares. The behaviour of the computer at any moment of time is determined by the symbols which he is observing and his "state of mind" at that moment<sup>36</sup>. We may also suppose that the operations performed by the computer are simple operations which are so elementary that it may not be divided further. Thus the computer can observe only one square at a time. Suppose that it takes quite a many steps and long hours to reach the end of the procedure owing to the restriction for using only simple operations. The computer then takes a break when necessary. But before leaving for his break, each time he leaves a note on what his state of mind is and which state he needs to be in for the next step on returning to his task. He does so because he may forget and does not want to start from the beginning. Now suppose further that this particular computer is a very stressful person and can compute only one step at a time before he needs a break. So each time the computer takes a break he leaves a note behind. The note is the counterpart of 'state of mind'. So if one were to see all the notes made by the computer before each break then one could take it along with the symbols on the tape as instructions to follow in each step till the whole procedure is completed. This is so because as mentioned earlier the state of mind and the symbols determine the behaviour. Now we can say that the state of mind at a given stage is completely determined by state of mind before the last step was made. The relation between the two states of mind is then expressible in functional calculus. Therefore, we can assume such axioms which express the rules governing the behaviour of the computer. The computer may be replaced by a machine whose  $m$ -configuration corresponds to the state of mind of the computer. The machine scans the symbols on the tape one square at a time just as the computer observed the symbols one square at

---

<sup>35</sup> Ibid., p. 230.

<sup>36</sup> Ibid., p. 250.

a time. Then following the axioms or rules the machine can be made to compute the required number. The axioms or rules may be compared to a machine table.

Shortly after Turing's paper on "Can a machine think?", Marvin Minsky and Dean Edmonds built the first AI computer based on a network of the neuron models of McCulloch and Pitts<sup>37</sup>. Minsky and Shannon, who considered the possibility of computer playing chess, started the first workshop in the field of AI subsequently leading to the classical foundations of the subject.

## 2. Middle Ages of AI Development

In 1960s Newell and Simon came up with the General Problem Solver program which was a multipurpose program aimed at simulating some human problem-solving methods<sup>38</sup>. The program was supposed to capture the feature of general intelligence as exhibited in human beings. Humans use intelligence across diverse sets of domains. As such it is thought that there is a general intelligence factor that underlies specific intelligence factor in a particular domain. It was assumed that such a computer program would be able to interface intelligently across domains just like humans do. Due to technical difficulties and limitations the project was abandoned. However, insights into AI research could be gained from their paper titled "Computer Science as Empirical Enquiry: Symbols and Search"<sup>39</sup>. AI research for them probably meant the study of how to build and/or program computers to enable them to do the sorts of things that minds can do, in this case to enable computers to perform intelligent actions. For them, neither machines nor programs are black boxes. They are artefacts that have been designed; both hardware and software. They can be dismantled, taken apart part by part, opened up, and each part or the inside could be examined. One can relate their structure to their behaviour and draw many lessons from a single experiment<sup>40</sup>. Newell and Simon claimed that symbols lie at the root of intelligent action, which is, the primary topic of AI. One fundamental contribution of computer science has been to explain at a basic level what symbols are<sup>41</sup>.

---

<sup>37</sup> Warwick, 2012, p. 3.

<sup>38</sup> Ibid., p. 3.

<sup>39</sup> Newell and Simon, "Computer Science as Empirical Enquiry: Symbols and search", reprinted in *The Philosophy of Artificial Intelligence*, ed. Margaret A. Boden, Walton Street, Oxford: Oxford University Press, 1990, pp. 105-32.

<sup>40</sup> Ibid., p. 106.

<sup>41</sup> Ibid., p. 107.

Intelligence of a system is measured by its ability to achieve stated ends in the face of variations, difficulties, and complexities posed by the task environment.

Newell and Simon say that intelligence is a composite, for no single elementary thing accounts for intelligence in all its manifestations. There is no intelligence principle, just as there is no vital principle<sup>42</sup> that conveys by its very nature the essence of life<sup>43</sup>. The duo instead argue that one requirement for intelligence is the ability to store and manipulate symbols.

Newell and Simon draws from laws of qualitative structure found in science<sup>44</sup> and forward their own qualitative law called the “Physical-Symbol System Hypothesis”. The hypothesis states that “a physical-symbol system has the necessary and sufficient means for general intelligent action”<sup>45</sup>. What is this physical-symbol system then? The term physical here denotes – firstly, that such a system obeys the laws of physics and secondly, that such a system is not restricted to a human symbol system but may be any other physical system. A physical-symbol system then consists of two things – a set of ‘symbols’ and a collection of processes. Symbols are physical patterns that can occur as components of ‘expressions’ (or symbol structure). The tokens of symbols within a symbol structure are then related in some physical way. A collection of processes are also contained in the system that operates on expressions in the system to produce other expressions within the system through creation, modification, reproduction and destruction. The notions central to this structure are that of ‘designation’ and ‘interpretation’. An expression designates an object if the system can either affect the object or behave in ways depending on the object<sup>46</sup>. The system interprets an expression if the expression designates a process that the system can carry out. Additionally the system needs to satisfy some conditions –

- i. the symbol can be used to designate any expression;
- ii. the expression can designate every process of which the machine is capable;

---

<sup>42</sup> Ibid., p. 107.

<sup>43</sup> For Descartes animals and lesser animate entities, that is, other than humans, could be accounted for or explained through mechanical explanations.

<sup>44</sup> Newell and Simon, Computer Science as Empirical Enquiry: Symbols and search”, reprinted in *The Philosophy of Artificial Intelligence*, ed. Margaret A. Boden, Walton Street, Oxford: Oxford University Press, 1990, pp. 107-109.

<sup>45</sup> Ibid., p. 111.

<sup>46</sup> Ibid., p. 110.

- iii. there are processes for creating any expression and modifying any expression in arbitrary ways;
- iv. expressions once created will continue to exist until explicitly modified or deleted; and
- v. the number of expressions that the system can hold is unbounded.

Thus, such a system is then called a physical-symbol system and such a system has the necessary and sufficient means for general intelligence action. Necessary here means that if any system shows general intelligence then that system will be a physical-symbol system<sup>47</sup>. Sufficient here means that any physical-symbol system of sufficient size can be organized so as to exhibit general intelligence. General intelligence is the same scope of intelligence as exhibited in human actions. This means the system can actualize behaviours appropriate to its ends and also at the same time adaptive to the demands of the environment of the system in a real situation.

Having sketched out such a symbol system and owing to the physical-symbol system hypothesis, it is then assumed that any intelligent system will then have to be understood and analysed as a symbol system of the kind described above. The task then becomes that, when one encounters any intelligent action by a system or once a system is classified as an intelligent system, one has to then search for and identify the symbol system in it, based on which then that system may be further analysed as a symbol system. An analogy may be drawn with a law of qualitative structure regarding diseases known as the ‘Germ Theory of Diseases’<sup>48</sup>. According to this theory, tiny single-celled organisms are the cause of diseases and diseases are transmitted due to transmission of these germs from one host to another. So the task here then becomes of first identifying the disease and then looking for the germ. Similarly in the field of AI, one has to first identify a task-domain that requires intelligence<sup>49</sup>. It is the task then to provide for algorithm for a system to handle those tasks in that domain. Thus, the task of playing chess, which was accepted as a domain that requires intelligence, was early on imbibed in AI research. But chess makes up for only one of the tasks that require intelligence. There are many such other tasks requiring intelligence. The programs performing these various tasks then may have common components. It may be then understood that there might be some general nature of the tasks or

---

<sup>47</sup> Ibid., p. 111.

<sup>48</sup> Ibid., p. 108.

<sup>49</sup> Ibid., p. 115-116.

problems that require general intelligence to be carried out or to be solved. Thus, there was a considerable interest in searching for mechanisms possessing generality and also for common components among programs performing a variety of tasks. One such program was the General Program Solver.

Having described the physical-symbol system that provides the basis for carrying out intelligent actions, the question is – how does such a system accomplish this? Newell and Simon give the ‘Heuristic Search Hypothesis’. They propose that a physical-symbol system behaves intelligently in problem-solving by ‘search’. The solutions to the problems are represented as symbol structures. The system then searches for the solutions by generating and modifying progressively the symbol structures until the solution is produced. But does it do so? Newell and Simon describes how even Plato described such a problem in the *Meno*. The question is how does one inquire into what one does not know? If one doesn’t know what one is inquiring about then there is no subject of inquiry. So what is the subject of inquiry here? How will one put it forward? And even if one finds what one is looking for, how will he know if that is what he did not know?<sup>50</sup> For Plato the solution to this problem was the theory of recollection. When one learns something then he is just recalling what he already knew. Newell and Simon describe this problem based upon the symbol systems:

“To state a problem is to designate (1) a *test* for a class of symbol structures (solutions of the problem), and (2) a *generator* of symbol structures (potential solutions). To solve a problem is to generate a structure, using (2), that satisfies the test of (1).”

A test defines the problem. To have a generator, there must first be a problem space where both the initial situation and the goal situation can be represented. Generators are processes for modifying one situation in the problem space into another. A physical-symbol system possesses both the problem space and the generator. The task then, that the symbol system has to perform, is to generate possible solutions one after the other until it finds one that satisfies the test. But to exhibit intelligence it should not generate solutions randomly. The system has to have some control over the generation of possible solutions and it must have some order of generation so that there is a high likelihood that the actual solution appears as soon as possible in the process.

---

<sup>50</sup> Ibid., p. 120.

One condition then, for the system to appear intelligent, is that the space of symbol structures exhibit some degree of order and pattern. Also, it must be the case, which is also the second condition, that the pattern should be detectible. Lastly, the third condition is that the generator be able to behave differentially, depending on the pattern detected.

There are some core mechanisms<sup>51</sup> used by symbol systems for intelligent problem-solving. Firstly, the generator produces a new expression by modifying the expression already present in the system and which expresses the problem. The next expression is again generated by modifying the previous expression and so on. This assures that the generated expressions are relevant to the symbol system and the problem designated by it. Secondly, the modification does not occur haphazardly. They depend on two kinds of information – the information that is already built into the structure of the generator, and the information on the difference in form between the present expression and the desired expression. The first kind of information is that all kinds of expression generated must not affect the solution for it to change. This information is in the generator and thus, this principle is constant throughout the whole process of generation (by modifying) of expressions. It guarantees that only a tiny number of possible solutions are generated. The second information is used for arriving at solutions by a succession of approximation. It is done by employing a simple form of means-end analysis to give direction to the search. Newell and Simon describe two kinds of search – serial heuristic search and tree search. In serial heuristic search the basic question is always – what shall be done next?<sup>52</sup> In tree search that question has two components – from which node in the tree shall the search begin and what direction from that node shall the search take? To answer the first question some factors may be considered such as the relative distance of different nodes from the goal. The closest node then gives the best possible node to start from so as the goal may be reached in least steps. The answer to second question depends on factors such as the specific differences between the current nodal structure and the goal structure described by the test of a solution. Action then may be selected based on the differences so that the difference may be reduced. This is same as described earlier. The technique is known as means-end analysis, which plays a central role in the structure of General Problem Solver.

---

<sup>51</sup> Ibid., p. 123

<sup>52</sup> Ibid., p. 125



In another instance, during 1960s, Lotfi Zadeh introduced his idea of ‘fuzzy’ sets and systems<sup>53</sup> which meant that computers do not have to operate in a merely binary, logical format. Warwick gives an account of fuzzy sets and systems<sup>54</sup>. It is found useful in certain circumstances for conclusion to be partially true or a confidence percentage be applied to results. *Fuzzy logic*<sup>55</sup> makes it possible to do so and not simply work limited to the binary of true or false. So the first step here then is to take an actual real world value and make it *fuzzy*<sup>56</sup>. The relation between the actual value and the fuzzy value needs to be well defined through graphical means or a look up table or even through mathematical relationships. Once the *fuzzification* of the value is done, it is passed to the rules for evaluation. It is normally the case that different rules will fire and thereby different value will be taken forward. To provide a single end value they must be aggregated. This is termed *defuzzification*<sup>57</sup>. In this process, each rule is assigned with an associated *weighting value* known as *Centre of Gravity*. The resultant values after fuzzification are multiplied with their respective weighting values, they are then added together and divided by the sum of all weighting values. This way weightage of a rule may change the output value. Such expert systems are termed *Fuzzy Expert Systems*<sup>58</sup>.

Besides General Problem Solver and fuzzy sets and systems, the 1960s saw mostly ideas of attempt to get a computer to copy human intelligence, to make computers understand and communicate in natural human language rather than machine code and so on. This was partly due to Turing’s idea of intelligence and partly due to desire for computers to interface with the real world. The discourse on AI and the relevant concepts so far may be termed as Classical AI.

### 3. The divide

The development in AI can be classified into two schisms. First there is the symbolic AI which is based on symbolic processing, is sequential and is based on propositional logic. This is often termed the Good Old Fashion AI (GOF AI)<sup>59</sup> and is considered the classical approach. Here

---

<sup>53</sup> Warwick, 2012, p. 3.

<sup>54</sup> Ibid., p. 39-44.

<sup>55</sup> Ibid., p. 39.

<sup>56</sup> Ibid., p. 39, 40.

<sup>57</sup> Ibid., p. 42.

<sup>58</sup> Ibid., p. 43.

<sup>59</sup> The term GOF AI was dubbed by Haugeland, John in *Artificial Intelligence: The Very Idea*, Cambridge, Massachusetts: MIT Press, 1985; Walmsley, Joel, *Mind and Machine*, Hampshire, England: Palgrave Macmillan, 2012, p. 27.

the truth values (T/F) of logic are mapped onto the 0/1 of individual states of Turing machines or even onto the on/off states of brain cells. This is a formal approach and thus the algorithms are syntactically driven rather than semantically. The core idea behind classical AI was that only one approach, namely, the Turing computation, could be applied to both human and machine intelligence<sup>60</sup>. However GOFAI wasn't the only one to have its basis on logic. Connectionism was one such approach based on the paper 'A Logical Calculus of the Ideas Immanent in Nervous Activity' by McCulloch and Pitts<sup>61</sup>. However its architecture was based on the neural networks of the brain. It was more biologically realistic. Also it was pointed out by McCulloch and Pitts that thermodynamics is closer to the functioning of mind than logic<sup>62</sup>. Thus, statistics and other approaches were used alongside logic. It was seen as an extension of the symbolic system not just an alternative to it. The second schism is that of cybernetics. The focus of the cyberneticians was mainly on biological processes such as adaptation, metabolism and so on. Their area of study was mainly the self-organization in biological systems. The core concept of cybernetics was goal-guided procedures; the current distance from the goal was used to determine the next step. The researches in this field mostly focused on engineering and developing analogue computers rather than logic and computation. However the distinctions between the two weren't always clear cut. Even symbolic AI had goal-guided problem solving algorithms.

From around 1960, the intellectual divide between the two schisms became more pronounced<sup>63</sup>. Those interested and working on *life* (self-organizing biological systems, also this area was later known as *artificial life*) stayed in cybernetics and those interested in *mind* stayed in symbolic computing. There may also be drawn a third schism, of those interested in both, the network enthusiasts. The term applies to those who were interested in areas such as *artificial neural network* which requires both symbolic computing and research in biological systems such as brains and other neural systems, to try and replicate them artificially. Before the divide there was a mutual respect among the schisms but since 1960s onwards the respect increasingly

---

<sup>60</sup> Boden, 2016, p. 10.

<sup>61</sup> McCulloch and Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity", in *The Philosophy of Artificial Intelligence*, Margaret A. Boden (ed.), 1990, pp. 22-39.

<sup>62</sup> Pitts and McCulloch, "How We Know Universals: The Perception of Auditory and Visual Forms", *Bulletin of Mathematical Biophysics*, 9, reprinted in S. Papert (ed.), *Embodiments of Mind*. Cambridge, MA: MIT Press, 1965, pp. 127-47.

<sup>63</sup> Boden, 2016, p. 17.

diminished among them. This was partly due to the huge success and popularity achieved by symbolic computing. This was followed by AI winter or the Dark Ages of AI. But it was a winter, more specifically, for a particular schism, that is, the neural network camp. Perhaps the least affected was the symbolic computing camp.

#### **4. Dark Ages of AI Research**

1970s is considered to be the Dark Age of AI due to failure of high expectations and promised results of 1960s<sup>64</sup>. Marvin Minsky and Seymour Papert attacked the field of neural networks on the inability of perceptrons to generalize in order to deal with certain types of simple problems. Such lacking was mostly due to the limited capabilities of computers in the 70s as compared to computers today. The main problem was limited computing power and not enough memory which are a requisite to communicate or understand just like how a human requires a lot of information and a lot of processing power. Besides such technical difficulties, the field of AI also became a topic of interest for philosophers such as John Searle who came up with his Chinese Room Argument<sup>65</sup> to show that a computer cannot be said to understand the symbols with which it communicates. Machine cannot be described as thinking, as Turing had previously postulated, purely in terms of symbol manipulation. While on one hand, researchers such as Minsky felt that such arguments should be ignored, on the other hand, John McCarthy considered that operation of human brain is not directly relevant for AI. For him what were needed were machines that could solve problems and not necessarily think the same way as people do.

#### **5. AI Renaissance**

The 1980s saw revival in AI owing to three factors<sup>66</sup>. Firstly many researchers continued to develop AI systems from a practical point of view. ‘Expert systems’ were developed which were designed to deal with a very specific domain of knowledge. Secondly philosophical discussions and practical AI work proceeded with their own thing. AI developers focused on practical industrial solutions without necessarily claiming that computers should or could behave like humans. Thirdly, parallel development of robotics started to have a considerable influence

---

<sup>64</sup> Warwick, 2012, p. 4-6.

<sup>65</sup> The Chinese Room Argument will be taken up in chapter 3.

<sup>66</sup> Warwick, 2012, p. 6.

on AI owing to the belief that in order to exhibit ‘real’ intelligence, a computer needs to have a body in order to perceive, move and survive in the world. Without such skills, how can a computer ever be expected to behave in the same way as a human? Without these abilities, how could a computer experience common sense? Such bottom up sort of approach to AI building was what was originally postulated by McCulloch and Pitts.

Many aspects of AI development were based on aspects of human brain functioning. The ability to reason was taken up as an approach to AI systems. Warwick describes this, “If given a number of facts, the human brain can make a reasoned assumption about a situation and decide on a conclusion”<sup>67</sup>. Based on this the first successful AI system was built. The concept of an *expert system*<sup>68</sup> is that of a machine being able to reason about facts in a specific domain and to simulate roughly the workings of an expert brain. For this the machine requires domain specific knowledge, some rules to follow when new information occurs and some way of communicating with a user of the system. Such knowledge based systems are called *rule based systems*, *knowledge based systems* or expert systems. In an expert system the basic form of each rule is – IF (condition) THEN (conclusion)<sup>69</sup>. In case that multiple conditions must be exist for a rule to be true or one of a number of conditions could exist for a conclusion to be drawn, the rule may have the following form –

IF (condition 1 AND condition 2 AND condition 3) THEN (conclusion).

The actual rules employed are obtained by questioning a number of human experts. In cases of several possible conclusions a decision has to be taken, this then is referred to as *conflict resolution*.

There may be cases where several conclusions are met but only one conclusion is required. A decision is necessary as to which of the rules take precedence. The rule that is selected depends on the following criteria<sup>70</sup> as described by Warwick –

1. High priority rule – each rule has a priority associated with it and if several rules apply, the one with highest priority is selected.

---

<sup>67</sup> Ibid., p. 32.

<sup>68</sup> Ibid., p. 32.

<sup>69</sup> Ibid., p. 33.

<sup>70</sup> Ibid., p. 34.

2. Highest priority condition – each condition has a priority associated with it. For a rule to be chosen, it must contain the highest priority conditions.
3. Most recent – the rule whose condition has most recently been met is chosen.
4. Most specific – the rule which has most conditions met is selected.
5. Context limiting – rules are split into groups. Only some are active a certain time. To be chosen a rule must belong to an active group. In this way the system can adapt over time. In case of multiple rules, rules are structured in layers. When all conditions are met for one rule such that its conclusion is drawn, that conclusion can in turn meet a condition for a rule in the next layer and so on. This can be in the following form<sup>71</sup> -

Layer 1

IF (condition 1) THEN (conclusion 1)

IF (condition 2) THEN (conclusion 2)

Layer 2

IF (conclusion 1 AND conclusion 2) THEN (conclusion3)

Warwick explains that in an expert system “a set of facts will be apparent at a particular time and these will fire a number of rules, realizing further facts which fire other rules and so on until an end conclusion is reached”<sup>72</sup>. This is termed as *forward chaining*. The reverse of this is also described by Warwick: “... when a goal has been achieved the rules are then searched to investigate what facts (data) occurred in order for the system to reach the conclusion that it did”<sup>73</sup>. This is termed *backward chaining*. Further, one can look backwards through the system to find out what facts must be realized as input by the system in order for it to realize a specific goal.

There are some advantages of expert system over other AI approaches<sup>74</sup> – firstly, they are easy to program into a computer; second, they are ideal for dealing with natural world information; third, the system structure is separate from data and hence separate from the problem area; fourth, expert systems can deal with uncertainties; and lastly, speed of response is

---

<sup>71</sup> Ibid., p. 34-35.

<sup>72</sup> Ibid., p. 35.

<sup>73</sup> Ibid., p. 36.

<sup>74</sup> Ibid., p. 36, 37.

good. However, there are also few problems<sup>75</sup> – firstly, standardization of expert opinions into rules may be problematic. Secondly, the expert system becomes too big leading to combinatorial explosion. Aim of such a system is to draw a conclusion no matter what the situation. But to deal with every eventuality, rules must be continually added to cover every possible situation.

Warwick gives an alternative of the definition of intelligence - “The variety of mental processes that act together to make up those necessary for life”<sup>76</sup>. This definition ties intelligence to mental processes and also gives it a central role in terms of life and living entities. This then raises the question – What is life<sup>77</sup>?

Warwick goes on to talk of on artificial life also known as **A-life**<sup>78</sup>. There are different approaches to A-life. What is common among them is that some aspects of life as we know it is taken as the basis for the A-life study. Thus, life can be modeled and used as a basis for AI technique. Two A-life approaches<sup>79</sup> -

- *Approach 1*: Computing AI methods such as Genetic Algorithm<sup>80</sup>, neural networks<sup>81</sup>, evolutionary computing<sup>82</sup> and software agents<sup>83</sup> are all based on what happens in nature, i.e., how brain works and how life evolves.
- *Approach 2*: Different aspects of life can be picked on and employed along with standard forms of AI to improve performance. For instance models of social and cultural change may be used.

*Cellular automata*<sup>84</sup> also known as finite state machine is a form of simulated A-life approach. This consists of cells. A cell is surrounded by neighboring cells in a particular relationship with each other. The relationship of a cell with one neighboring cell may be different from its relationship with another neighboring cell. As time progresses the state of a cell changes depending on its previous state and the state of its neighbors. As such when simple operational

---

<sup>75</sup> Ibid., p. 37, 38.

<sup>76</sup> Ibid., p. 116.

<sup>77</sup> This may be compared later on with the concept of life-force or vitality principle.

<sup>78</sup> Warwick, 2012, p. 117.

<sup>79</sup> Ibid., p. 118.

<sup>80</sup> Ibid., p. 102 – 107.

<sup>81</sup> Ibid., p. 92 – 101.

<sup>82</sup> Ibid., p. 102.

<sup>83</sup> Ibid., p. 109 – 111.

<sup>84</sup> Ibid., p. 119.

description is viewed over a sequence, extremely complex patterns may emerge. This setup allows the study of effects of society in that the state of an individual not only depends on itself but also on those around it. Such cellular automata also exhibit evolutionary behavior. Also arrangements such as wrap-around<sup>85</sup> may be applied so that a two dimensional layer can have its edges linked. A two dimensional layer of cells as described earlier may be imagined. A cell in the corner has only three neighbors while cells on edges have five neighbors which may lead to a biased effect on the whole population of cells. So a wrap-around is applied. Here wrap-around is used in the sense of a piece of paper that may be folded so that its edges are touching. But in this case, the fold is diagonal so that the cells in the diagonally opposite edges become each other's neighbor and a relationship is formed between them. This way the biasness due to unequal numbers of neighbors is mitigated. Such A-life simulation may be used to bring about real-life modification and are themselves based on real life inspirations. Also there are cases of cellular automata which are reversible. These are useful in studying physical phenomena and obey the laws of thermodynamics. Based on the present state the previous state can be found out. For non-reversible cellular automata, patterns can exist for which there are no previous states. Such patterns are referred to as *Garden of Eden Patterns*<sup>86</sup>. Also such simulated A-life may make use of software A-life<sup>87</sup> to create virtual worlds which is a representation of the real world. Based on these representations a machine can quickly actualize movements and modifications in real-life world.

In true sense of A-life what is needed are physical entities that exist in the real world. This is where hardware robots come in. Real world affectations are realized by hardware A-life entities by means of a robot base. A form of collective or collaborated intelligence is *Swarm Intelligence*<sup>88</sup>. It is typically made up of simple robots or agents interacting with one another as well as the environment. There is no centralized control structure dictating how individual agents ought to behave but the interactions between such agents lead to emergence of apparently intelligent behavior.

---

<sup>85</sup> Ibid., p. 121.

<sup>86</sup> Warwick, 2012, p. 124.

<sup>87</sup> Ibid., p. 125 – 127.

<sup>88</sup> Ibid., p. 136.

## 6. Knowledge representation

For any system to function in a given environment requires for representation of the problem as well as knowledge or information on the environment to the system in a way that the system can understand or deal with. The environment provides the condition and the system based on the conditions performs some action relevant to the environment and the problem at hand. For instance one may consider the expert systems described earlier. They not only require domain-specific knowledge but also new information or knowledge has to be represented to it in such a way that it may act upon it. Also one may consider the case of A-life where real world affectation to be realized by the system has to be represented in a way that the machine actually gets affected. Also AI learning and methods such as heuristic search requires that relevant knowledge should be represented to an AI computer in a manner that it can operate upon it. One of the core issues of AI is solving the problem of such representation to the machine. The condition need not be that of environment but of the machine itself, as in a theoretical model, such as the Turing Machine, the machine is considered standalone. The knowledge representation then will partly depend on the nature of the system and partly on the environment/conditions. In a symbolic system, which is rule based, a body of knowledge may be represented as a set of IF-THEN rules linking conditions to actions – IF this condition, THEN take this action. This kind of knowledge representation is based on formal logic. In rule based programming, a rule is triggered only when its given condition or conditions are met. This form of knowledge representation allows programmers to build the program gradually as he/she learns more about the problem area. A new rule may be added anytime as and when needed. Instances of such knowledge representation are evident in the expert systems discussed earlier in section 1.1.5. Another method for knowledge representation is done by specifying a *frame*, such as the frame of a room, specified by the programmer by specifying a hierarchical data structure. For instance the programmer may specify that a room has four walls, a ceiling, a floor, a window, a door, furniture and so on. Of course, actual rooms may not have furniture and then specifications such as four walls, a ceiling, a floor etc may be arranged in hierarchical order such that the machine may understand the frame even when some specifications are not met. Another way of knowledge representation is when concepts may be represented by word vectors rather than words. In this kind of representation, first semantic features that connect many different concepts



are identified by the system (such as deep learning) itself. These features (which now serves as vectors or directing features) are then used to predict other words or following words, for instance in machine translation. Scripts are data structures that denote familiar action sequence. Basically, these data structures provide a storyline that is to be followed, for instance, cooking noodles require that the pot be placed on the stove and the stove be turned on, water to be poured into the pot and noodle to be put into the pot. Such data structures may be used in question-answering. This form of knowledge representation may then enable machines to converse (to have a conversation or at least perform a simulation of it) with humans (just like in the Turing's test or the Imitation Game). Suppose if the water is not poured into the pot, a question may be asked as to what is wrong and the machine may give the answer that water has not been poured yet. Also if asked what happens next, the machine may tell you the water boils. Conversely the machine may also raise questions. An alternative to word vector for representation of concepts is the semantic network. Here, a semantic network links concepts by semantic relations such as synonymy, antonymy, part-whole and so on. The network may link words as well as concepts. However, semantic networks are not the same as neural networks.

In neural network representation of knowledge concepts are represented not by a single node in a defined associative net but rather by the changing pattern of activity across an entire network. A category of neural network is the Parallel Distributed Processing (PDP) network. A PDP system is made up "... of three or more layers of interconnected units, each unit capable of computing only one simple thing"<sup>89</sup>. What and how a unit computes may differ and thus the units differ. It differs in the sense that a unit may stand for only a micro-feature of a particular concept. But otherwise a unit's activity is mostly limited to a state of on/off. Each unit can be part of many patterns and so can be part of (and contribute to) other concepts. Among the three or more layers, there is an input layer and an output layer<sup>90</sup>. The rest of the layers or the middle layer(s) is/are called hidden layer(s). Correspondingly, a unit is either an input unit or an output or a hidden unit. Unlike the input and output units, a hidden unit has no direct contact with the outside world. A PDP system then involves distributed representation of concepts. Each concept is represented by the state of the whole network. For different concepts there can be then a corresponding state of the same network. Such a network may even learn and thereby represent

---

<sup>89</sup> Boden, Margaret A., *AI: Its Nature and Future*, 2016, p. 82.

<sup>90</sup> *Ibid.*, p. 83.

new concepts. To represent a new concept all that is needed is to modify the interactions between the units so as to create new stable pattern of activity. This modification may occur gradually over many separate occasions.

## **7. The Present and The Future**

After the renaissance period, the field of AI once again gained momentum in areas such as industrial applications, financial systems and the military<sup>91</sup>. In these areas AI applications were not only a replacement of human operatives but were also much better in performance than their human counterparts. In 1997, Deep Blue became the first chess-playing computer system to beat a reigning, world chess champion. In 2002 Kevin Warwick successfully linked the human nervous system directly with a computer to realize a new combined form of AI. Many important successes were not due to new invented form of technology but rather due to pushing the limits of technology available mostly in terms of computing and memory. Increase in computing power is followed and predicted by Moore's law which indicates that the speed and memory capacity of computers doubles every two years. Earlier problems are being overcome by sheer computing power. Novel approaches to AI have also emerged, for example intelligent agents, a modular approach which could be said to be mimicking the brain in some ways by bringing together different specialist agents to tackle each problem. An intelligent agent is a system in itself in that it must perceive its environment and take actions to maximize its chances of success. Some other approaches such as probability and decision theory are mathematical in nature. Neural networks and concepts from evolution such as genetic algorithms also have played an influential role. With the advent of wireless technology emerged what is termed distributed intelligence<sup>92</sup>. Until that time what existed were standalone computers. With networked computers it became realistically necessary to consider the entire network as one large intelligent brain.

Most of the discussion on AI approaches has been with the concept of AI based on a machine base. But recent approaches are focusing on cultured biological brains. It is mostly due to the idea of consciousness as being of emergent nature of a collective of biological neurons. A

---

<sup>91</sup> Warwick, 2012, p. 7.

<sup>92</sup> Ibid., p. 9.

cultured brain<sup>93</sup> is created by collecting and separating the neurons found in cortical brain tissues using enzymes and then growing them in an incubator in a suitable environmental conditions and nutrients at a constant temperature. The culture is connected with its robot body by allowing the neurons to growing a small dish on the base of which is an array of flat micro electrodes. This provides an electrical interface with the neuronal culture. The neuronal cultures grow and form a layer over the electrode array, effectively growing into a two-dimensional brain. The electrodes enable output voltages from the brain to be monitored and for the brain to be simulated by the application of appropriate voltage signals. In this way both motor output and sensory input can be achieved. A closed-loop feedback structure is formed between the robot body and its cultured brain.

Such biological AI brain research are receiving focus in terms of learning, growing such structures from human neurons, incorporating the culture directly into a robot's head instead of wirelessly linking incubator and robot body and to increase the overall size of the culture in terms of the total number of neurons contained. A primary step is to shift towards a three dimensional growth rather than two dimensional. Questions may arise – *will it be endowed with genuine understanding and therefore genuine intelligence?*<sup>94</sup> The important point here is that placing a biological brain within a robot body bridges the gap between the operation of human brain and that of a computer or machine brain. This in turn will raise a new paradigm of questions.

Point of interest today is the effect of the body on the intellectual abilities of that body's brain. Ongoing research aims at realizing an AI system in a body so it can experience the world. This is embodied AI<sup>95</sup>. Also researches are focused on, as mentioned above, AI brains grown from biological neural tissue. AI may no longer be based on a computer system but rather on a biological brain that has been grown a fresh. Such ideas provide new areas of study in terms of its questioning of many of the philosophical assumptions regarding AI, such as the question of difference between human intelligence and machine intelligence. Current researches are also going on regarding cyborgs<sup>96</sup> (cybernetic organism), i.e., biological AI brain being given a

---

<sup>93</sup> Ibid., p. 140.

<sup>94</sup> Ibid., p. 143.

<sup>95</sup> Ibid., p. 10.

<sup>96</sup> Ibid., p. 11.

technological robot body, thereby, an embodied brain.

At any stage of AI development, the enthusiasts of AI had always hyped its future prospects. One such prediction is that of *Singularity*. Singularity is some proposed time in future when machines become more intelligent than humans<sup>97</sup>. It is predicted that, first AI will reach human level intelligence. It is assumed that it will be real intelligence. Eventually, once it has reached human level, it will surpass the humans, by creating copies of itself or even programming better AIs that surpass the current level. AI will then be able to out-think humans and thus provides a conflict of interest. This conflict of interest is apparently due to the gained ability of AI to think for itself and its survival. It has been observed that the planet has mostly been dominated by humans due to their intelligence. Once there is a more intelligent being, then the control will be wrested from human hands. As such the future of AI seems to be a threat to human civilization. On the other hand there are a few optimists around too. Contrary to the above prediction they predict that once that point in time has been reached, AI will then be able to solve all the problems of human and doesn't need to necessarily lead to the doom of human civilization but rather leads to a better world. Another prediction post singularity is that a point will come at which biological humans will be integrated with machines. As a result, personal death may be eliminated and the cyborgs will be far superior to a normal biological human. However these are all too hyped and optimistic claims of AI. The immediate problems that the AI faces are that of reaching the general intelligence level, knowledge representation and so on.

---

<sup>97</sup> Boden, 2016, p. 147.

## Chapter 2

### Philosophical Discourse on the Nature of Mind

The aim of this chapter is to glean out various theories on mind from within the philosophical discourse on mind. The discourse on mind and thereby also the body is perhaps one of the oldest discourses in philosophy. The genesis may as well be traced back to the first of philosophers or even prior. The path back is one of admirable complexity. Not only the discourse seems to have evolved but also the terms and concepts used seem to have undergone considerable changes from philosophers to philosophers and epoch to epoch, leading almost to a sort of incommensurability among the different theories on mind in different periods. Also in the history of philosophy, this discourse had to be often cut out of some other subject domain (ethical or religious) where the core issue is not always the mind. Much had to be gleaned out of discourses of different nature but still relevant was the concepts of mind (soul). However, a narrative may be composed out of this complex historical flow centered on the question- “What is the nature of mind?” And almost like some kind of entailment or deduction, another subsequent question follows – how is the mind and body connected or related? A more specific question pertaining to the relation between mind and body is that of causation – “How does the mind causally interact with the body and vice-versa?” Another such question may be – “How are mental properties related (causally or otherwise) to physical properties?” This set of problems is what has been termed, *the Mind-body Problem*. This kind of questions usually arises when a theory posits fundamentally distinct nature for the mind as opposed to the nature of the body. To delve further into this problem it is only appropriate to start at the concepts employed in this problem. It may also be pointed out here that the discussion of mind-body problem has not always been so in terms of mind and body. Often the distinction has been between soul and body or soul and matter.

#### 1. Ancient Notions of Mind/Soul in Western Philosophy

Perhaps it might be prudent to start with the notions of mind in ancient Hebrew discussions and the Archaic Greek ideas found in Homeric literature. In Hebrew texts such as the *Old Testament*, the term *nepesh* may be considered as appropriate for soul. Besides *nepesh* there are other terms such as *ruach* and *leb*. These Hebrew terms had subtle nuances to them which

would alter their meaning based on whether it is predicated to a human or animal, whether it is used as a subject or an object and so on. Also these terms had undergone evolution. For instance, *nepesh* is used to refer primitively to the throat. As it evolved, a secondary meaning of it refers to desire or longing. Subsequently, its abstract meaning refers to life-force, that which sustains desire. Also the term could refer to the “I” or the self, that is the individual. However, though the transition of meanings of such terms vectored in various directions, Daniel Lys in his investigation of the Greek translations of the term *nepesh*<sup>98</sup> in the Greek Version of the Old Testament, LXX (also known as Septuagint, c. 250 BCE), had remarked “the LXX never goes in the direction in which “soul” would be understood as opposite to “body” (as in Platonic dualism).”<sup>99</sup> This points out that one of the earliest ideas of soul and body was not as two distinct entities governed by distinct laws or principles of their respective nature, but rather they were thought of consisting of the same nature. Even the reference of the terms used and thereby their meanings had the sense of both psychological and physical. This presents us with a kind of conceptual framework whereby soul/mind could be talked of in terms which doesn’t shed off the physical senses of the psychological concepts. Perhaps such an understanding of the concepts of mind may then be applied to posit a theory which then may serve as a theoretical basis to address various issues such as consciousness, intentionality and mental causation.

While the ancient Hebrew notions of mind were not opposed in nature to the body, a hint of dualism may be found in the Homeric notions of the mind. MacDonald describes Homer’s concept of human nature by distinguishing between two sets of central psychological terms: “One set is comprised of *psychē*, *thymos*, *menos* and *nous*, each of which is soul-related (psychical) function necessarily dependent on an animal body... and as a component of the whole being (an ensouled animal).... The other group comprises *kēr* (or *kradiē*), *phrenes*, sometimes *hēpar*... each of which is an organic internal component of the animal body, occupies a definite... internal site, and upon which the first set of psychological functions has a necessarily dependent relation...”<sup>100</sup> Further MacDonald asserts that “our modern post-Cartesian concept of mind has

---

<sup>98</sup> Lys discovered that for 754 passages in the OT that use *nepesh*, the Greek text used *psychē* in 680 cases; see MacDonald, Paul S, *History of the Concept of Mind*, England: Ashgate Publishing Limited, 2003, p. 11.

<sup>99</sup> Lys, Daniel, *Ruach: Le soufflé dans l’Ancien Testament*, EPhR, annual issue, 1959, p. 227; see MacDonald, 2003, p. 11.

<sup>100</sup> There are various standpoints of the tracing of ancestry of concepts in Homeric ideas of soul, see MacDonald, 2003, pp. 12-22.

its *closest ancestor* in Archaic Greek thought in Homer's *nous*"<sup>101</sup>, while "The closest ancestor in Archaic Greek to ... pre-Cartesian concept of soul or *anima* is *psychē*"<sup>102</sup>. This point by MacDonald gives insight into the beginning of how the concepts and the terms regarding "soul" or "mind" may have evolved into shedding their physical connotations. Also at the same time it provides for concepts that, like the Hebrew ideas on soul, might make up a conceptual framework that provides for a theory that does not inherently distinguish the nature of soul in opposition to the nature of body.

The transition of the discourse on mind from Homeric notions to pre-Socrates notions is characterized by the continuation of soul terms that were more or less in line with the earlier notions of their predecessors. For instance Thales (c. 624 – c. 546 BC), who is accorded the title of the first Greek philosopher<sup>103</sup>, asserted that the soul pervades the whole world. The subsequent pre-Socrates philosophers, like Thales, more or less echoed the idea of a single entity comprised of some basic fundamental element as constituting the world and the individual, the living and non-living, the psychical and the physical. However as a precursor to Platonic dualism, pre-Socrates philosophers also brought the notion of soul as distinct from the body. For instance, Anaxagoras (c. 510 – c. 428 BC) and Empedocles (c. 495 – c. 444 BC) both posited their own notions of soul which were similar to a dualist notion of soul. An important mark of pre-Socrates discussion of soul or *psychē* or *nous* is the notion of soul or mind as an entity and yet a quality of the matter that it inhabits or is related with. It is an entity in the sense that it is not a quality of an object that serves as the basis of that quality's existence. Usually a quality depends on the object it is a quality of and the object serves as a subject for that quality of which it is predicated of. It is not so in the case of soul, the object doesn't serve as a subject. However, the soul or mind is usually thought to exist in material objects (living ones) and thus, seems to be a quality of those objects. The soul is not typically talked of existing independently of the body. It is, then, understood as both a quality and a thing. Thus, the pre-Socratic idea of the nature of mind may

---

<sup>101</sup> However, MacDonald points out, "where Descartes and his descendants equate mind with the faculty of reason, Homer characterizes *noos* (mind or intellect) as one dimension of the human being." See MacDonald, 2003, p. 22.

<sup>102</sup> MacDonald points out that the ancestry can be established "... in so far as Homeric *psychē* characterizes the uniquely human life-force or vitality which does not survive bodily death." See MacDonald, 2003, p. 22.

<sup>103</sup> Russell, Bertrand, *History of Western Philosophy*, 1996, Special Indian edition, New York: Routledge, First Indian Reprint, 2013, p. 15.

be referred to as a “quality-thing”.<sup>104</sup> One may then consider that the pre-Socrates discussions of soul marks an intermediary phase in the transition of the concept of soul from one that is not distinct in nature from its physical aspect to one where soul is considered as fundamentally of different nature than the body. The idea of human-nature thus, moves from a non-dualistic picture to a dualistic picture of Platonic dualism.

## 2. Dualist Approaches to Mind

It was seen that the path to dualism had already been laid out during the pre-Socrates era. However, Plato (4<sup>th</sup> century BC) was the first one to present written arguments in favor of dualism. Among the several arguments that Plato provides in favor of dualism, Maslin picks out one specific argument<sup>105</sup>. The argument is formally stated by Maslin as follows:

1. The user of a thing and the thing used are two numerically different and distinct things, i.e. two logical substances.
2. A person uses his or her body.
3. Therefore a person must be numerically different and distinct from his or her own body.

But if a person is different from his or her body, he or she must be a non-bodily logical substance, i.e. a soul.<sup>106</sup>

Maslin picks out this specific argument for the reason that this argument exhibits a common error. He analyses the above argument. The first premise can be accepted as true if the thing used is a utensil, knife or some such tools. The question then is raised – do people use their limbs and bodies in the same way as these tools? Maslin points out that one may as well say, “Use your mind” and this doesn’t mean that the mind is a tool and hence a numerically different and distinct thing from the person. This kind of arguments, says Maslin, lands dualists into trouble. Thus, argument such as this is not reliable at all in supporting their assertion of dualism.

---

<sup>104</sup> For more details, see MacDonald, 2003, p. 34; also see Granger, Herbert, *Aristotle’s Idea of the Soul*, Dordrecht: Kluwer Academic, 1996, pp. 149-50.

<sup>105</sup> See Maslin, K. T., *An Introduction to the Philosophy of Mind*, Cambridge: Polity Press, 2001, pp. 36-41.

<sup>106</sup> *Ibid.*, p. 39.



Whatever the objection to Platonic dualism may be, what Plato presents is the concept of soul that is understood as immortal, intelligible, uniform<sup>107</sup> and distinct from the body. Plato posits two worlds. In the first world reside the *forms* which are perfect, eternal, immutable essence or archetypes. And it is the forms which make the other world, the material world possible and intelligible. Dualism as such then presents a possible picture of understanding the mental phenomena whereby the soul is understood as similar to the forms and thereby immaterial, immortal, intelligible and so on. And it is due to its resemblance to the forms that the soul is able to make sense of the world.

Contrast to this notion of dualism of form and matter, where form exists separate from matter and matter becomes intelligible only through form, is Aristotle's (384 – 322 BC) notion of form and matter, which asserts that though forms are distinct from matter, yet its existence cannot be without matter<sup>108</sup>. If there is form then it is only as the form of matter. Form is the predicate that is predicated of a subject which is matter. In *De Anima* Aristotle draws a similar picture of soul and body. He asserts that soul is predicated of a subject that is the body. The body cannot be predicated of the soul. The body can only serve as the subject that has soul. Common themes between Platonic and Aristotelian notions are, firstly, the comparison of the notion of soul with the notion of form and secondly, that of the distinction of soul from body. In Platonic idea, the soul is similar to form and hence the soul is immaterial, intelligible and immortal, while in Aristotelian idea, the soul is similar to form and hence it is immaterial, a predicate and thereby cannot be thought of existing without its subject, the body. It is due to this similarity of soul with the form that the second common theme also follows. In Plato's case the soul in being similar to form is immaterial and hence is separate from the body which is material. Similarly so, in case of Aristotle, the soul in being similar to form (or rather soul is taken to be a form) is immaterial and thereby different from the body. However, the difference between the two notions lies in the relation between soul and body. In Plato's case body depends on soul to be intelligible, while in Aristotle's case the soul depends on the body to exist.

As distinct from the body and yet depending on it for existence, the Aristotelian notion of soul is sort of like a hybrid entity, which is termed as “property substance” or “attribute-

---

<sup>107</sup> Plato in *Phaedo*, 80a-b in E. Hamilton and H. Cairns (eds), *The Collected Dialogues of Plato*. Princeton: Princeton University Press, 1961.

<sup>108</sup> For a discussion on Aristotle's doctrine see MacDonald, 2003, pp. 54-71.

substance”. Here substance has a restricted sense and hence does not mean an entity which can exist independently on its own. A. O. Rorty explains<sup>109</sup> that, in one sense the physical and the cognitive can be thought of and inquired upon separately from each other, however, they are not separate in their being. A further comparison of soul and body that Aristotle seems to make is with the eye and sight. If eye were an animal then sight would be its soul<sup>110</sup>. Sight is the function of an eye and similarly, soul is the function of an animal. Aristotle’s notion of soul then seems to be a functional notion and as such his idea of soul is similar to functionalism. In fact, antecedent to modern idea of functionalism may be traced back to Aristotle’s notion of soul as function<sup>111</sup>.

Another form of dualism is the Cartesian dualism. Descartes (1596 – 1650) through his method of doubt comes to the conclusion that in doubting lays the evidence that this doubting being exists. Maslin describes three arguments<sup>112</sup> forwarded by Descartes in support of his notion of mind as a distinct substance from the material substance. First, is the argument from doubt which is formally stated Maslin as follows<sup>113</sup>:

1. I doubt that my body exists.
2. I cannot doubt that I exist.
3. Therefore I must be different and distinct from my body.

In this argument from doubt, Descartes doubts everything that is given to him through sensory means. The reason he gives for doubting them is that it has often been the case that the sensory knowledge have deceived him or turned out to be false. Also, he invokes the idea of an evil demon whose intention is to deceive us and provide us with a sensory world that is false. And since it is possible to doubt the body’s existence but not the existence of his self, the thinking or doubting being, he concludes that he, the mind, is distinct from the body. The second argument is the argument from clear and distinct perception<sup>114</sup>. The argument may be stated briefly as follows:

---

<sup>109</sup> A. O. Rorty in Nussbaum, Martha & Rorty, A. O. (eds) *Essays on Aristotle’s De Anima*. Oxford: Clarendon Press, 1992, p. 8.

<sup>110</sup> MacDonald, 2003 p. 62.

<sup>111</sup> For such an analysis of soul as a function see MacDonald, 2003, pp. 61, 62.

<sup>112</sup> For detail discussion of the three arguments, see Maslin, 2001, pp. 50-64.

<sup>113</sup> Maslin, 2001, p. 53.

<sup>114</sup> *Ibid.*, p. 56.

*If I can clearly and distinctly understand one thing apart from another thing, then this is enough for me to believe that the two things are distinct from each other. Also, I know that I exist (from the first argument) and that I can perceive or understand myself only as a conscious being. And though I can perceive a body closely bound with my conscious being, yet I can clearly understand or perceive my conscious being as a distinct self, conscious and not extended. On the other hand I have an idea of my body as extended. And since I have distinct idea of myself and a distinct idea of my body, it may be concluded that I am really distinct from my body.*<sup>115</sup>

Objections to these arguments are raised but shall not be discussed here. Lastly, the third argument that Maslin describes is the argument from divisibility<sup>116</sup>. This argument is formally stated by Maslin as follows:

1. The body is divisible into parts.
2. The mind is not divisible into parts.
3. Therefore the mind must be of entirely different nature from the body, i.e. it must be essentially non-physical.

With these arguments, Descartes establishes the distinction of mind and body has two separate substances.

It is with Cartesian dualism that the mind-body problem is first presented in its present-day form. It is a matter of fact that, whether they are distinct substances or not, the mind and body interact. Mind causes physical actions. Thus, positing mind and body as two different substances, one also has to explain how the mind causes physical actions at all. How is the causal interaction between mind and body possible? How can two substances with different sets of properties and nature and obeying different sets of laws interact at all? And if there is a connection between mind and body where is it to be located? These questions form the present-day mind-body problem. Apart from the question of how mind causes bodily action, in contemporary theories on mind, the converse of this question that how physical brain causes

---

<sup>115</sup> See, Geach, P. and Anscombe, G. E. M. (eds.), *Descartes: Philosophical Writings*, Nelson University Paperbacks for The Open University, 1970, p. 128.

<sup>116</sup> Maslin, 2001, p. 63.

mind is another explanatory gap that is included within the mind-body problem. Descartes in facing the question how mind causally interact with the body, replied that it does so through pineal gland. In addition to positing pineal gland as the link between mind and body, according to Baker and Morris, he also asserts that “... *to have a Nature is to be a substantial union of mind and body, it follows that only a soul united to a body can possess the faculty of sensory perception, and only a body substantially united to a soul can possess the power of the body to act on the soul*”<sup>117</sup>. Descartes asserted that there are primitive notions that are most general and simplest ideas, not reducible to more simple ideas that explain them<sup>118</sup>. Three such primitive notions are evident in his doctrines – mind, body and mind-body union. All three are equally primitive in the sense that they are all equally simple and not further divisible into yet more simple ideas. These three notions correspond with three aspects of the natural world – the mental, the physical and their interrelation. The ability of mind to set bodies into motion, that is to cause physical movements, is evident through our everyday normal experience and sensation. Thus, the interrelation of mind and body itself is simple. To further explain it in terms of only physical and thereby mechanistic explanations is erroneous. Thus, Descartes’ position is away from that of a materialist’s though he does give scientific account for simple animated beings such as animals. Also in Descartes’ view the mind-body interrelation is something ordained by Nature (action of God), and thus, for him the matter is beyond metaphysical and scientific explanations<sup>119</sup>.

Descartes’ thesis may finally be understood in two ways. Firstly according to reason the mind is essentially a thinking, immaterial thing and the body is essentially an extended material thing. But as a being, human is a composite of two dependent wholes – mind and body. Problems arise when one tries to account for it in terms of only one of the conceptual schema.

An account of mind and body that maintains their distinct characteristics but posits only one substance and therefore only one nature or reality is that of Spinoza (1632 – 1677). His theories present a picture of the reality where there is only one single substance which is infinite

---

<sup>117</sup> Baker, Gordon & Morris, Katherine, *Descartes’ Dualism*, London: Routledge, 1996, pp. 172-4.

<sup>118</sup> This is suggested by David Yandell; Yandell, David, ‘What Descartes Really Told Elizabeth: Mind-Body Union as Primitive Notion’, in *Brit. J. Hist. Phil.* 5, pp. 249-73, 1997, pp. 250-53.

<sup>119</sup> Baker & Morris, 1996, p. 154.

and eternal<sup>120</sup>. This substance then has two kinds of attributes – thought and extension. Thus, mental properties and physical properties are just two attributes of the same substance. Spinoza's theory of one substance and thought and extension as the attributes of this one substance provides for a unique kind of explanation of the mind and also the world. The mind-body problem no longer occurs as there are no two distinct substances interacting causally with each other. Spinoza's theory provides for a picture of the world where the entire world can be seen, firstly as physical bodies ordered in causal linkages and secondly as series of ideas ordered in intelligible sequence. These two series correspond exactly with each other not due to some external contingent connection but due to them being the same ordered sequence viewed in two different ways.

A similar theory is also given by Leibniz (1646 – 1716). However for Leibniz, unlike Spinoza's one single infinite substance, reality comprises of simple units called monads which are infinite in numbers. These substances are of the same kind as far as they are all monads but each monad itself differs from all other monads, not in substance but in its qualities and internal principle<sup>121</sup>. This picture of reality then presents monads as constituting both mind and body. An important feature of monads is that they do not affect each other. How is it then that the two kinds of monad affect each other? According to Leibniz there is no affectations between the monads, they are window-less<sup>122</sup>. Everything that occurs is due to *pre-established harmony*<sup>123</sup>. An analogy may be drawn with two perfectly synchronous clocks. Just as there is no causal connection between the two clocks and yet the changes in each of them correspond to the changes in other, similarly so is the case with monads. There is an exact correspond between monadic changes and the sentient beings' ideas about those changes.

Both Spinoza's and Leibniz's theories is often termed parallelism. They both posit distinguishes between the mental and the physical and yet accepts only one substance or one kind of substance (in case of Leibniz) as the reality. The mind-body problem then more or less

---

<sup>120</sup> For discussion on Spinoza's substance monism see, MacDonald, Paul S, *History of the Concept of Mind*, England: Ashgate Publishing Limited, 2003, pp. 291-300.

<sup>121</sup> Leibniz, G. W., *The Monadology: An Edition for Students*, Nicholas Rescher (ed.), London: Routledge, 1991, pp. 72-4.

<sup>122</sup> Ibid., p. 58.

<sup>123</sup> Ibid., pp. 235-60.

disappears as there is no assertion of dual substances of distinct nature. Such theories too then provide for a picture of the world which attempts to account for the mind.

### **3. Varieties of Materialism (From Identity Theory to Functionalism)**

It is a contention of dualism that mental events can never be identical with physical events. Mental events are completely non-physical in all their aspects and consist in changes in non-physical states of an immaterial entity, which is the soul, and mind in case of Cartesian dualism. The mind-brain identity theory denies precisely this thesis of dualism. It contends that mind is not an entity that is separate and distinct from the brain. Mind is identical with a living brain and mental events are just brain events. Everything, consciousness, thoughts and so on, is purely material or physical. The theory, thus, seems to be an instance of materialism. The mind-brain identity theory came into prominence in the early 1950s and 60s and was advocated by philosophers such as U. T. Place, Smart and Armstrong. The mind-brain identity theory takes the stand that opposes the belief in minds as non-physical in nature. This stand can be categorically stated as – ‘Mental phenomena is not an existence over and above physical phenomena, that they are physical phenomena’. Thus, the mind-brain identity theory is a form of material monism. At the same time, owing to its slight distinct notion of what is physical as juxtaposed with what is material the theory may be taken as an instantiation of physicalism. Physicalism claims that human beings are fully material entities whose workings and properties may be completely explicated by the concepts and theories drawn from physical sciences. There is no room for immaterial or ghost entities or non-physical entities and concepts in the physical causal chains that runs through a human’s central nervous system and connects inputs in the form of physical stimuli to behavioural outputs.

However, the mind-brain identity theorists are not claiming that the talk about mental states gives the same meaning as the talk about brain states. An example may be given here. Suppose that A breaks his legs. He is in pain. Suppose that whenever A is in pain, there is neural firing of c-fibres in his brain. Therefore, pain is identical to c-fibre firing. A statement will be “A is in pain”. This is talking about A’s mental state. Another statement when A breaks his legs will be, “A’s c-fibres are firing”. Obviously the meaning here is different. The claim that pain is identical to c-fibre firing is not like claiming “A bachelor is unmarried”. A bachelor has the same meaning as being unmarried. To call someone bachelor but to deny that he is unmarried is a

contradiction. All talks about bachelor can be translated into talks about unmarried men without any loss of meaning. This is analytical reduction. In case of A breaking his leg, it may be observed that it is not contradictory to claim that A is in pain but c-fibres are not firing, though it may be false. 'Pain' does not necessarily mean 'firing of c-fibres'. Thus the mind-brain identity theorists claim that they are not asserting an analytical theory. It may be objected that if the meaning of the talk about mental states is not identical with the meaning of the talk about brain states then mental states cannot be identical to brain states. The mind-brain identity theorists may counter argue that just because the meaning is different doesn't mean the reference of those talks is different too, for instance the case of Morning Star and Evening Star. Here talk about Morning star has different meaning from the talk about Evening star. But ultimately both refer to the same thing, that is, Morning star is identical to Evening star, as they both refer to Venus.

Morning Star = Venus = Evening Star

Thus, the meaning of an expression may be distinguished from its reference. According to Frege, meaning of a word, phrase or sentence, basically a designator, is to be found in their sense<sup>124</sup> and not just reference. Similarly for the mind-brain identity theorists, reality is describable in two vocabularies but there is only one reality, that of the physical. But a counter argument is that when talking of mental events it is precisely the case that even the reference is different from the reference in talks about brain events. For instance, in the above case of A's broken legs, when A says that he is in pain he is not referring to the c-fibres firing in his brain. He is precisely referring to his mental state and not his brain state.

The type-type identity theory claims that each type of mental state will be identical with a given type of brain state. So now pain is no longer identified with a particular physical event but the type of mental state such as pain is identified with a particular type of physical event. It is the case that mind-brain identity theory rejects analytical reductions of mental to physical. So does type-type identity theory. But the type-type identity theory argues that events that fall under a given type of mental description must also fall under the same type of physical description. This

---

<sup>124</sup> For more on this see Frege's "On Sense and Meaning" in *Collected Papers on Mathematics, Logic and Philosophy*, ed. Brian McGuinness, New York: Basil Blackwell, 1984, pp. 157-77; for Kripke's discussion on it see Kripke, Saul A. *Naming and Necessity*. Cambridge, Massachusetts: Harvard University Press, 1980, p. 58-59; also see Klement, Kevin C., "The Theory of *Sinn* and *Bedeutung*", *Frege and the Logic of Sense and Reference*, London: Routledge, 2002, pp. 8-14.

means that the meaning might be different but the underlying mental and physical concepts will exactly coincide. Thus if M is a mental state identical with the brain state B, then M will be obtained only if B is obtained and B will be obtained only if M is obtained. This leads to the claim that mental properties too are reducible in the same way to physical properties. The properties of a given type of mental event will be reducible to the properties of the type of brain event. One may then claim that the painfulness of pain is identical with the behaviour of certain sorts of neurons in the central nervous system.

The token-token identity theory insists that each type of mental state will be identical with a given type of physical state that is the physical arrangement. The claim is, for instance, that pain might be actualized in a given brain state in humans, but it may also be actualized in a different physical state in different organisms like animals or even a Martian. Here then the pain is not identical to the c-fibres firing or even the neuronal behaviour of particular type. The pain here is identical to a particular state of the physical brain. Now in insects for example, there might be no such thing as brain but then they might actualize the same mental state, that is, pain in a physical state similar in arrangement to the physical state of the brain in humans but through a different medium not similar to the brain. For instance in the case of A's broken legs, pain is identical to c-fibres firing but in another case, say of B's broken legs, it might be that d-fibres are firing. The token of pain in A is identical to a token of the type of brain process involving c-fibres. The token of pain in B is identical to the token of the type of brain process involving d-fibres.

The advantage of identity theories over dualism is that it does not have to deal with the mind-body problem<sup>125</sup> as in case of Descartes' dualism. The mental are directly identical with physical and thereby can be explained in terms of physical causation. Also identity theories are comparatively simpler and involve fewer entities than dualism. If one were to apply Occam's Razor then identity theory is preferable. A serious weakness is often pointed out. A characteristic of identity statements is that they are symmetrical. If Morning star is identical with Evening star then one might as well say that Evening star is identical with Morning star. Applying this in the

---

<sup>125</sup> See the introduction of this chapter for mind-body problem. Mind-body problem typically arises when a theory posits duality of human nature – the mind or the mental and the body or the physical. Descartes had posited mind as separate from the body and hence had to account for the relation between them. For Descartes account of mind-body relation one may refer to section 2 of this chapter.



current context it may be then said that if mental states or events are identical with physical states or events then one might as well claim that physical states or events are identical with mental states or events. It then leads to the claim that physical phenomena are explainable in terms of mental phenomena and ultimately everything physical boils down to the mental. So the question posed to the identity theorists is – why is the physical privileged over the mental? Maslin<sup>126</sup> argues that evolution gives the reason for privileging the physical. It is reasonable to assume that when universe was formed there were only physical particles and life is a later phenomenon which developed gradually. Mental is then taken to have evolved only after certain level of physical complexity had been reached. Thus, there is a dependence of mental on the physical and thereby the privilege. Another argument against the above objection is that if mental is identical with physical states especially the brain states then what about those brain states or physical states which have to do with unconscious activities. Such brain states do not give rise to any mental states. Thus, mental states are dependent on physical and not the other way around. It is the case that all mental events or states are identical to some physical events or states but all physical events or states are not identical to mental states or events.

But there are other arguments which the identity theories fail to account for. One argument is the claim that just by observing brain states one won't be able to know or arrive at corresponding mental states. For instance, scientists might look at A's brain activity as much as they want, but it won't tell them what it is like to have the pain of broken legs unless they themselves have legs broken or some part of their body broken. The qualitative feel of an experience is termed *qualia*. Thus, mental is not reducible to the physical. Another problem that the identity theorists are yet to give a coherent and solid explanation for is that of intentionality. Some mental states are intentional, that is, they are *about* something, which means they are directed towards something not-themselves. Therefore, intentional mental states are about or directed towards something external, including cases where that something might not exist. Also intentionality are relative in the sense that one intention might be related to another intention or a group of intentions may together form or give rise to an intention, all in a harmonious way.

---

<sup>126</sup> Maslin, *An Introduction to the Philosophy of Mind*, 2001, p. 86.

Basically intentions are holistic<sup>127</sup>. So how does then the identity theorist account for this added dimension of intentionality to mental states?

An argument against the token-token identity theory is the charge that it does not explain the reason for the identity of a token mental state with a token physical state. For instance, in the above example of A and B feeling pain, no reason can be given as to why in A's case c-fibre fires and not d-fibres like in the case of B. One has to then accept it as a brute fact that a token mental state is identical with a token physical state. It does not explain why a token mental state say 'pain' is identical with a token physical state, say 'c-fibres firing' in case of A and 'd-fibres firing' in case of B. Why A's or B's pain is identical with physical arrangement of their brain rather than not identical with it? Another argument against token-token identity theory is that if a token mental state is identical with a token physical state and yet realizable and thus identical with a different token physical state then it breaches the law of identity. According to law of identity, if M is identical to B then M obtains only if B obtains. Also if M is identical to B and M is also identical to X then B is identical to X. But in case of token-token identity this is not the case. In case of A's broken leg, pain is identical to c-fibres firing. So pain is obtained only when firing of c-fibres is obtained. But in case of B, the pain is not obtained when c-fibre fires but when firing of d-fibres is obtained. Also, if pain is identical to c-fibres and pain is also identical to d-fibres, then c-fibres must be identical to d-fibres. But it is not what token-token identity theorists are claiming. For them c-fibres might not be and therefore is not identical with d-fibres. A token physical state may be sufficient for a token mental state but not necessary. This violates the law of identity. Thus, this poses a serious problem for token-token identity theorists.

Another argument against the identity theory is given by Kripke. The Cartesian argument is usually that since mind could exist without the body, therefore, mind is distinct from the body. Descartes of course privileged mind over body but the Cartesian argument might have as well said that since body can exist without the mind<sup>128</sup>, therefore, body is distinct from the mind<sup>129</sup>. Kripke asserts that if this Cartesian conclusion that mind is distinct from body is to be rejected, then one must also reject the premise of it. It cannot be that one holds the premise and yet rejects the conclusion. Thus, against the type-type identity theory argument is as mentioned earlier,

---

<sup>127</sup> Ibid., p. 28.

<sup>128</sup> Such bodies do exist and are called corpse.

<sup>129</sup> Kripke, Saul A., *Naming and Necessity*, Cambridge, Massachusetts: Harvard University Press, 1980, p. 144.

there may be pain and yet no c-fibres firing (d-fibres are firing instead) and also there may be a case where c-fibres are firing but there is no pain (anaesthesia or may be only d-fibres firing gives pain in a particular individual who has both c-fibres and d-fibres).

Kripke makes the distinction between rigid and non-rigid designators. Non-rigid designators are those words or names (that are descriptions) which does not designate an object in all possible worlds. To designate in all possible world means to designate necessarily. Rigid designators refer to objects necessarily. Kripke gives the example of Lewis Carol as the rigid designator of an individual in all possible worlds. Similarly Charles Dodgson, necessarily refers to that same individual. Thus, Lewis Carol = Charles Dodgson. 'The author of Alice in Wonderland' is not a rigid designator. Thus, the identity statement of Lewis Carol = the author of Alice in Wonderland is not true in all worlds. Kripke applies this even to the names of natural kinds such as gold, water and so on. Water is a rigid designator of a substance. H<sub>2</sub>O is a rigid designator of the same substance. Thus, Water = H<sub>2</sub>O in all possible worlds. Suppose in a possible world, perhaps a twin planet, there is a substance of XYZ composition which has the same qualities as water, such as quenches thirst, is a colourless liquid, has a particular degree of viscosity and so on. Will XYZ be water then? Kripke says that in this case XYZ is not really water but just a water-like stuff. Now the same may be applied in case of mental states like pain. Pain is identical to c-fibres firing according to identity theorists. But is it possible to experience pain in cases where c-fibres are not firing? It is very much possible to do so. The identity theorists will then have to account for it by saying that when one is imagining the pain without c-fibres firing then it is only a pain-like sensation that we are imagining. It is pain-like precisely because its constitution is not the right one just like in case of water-like stuff composed of XYZ. But Kripke here makes an essential point, precisely that the cases of water, gold and so on are radically different from cases of pain. The only identifiable property of pain is that it is pain-like. So something pain-like is pain itself. There cannot be a distinction between pain and pain-like feeling<sup>130</sup>. A person cannot be in a pain-like situation and not feel pain. But we can imagine pain in cases of c-fibres firing and also in cases of d-fibres firing or in cases where both of them are not firing. Thus, pain is not identical with brain states. Therefore, mental states are not

---

<sup>130</sup> For Kripke's objection see, Kripke, 1980; also see Maslin, 2001, p. 99, 100.

necessarily identical with brain states because we can genuinely imagine mental states without the corresponding brain states.

Materialism is the claim that this world, the actual world, is physical and the physical nature exhausts all its nature. The mental nature is then also a physical nature or due to the physical nature. Identity theorists are no doubt then some kind of materialists. Yet, they do accept that there are mental states. They only claim of an identity relation between the mental and physical. This in a sense is reduction of mental to physical. But there is a more radical position that completely eliminates or denies the existence of mental states in totality. This is called eliminative materialism<sup>131</sup>. Some proponents of such a view are W. V. O Quine, Paul Feyerabend, Patricia and Paul Churchland and Richard Rorty.<sup>132</sup> According to this kind of materialist theory, the concept of the mental is owing to our vocabulary. What we are talking about when we talk of mental states are in reality only brain states. It is only due to our vocabulary conventions and the conventions of expression that the mental seems to exist. Thus, according to eliminative materialists we need to remove the very concept of the mental and the vocabulary which supports it from our thoughts and speech. Consciousness thus does not exist; it is all just brain processes. Maslin feels that such a claim is absurd. One is more certain of one's consciousness and one's mental life than the truth of any theory that denies the undeniable facts of one's personal experience<sup>133</sup>.

“Behaviourism is the doctrine that mental states are behavioural dispositions”<sup>134</sup>. A behaviourist need not necessarily be a physicalist. But the version of behaviourism that is most plausible holds that mental states are dispositions of purely physical bodies – our bodies. The principal motivation for behaviourism came from physicalism. Behaviourism denies any of the mental stuff. Physicalism is the doctrine that physical nature fully determines psychological nature. Behaviourism is that part of physicalism which argues that psychological features are just behavioural dispositions of certain physical bodies. Sometimes behaviourism is described as the doctrine that there are no mental states, though mental language is fine because it does not refer

---

<sup>131</sup> Maslin, 2001, pp. 100-2.

<sup>132</sup> However, their theories may further be classified as eliminative behaviourism, analytical behaviourism and so on. They are more or less some or the other form of eliminative materialism.

<sup>133</sup> Ibid., p. 102.

<sup>134</sup> Braddon-Mitchel, David and Frank Jackson, *Philosophy of Mind and Cognition*, UK: Blackwell Publishing, 1996, p. 29.

to such non-existent mental states but instead refers to tendencies to behave. Other times behaviourism is described as the doctrine that there are mental states but they are just tendencies to behave. In both cases it is agreed that a) mental states are not inner, categorical states of persons, and b) what makes psychological claims true are subject's behavioural dispositions.

Analytical behaviourism states that statements about mental states turn out to be equivalent to statements that describe a person's actual and potential public behaviour<sup>135</sup>. The advantage of this doctrine over Cartesian doctrine or doctrine of dualism is that it does not have to deal with the mind body problem. All talks on mental are ultimately just about behaviours. The problem of mind-body is purely scientific and is reduced to just patterns of behaviour. There are physical causes (stimuli) that are responsible for a subject's behaviour. Also the problem of other minds is eliminated<sup>136</sup> here as one can get access to or know the other mind based on the other's behaviour. In case of identity theory it was seen that mental states are just brain states but the mental does exist. Also talks about mental has a meaning distinct from talks about brain states, the reference is the same though. The mental facts are not over and above physical facts. The identity they were talking of was not an identity of analytical kind. In contrast, analytical behaviourism is the theory that states statements about mental states can be translated to statements of behaviour or possible behaviours without any loss of meaning. The statement "A is in pain" may be translated to "A shouts 'ouch'". So it may be asserted that if A says 'ouch' or anyone says 'ouch' (the other mind is not a problem for the analytical behaviourists), then it means s/he is in pain. It is not necessary that it is only 'shouting ouch' that characterises pain. One may include various possible behaviours such as wincing, crying that s/he is in pain and so on.

However, an objection here might be made. Maslin considers two statements<sup>137</sup> – a) Martin raised his arm and b) Martin's arm went up. Here statement a) entails statement b), but statement b) does not entail statement a). If Martin raises his arms, his arms goes up, which is what statement b) means. But if his arms went up, then it does not necessarily follow that Martin raised it. It may be the case that someone else raised his arm and that's why it went up. In the

---

<sup>135</sup> For a very precise formulation of analytical behaviourism Foster, John, *The Immaterial Self: A Defence of the Cartesian Conception of the Mind*, London: Routledge, 1991, pp. 33-46.

<sup>136</sup> The problem of other minds no longer arises.

<sup>137</sup> Maslin, 2001, p. 111.

first case there was a mental state, that is it was Martin's intention to raise his arm, but in the second case there was no such intention and therefore no such mental state. So if only the behaviour were to be considered for Martin's mental state then in both cases we will have to attribute the same mental state which is false. Another problem with such position is that there can be almost any number of behavioural descriptions for a particular mental description and conversely there may be a number of mental descriptions that has similar behavioural description. Also the case of an individual who intends to hide his mental state by behaving differently than usual may be considered. For instance in case of A breaking his leg, he winces and sheds tears, which is his usual behaviour when he is in pain. Now if he intends to hide his pain, then he won't wince or shed tears. A behaviourist may claim that the fact that he does not wince or cry is a behavioural description of his intention to hide his feelings within the context of breaking his leg which might be described in behavioural terms: "he cannot walk and his leg is dangling at odd angles". But then given the same context, it may also be the case that he has been given anaesthesia and he actually does not feel pain and therefore the absence of the behavioural description. It kind of becomes difficult for behaviourists to account for such problems without falling into some kind of circularity. Also to verify ones mental state based on behavioural descriptions becomes difficult.

Behaviourism denies causal role to mental states due to hostility towards unobservable events. Since mental states cannot be observed, which denies them of any empirical verification in principle, they cannot be a part of the scientific inquiry. Behaviourists were influenced by *verificationism*<sup>138</sup>. Braddon-Mitchell and Jackson explains verificationism as "the doctrine that the meaning of a sentence is the method you should use to determine if it is true."<sup>139</sup> Behaviourists concluded that sentences about mental states could only mean something about behaviours, since it is by other's behaviours, which are observable, that we verify claims about their mental states. But even in the domain of science, scientists too go beyond what is directly observable. Electrons, the planet Pluto, black holes, none of these are observable but they are much included in science.

---

<sup>138</sup> Braddon-Mitchell and Jackson, 1996, p. 35.

<sup>139</sup> *Ibid.*, p. 35.

Another form of behaviourism is that given by Ryle known as soft behaviourism. It doesn't appeal to verification like in analytical behaviourism. Also Ryle did not attempt to reduce psychological to physics<sup>140</sup>. He felt that Cartesian dualism is one big mistake, a mistake in principle at that<sup>141</sup>. He gives an analogy. A person visits a university. He is shown the different academic buildings, the libraries, the hostels and so on. At the end of the tour, he says that he has seen all these but where is the university. Here Ryle points out that the mistake committed by the person is supposing that a university is an individual building just like the library or other academic buildings. He doesn't realize that university is the totality of all that he has seen. Similarly in case of Descartes' claim that there is a non-physical mind existing over and above physical body, he seems to have committed a mistake. According to Ryle mind is nothing but the totality of the overt public physical behaviours of individuals. The behaviours are not clues to the working of some ghost-like entity called mind hidden from public but rather the behaviours are those workings themselves. The mental processes are manifest in the behaviours that are open to public view and introspection. To counter some problem mentioned earlier against behaviourism, Ryle invokes the notion of disposition. This however merely means that one has to include in behavioural analysis not only actual behaviour but potential behaviour too. A hides his feeling of pain, so for A to have pain does not mean he has to actually show overt behaviour or wincing and crying. It is enough to claim that A has a disposition to wince and cry. It is objected against this position that it is usually the case that one explains one's behaviour by appealing to one's internal states. If Ryle's position were to be accepted then one will have to explain one's instance of behaviour in terms of another instance of behaviour. Behaviour itself cannot serve as the underlying cause of the surface behaviour.

All the varieties of behaviourism face some common difficulties. In behaviourism third person perspective gains more ground than first person perspective<sup>142</sup>. But the idea that someone else will have better access to one's mental states than that individual herself sounds absurd. Another objection is that one has direct access to his/her mental states. Thus, one can actually

---

<sup>140</sup> Ryle says, "... contrast between mind and matter will be dissipated, but dissipated not by either of the equally hallowed absorptions of Mind by Matter or of Matter by Mind, but in a quite different way." See Ryle, Gilbert, *The Concept of Mind*, London: Hutchinson, 1949, p. 23; quoted in Maslin, 2001, p. 119.

<sup>141</sup> Ryle, 1949, p. 17.

<sup>142</sup> Braddon-Mitchell, 1996, p. 33.

introspect them<sup>143</sup>. A further objection to analytical behaviourism is that it has never been able to give a plausible analysis of a single psychological state in terms of behaviour and dispositions to behaviour.

Functionalism rejects mind as an entity. Mind is reduced to a function<sup>144</sup>. Some the proponents of functionalism are Hilary Putnam and David Armstrong. A function of a thing is the job or task that it performs<sup>145</sup>. The function of a thermostat is to regulate temperature be it a room, a building, or even of water in a tank. It takes certain input from the surrounding environment, namely the temperature of the surrounding and switches off or on the heating system to regulate the temperature to what is desired. The outputs here are – 1) the heating system is turned off, and 2) the heating system is turned on. An abstract notion of a function may be understood through the idea of a mathematical function. Functions are not numbers but things done to numbers<sup>146</sup>. Addition is a basic arithmetical function done to numbers. It has inputs and outputs. For instance the addition of 3 and 5 gives us 8. The numerals 3 and 5 are then the inputs to the addition function and the number we arrive at, that is 8, is the output. This addition sum is represented as:  $3 + 5 = 8$ , where ‘+’ represents the addition function. Now 3 and 5 may be substituted with any number and the addition function will arrive at a unique number as the output. The whole process may be represented as  $x + y = z$ . Here  $x$  and  $y$  are variables. The inputs may be called the *arguments* of the function and the output is the *value* of the function<sup>147</sup>. Similar to addition function there are subtraction function, division function and so on.

A function may be in the form of ordinary expression. For instance, the expression ‘the natural father of  $x$ ’ may be taken to be a function. The arguments that this function accepts then will be people and the value it gives is the father of those people. The idea of a function is a very general one. It can be specified abstractly and therefore is independent of whatever it is that enables a particular function to be discharged. One may wonder as to which category does a function belong to – is it physical or non-physical. A function is also not identical with the physical arrangements that embody them. But they do require some kind of embodiment to be

---

<sup>143</sup> Ibid., p. 34.

<sup>144</sup> Maslin, 2001, p. 130.

<sup>145</sup> Ibid., p. 131.

<sup>146</sup> Crane, Tim, *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*, first published in 1995, second edition, London: Routledge, 2003, p. 85.

<sup>147</sup> Ibid., p. 86.



actualized, that is, if they are to be performed. So then a functional analysis of mind may be given in the same manner. In case of addition function, there were inputs and outputs. In case of a mental state, say, pain, there is an input and output too. The input is the tissue damage and the output is wincing and crying and may be also the desire to get rid of the pain. Taking up again the case of A breaking his leg, the input is the tissue and bone damage in form of broken leg, the output is wincing and crying and a desire to get rid of it. Pain is then the function which when given some input (physical stimuli) will give a particular output in terms of behaviour. The assertion is that mental states, just as described above, are nothing but functions whose job is to cause some kind of behaviour (output) when some kind of stimuli (input) is provided. Here it may be observed that the talk of the mental is neither completely denied nor exactly reduced to physical behaviour (like in case of behaviourism). There is some kind of reduction in the sense that the mental states are considered to be functions of brain states. The talks about mental states are reduced to talks of input-output structures<sup>148</sup>. But it is not identical to brain states and can be talked of distinctly from brain states. Mind is then but a function of the brain. Mental states are then characterized in terms of inputs, outputs and relation to other mental states. Also as pointed earlier due to the abstractness of function and its relative independence from any particular physical arrangement, it may be realized in multiple ways. This means that a given function may be instantiated in different physical systems or arrangements in different cases. This is known as *multiple realizability*.

Also thinking of a mental state in terms of causes and effects fits well with the common-sense understanding of mind. Commonsense functionalism then is the claim that the roles that matter for having a mind and matter for being in one or another mental state is given by what is common knowledge about mental states. The functional roles are extracted from the common knowledge about pain, beliefs and so on. Commonsense functionalism is also known as analytical functionalism. The mind-body problem also disappears in the functional analysis of mind.

However, there are still some aspects of the mind which functionalism fails to account for<sup>149</sup>. One such aspect is the subjective experience aspect of mental states. Consider, for

---

<sup>148</sup> Maslin, 2001, p. 143.

<sup>149</sup> Ibid., p. 149.

instance, that A is drinking wine with B. There is a subjective experience of what it feels like to drink wine in case of A and similarly so for B, but both being subjective experience is different. The subjective experience of what it feels like is known as qualia. It is this qualia that the functionalism fails to account for. The qualia is absent in the functional mental states. This is demonstrated by Ned Block in his 'Chinese Mind Argument'<sup>150</sup>. Suppose that each of the billion inhabitants of china were to play the role of a neuron in the brain. They are each provided with a two way radios which connect them with each other and allows for communication. The current mental state of the China brain is then displayed on satellites which can be seen from all over China. The whole setup, which is the China brain, is then connected by a radio to an artificial body. This body may be similar to a human body which then provides the sensory inputs and behavioural outputs. This setup satisfies all the descriptions of the functional analysis of mind. Such a system could be functionally equivalent to a human being. But can it be said that such a mind would then enjoy the subjective experience as a human being does. As can be seen in case of China brain, there is no such thing as qualia. One may refer to the qualia of each Chinese individual, but that is irrelevant here. The qualia to be experienced by the whole brain are missing. There is a brain of course and thereby brain states, but where is the mind and thereby the mental states. The subjective mental experience is absent. The point that Block makes is that there is nothing like a Chinese mind in case of the Chinese brain. What the functionalist refers to as Chinese mind here is no mind at all.

Coming back to the example of A and B drinking wine, suppose that there are two bottles of wine, X and Y. Both bottles are of the same design, same labelling and both contain red wine. But suppose again that A has the distinct ability to differentiate even the slightest of difference in the shade of the colour red. This ability may be due to some unique neural difference between A and B. Also the wine in X is red of a different shade from the redness of the wine in Y. Let these two different red be referred to as red1 and red2. Now when the wine from each bottle is poured into two different glasses, A can easily tell which glass of wine came from X and which one came from Y. This is owing to his ability to see and match the shade of redness of the wine in the glass to the shade of redness of the wine in the bottle. He matches the red1 in the glass with red1 in the bottle. Now the same challenge is posed to B. But she does not have this ability herself.

---

<sup>150</sup> Also known as China Brain argument given by Ned Block in "Troubles with Functionalism", in Ned Block (ed.) *Readings in the Philosophy of Psychology*, vol. 1, London: Methuen, 1980.

For B the red in both bottles look the same, perhaps only as red1. B might be able to access all the neural activities of A's brain and combine it with the inputs and outputs of his brain processes. But having analysed whole of A's brain's functioning; it still won't help her distinguish the R1 from R2. The point of this is that B may know everything there is to know physically about what goes in A's case, but A still knows something that B doesn't know, that is, R2. Such an objection to functionalism is made by Frank Jackson<sup>151</sup>. Another example given by Jackson is of Mary<sup>152</sup>, who is a brilliant scientist specializing in neurophysiology of vision. She is forced to investigate the world from a black and white room via a black and white television monitor. Suppose that she acquires all the physical information that perception involves. She discovers which wavelength combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that result in the uttering of the sentence "The sky is blue". What will happen when Mary is released from the black and white room? Will she learn anything or not? She looks at the sky and can know what the sky looks like in terms of colour. Can it be said that Mary knows something more and unique than what she knew earlier? The answer will probably be yes. Thus, functionalism fails to account for such cases.

Some other aspects of the mental such as intentionality, consciousness and so on are not accounted for by functionalism. One such argument against the functionalist theory of mind is given by John Searle and is known as the Chinese room argument. But this will be dealt in the following chapter.

#### **4. Computationalist Theory of Mind**

As can be seen from earlier discussions, Turing's thesis paved the way for a computational understanding of mind. Of course it was not in isolation from other development in diverse domains that he came up with his computational model of a universal machine which could compute what a human computer could. Already it may be observed that in his imitation game, all that the machine has to do is behave (in the form of responses through the teleprinter) in a way that will make the interrogator believe that the conversation he is having is with a

---

<sup>151</sup> Jackson, Frank, "Epiphenomenal Qualia", *The Philosophical Quarterly* (1950-), vol. 32, no. 127, 1982, pp. 128-30.

<sup>152</sup> *Ibid.*, p. 130.

human. Also the machine that was proposed by Turing was all out physical. So a parallel development of physicalist theories was taking place at the same time. In the imitation game, corresponding to the machine's aim of making the interrogator think that it is a human being, there are the functions whose task is to take in the questions or inputs that the interrogator provides and give a suitable response (output) that is suitable for its goal. Functionalism asserts that mind is a function of brain. Also a function's characteristic is that it can be realized in multiple systems or arrangements (for details and example see section 3 of this chapter). So functionalism asserts that mind as a function or mental states as functional states are also multiply realizable. A human computer can compute functions. Now according to Turing's claim a machine or a digital computer can compute whatever a human computer can compute. Thus, a computer (in a non-human mechanical or digital sense) can compute functions. The whole purpose of such computers (based on Turing's model of a computing machine) is to compute functions. Now if mind is a function then the claim is that a computing machine will compute such a function too.

A method for computing<sup>153</sup> a function is known as an *algorithm*. Algorithms are also known as 'effective procedures', effective in the sense that the procedures are effective in bringing about the results. A function may have more than one algorithm for finding its values for any given inputs<sup>154</sup>. For instance when we multiply 22 by 13 we follow a procedure described as follows –

1. Add 22 to itself the number of times equivalent to the number that is the first digit from the right side of the given number which is 13. We obtain 66 by doing so.
2. Add 22 to itself the number of times equivalent to the number that is the second digit from the right side of the given number which is 13. We obtain 22.
3. Add a 0 at the end of the result obtained in step 2. We obtain 220
4. Add the result from step 3 with the result of step 1. We obtain 286.

---

<sup>153</sup> Computing here means calculating the value of a function.

<sup>154</sup> Crane, 2003, p. 88.

This same function of multiplying 22 by 13 may be done by adding 22 to itself 13 times. There may be more ways of doing multiplication. Thus, a function may have different algorithms. In mathematics, to say that an arithmetical function has an algorithm is not to say that it will always give a number as an output. What an algorithm for a function means is that it always gives you a procedure for finding out whether there is an answer or not. When there is an answer, that is, when an algorithm gives the value of a function for any input then the function is taken to be computable. The notion of an algorithm is extremely general. An algorithm has to satisfy the following conditions –

1. At each stage of the procedure there is a definite thing to do next. There is no special guesswork, insight or inspiration or feat of imagination at play while moving from a step to the next.
2. The procedure is specified in a finite number of steps.

Crane demonstrates an algorithm for multiplying two whole number,  $x$  and  $y$ , which works by adding<sup>155</sup>  $y$  to itself. The procedure is performed on three pieces of paper, one piece of paper, suppose it is called  $X$ , for the first number, second one called  $Y$  for the second number and the third piece called  $Z$  for the answer. The following steps may be given:

- Step (i): Write '0' on  $Z$ , and go to step (ii).
- Step (ii): Does number written on  $X = 0$ ?
- a. If YES, then go to step (v)
  - b. If NO, then go to step (iii)
- Step (iii): Subtract 1 from the number written on  $X$ , write the result on  $X$ , and go to step (iv)
- Step (iv): Add the number written on  $Y$  to the number on  $Z$ , and go to step (ii)
- Step (v): Stop

Now one may apply this calculation to 4 times 5. So one begins by writing 4 on  $X$  and 5 on  $Y$ . Step (i) is applied and 0 is written on  $Z$ . Next, following step (ii) it is asked if the number written

---

<sup>155</sup> Addition is itself is a function and thereby, an algorithm may be given for it. It is not a problem that such a function is used within another function, as long as the former can be given specified as an effective procedure.

on X is 0. There is 4 written on X, so one moves to step (iii). In this step 1 is subtracted from the number written on X, which gives 3, then 3 is written on X and one moves to step (iv). The number written on Y is added to the number written on Z, which gives the answer 5 and then one goes to step (ii) and asks again if the number on X is 0. It is 3. So one goes to step (iii), subtracts 1, the result 2 is written on X and then moves to step (iv). The number on Y is then added to the number in Z, which is 5 this time. So 10 is obtained and one moves on to step (ii). Again the question is asked if the number on X is 0. It is 2 this time. So one goes to step (iii), subtracts 1, writes the result 1 on X and moves to step (iv). The number on Y is again added to the number on Z which is 10 this time. The result then is 15 and then step (ii) is again applied. The number written on X is 1, that is not 0. One goes to step (iii), 1 is subtracted giving the result 0, written on X and then step (iv) is reached. The number written on Y is again added to Z which is 15, and the result is 20. One moves to step (ii) again. This the time number on X is 0. So the next move is to go to step (v). The procedure stops. The number written on Z is 20, which is the result of multiplying 4 by 5.

Similarly, algorithms can be given for all sorts of things. For instance an algorithm for cooking a bowl of Maggi may be given as:

1. Turn on the stove.
2. Fill the pan with 210 ml water.
3. Place the pan on the stove.
4. Break the noodle cake into four parts.
5. When water boils, put the noodle and tastemaker into the water.
6. Set the timer for 2 minutes.
7. When the timer rings, turn off the stove.
8. Serve the noodle in a bowl.
9. Result: one bowl Maggi noodle.

Such an algorithm may then be given to a machine and it may, at the end of the procedure, give a bowl of Maggi noodle. Now cooking may be considered as unique to human being as intelligence, may be even more unique, at least among the living species. It may be that cooking itself requires intelligence. If in the imitation game the questions were replaced by requests for cooking dishes and the responses were given by serving the cooked dishes then, on an

interrogator's request, a machine might as well serve the appropriate dishes. Based on such successful behaviour the interrogator may judge the entity behind all the dishes as human.

Computationalism asserts that mental states are computational state. According to Turing's thesis, functions are computable by a Turing machine. A Turing machine's algorithm is given by its machine table. According to computationalism, then, mind is an algorithm that is computable, and since it is computable it may be precisely computable by a Turing machine. Mind is then compared with an enormously complex machine table. Now, human beings, brain more specifically, can be compared with a Turing machine. The inputs or arguments are provided by an individual's surrounding through his/her sensory faculty and then based on the algorithm provided, that is the machine table which is a list of mental states and as a whole is then mind, computes and gives the output as specific behaviours. Computationalism is then a form of functionalism. Functionalism argued for the multiple-realizability of mental states. Computationalism is thus that form of functionalism that claims that mind is a function that is computable and thereby, can be multiply-realized even in a machine.

It is claimed that Descartes anticipates several developments in twentieth century AI<sup>156</sup>. He draws from the obvious difference between humans and animals in their capacity for language and asserts that language is what defines the line between mere behaviour and genuine cognition. Language capable AI was one of the early attempts at developing human level AI. Descartes also gives an explanation that to think is to reason which is but manipulation of the symbols of language in a way that is meaningful. For Descartes, a machine wasn't capable of doing this and therefore a machine could not think.

Walmsley describes Thomas Hobbes as the "Grandfather of AI"<sup>157</sup>. He asserts that thinking, or rather 'exact thinking', can be considered in the same line as mathematical operations of addition, subtraction and so on. He claims that in matter, where there is place for addition and subtraction, there is also place for reason. Where there is no place for these then there is no place for reason either. Like Descartes, Hobbes too rests reasoning, as computation,

---

<sup>156</sup> Walmsley, Joel, *Mind and Machine*, Hampshire, England: Palgrave Macmillan, 2012, p. 6.

<sup>157</sup> Walmsley, 2012, p. 7; the reason given for this by Walmsley is that Hobbes, in his (1655) *Elements of Philosophy*, declares that by ratiocination he means computation and goes on to explain that thinking or reasoning can be considered along the same lines as the standard mathematical operations such as addition and subtraction. Also Hobbes is described so by Haugeland, see Haugeland, John, *Artificial Intelligence: The Very Idea*, Cambridge, Massachusetts: MIT Press, 1985.

on language. Language bridges the gap between our thoughts and our surrounding as linguistic items can be both internal and personal<sup>158</sup> to us and at the same time is external and public.

In chapter 1, the advent and development of AI has been traced out with discussions on some important concepts of AI. In chapter 2, various philosophical concepts and theories of mind have been dealt with briefly. It may be observed from the discussions carried out in these chapters that there are certain philosophical antecedents to AI. As such AI may be considered as presenting some theory on mind. As presenting a theory on mind and as a new perspective, which was not there in the earlier discussions (from ancient discourses to physicalism), problem and themes from the discourse on mind are then equally presentable in this fresh perspective of AI. AI “has” to work on the basis of some theory of mind. Like any of the sciences, it pre-assumes some fundamental concepts. Computationalism, then, is the theory that serves as the foundation of AI, at least from a philosophical perspective.

But how does computationalism tell us anything about mind? As far as computationalism is a form of functionalism, it considers mind as a function. A function can be computed, which means an algorithm can be given for it. Therefore, computationalism works with the assumption that an algorithm for mind can be given which when uploaded on a machine or device of a certain kind (Turing machine kind) can instantiate a mind. One such device is the digital computer. In section 1 of chapter 1, a human computer was described. It is a human being who computes. Similarly, a digital computer is a digital machine that computes. But what is essential to being a computer? Crane gives a rough definition of computer as *a device which processes representations in a systematic way*<sup>159</sup>. What is essential to being a computer is processing representations in a systematic way. Two questions may arise – what is representation? And what is systematic? A very general and obvious idea to the first question will be to say that a representation is something that represents something<sup>160</sup>. It may represent something else and even represent itself. Representation does provide a philosophical problem. For instance one may ask – how does a representation manage to represent anything at all? But for the purpose of this

---

<sup>158</sup> Personal in the sense of “first-person” or phenomenal in so far as linguistic terms often refer to or represent our thoughts to others. See Walmsley, 2012, p. 8.

<sup>159</sup> Crane, *The Mechanical Mind*, 2003, p. 85.

<sup>160</sup> *Ibid.*, p. 11.



thesis, representation will be taken at face value. In fact given that computation is processing representations, computationalism takes representation as granted. To answer the second question it may be said that to process in a systematic way is to follow effective procedures<sup>161</sup>. Effective procedures or procedures in general, can be understood as rules to be followed. It may be said that a computer is a device that processes representations according to rules. Can such a machine think? If the answer is yes, then AI research will not only borrow from research on mind but AI research itself will provide insights into the nature of mind. But if the answer is no, then AI for all its technological promises, will still be far away, in principle, from making machines think or to provide any substantial insight into the nature of mind. In this thesis, two such objections to AI will be taken up. The objections provided are as given by John Searle in his Chinese Room Argument and Roger Penrose's rejection of any of the current sciences being able to account for mind. These objections are taken up in the next chapter.

---

<sup>161</sup> As given in section 4 of this chapter.

## Chapter 3

### Philosophical Debates on AI

Earlier we had discussed how functionalism asserts that mind is multiply realizable. Computationalism also shares this assertion since a particular algorithm may be realized in any number of machines that are based on the theoretical or mathematical model of a possible machine as given by Alan Turing and known as Turing machine<sup>162</sup>. In addition, computationalism claims that mind is not only multiply realizable but also realizable in sorts of systems whose composition lacks any biological qualification. Digital computer is one such system. This leads to an analogy between mind-brain relation and software-hardware relation. As a consequence, it is claimed by computationalists that thinking is possible in computers generally, and more specifically in digital computers. It is then possible for computers to have an artificial mind of some sort. Computers then can have mental states and thereby, they may be said to have thoughts, beliefs, desire, hope, feelings and so on. This view is termed Strong AI by John Searle. Mind is to the brain as the program is to the computer hardware<sup>163</sup>. And it is this view that Searle criticizes and rejects.

It may, perhaps, be prudent here to give a brief explanation of the two methods in AI known as *top-down* and *bottom-up* approach. An algorithm may be constructed according to some well-defined and well understood fixed procedures<sup>164</sup>. The computer programmers already understand what knowledge is to be provided and what rules are to be specified to make a computer perform a particular task. This is known as the top-down approach. In this approach an overview of the system is considered and based on the requirements of the system, further sub-systems are specified. The algorithm or specific parts of it may be modified as and when the programmer learns more about the task at hand or finds out better rules and knowledge store to improve the computer's performance. Modifying a specific part or sub-system is done with the whole system and what its goal or task in mind. Hence the modification doesn't lead to much change in the structure of the overall system. Whatever rules are specified, they hold strictly for

---

<sup>162</sup> One may refer to chapter 1, section 1 for discussion on Turing-machine.

<sup>163</sup> Searle, John, *Minds, Brains and Science*, Cambridge, Massachusetts: Harvard University Press, 1984, p. 28.

<sup>164</sup> Penrose, Roger *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Oxford: Oxford University Press, 1994, p. 18.

the computer. The top-down approach may be contrasted with the bottom-up approach, where the knowledge store and the rules are not specified in advance. Instead a procedure is given for the system to learn and improve its performance according to its experience. The system here is made up of small and simple units or sub-systems which are in relation to each other. The inter-relations between units then build up to form more and more complex structure, until the system reaches a level where the desired goal is achieved or the desired task is performed. The details of how a computational system learns or improves or experiences is not of much importance here. A brief idea may be given. One such system is that of artificial neural networks. A category of neural network is the Parallel Distributed Processing (PDP) network<sup>165</sup> which consists of layers. Each layer is made up of inter-connected units. In a PDP system a concept is represented by the whole system. For different concepts there can be corresponding state(s) of the same network. Such a network may represent new concepts. All that is needed is to modify the interactions between the units so as to create a new stable pattern of activity. Modifications occur gradually over many separate occasions. The network then learns based on which of the patterns are more stable. The realization of more stable patterns may be understood as improvement over the lesser stable states. A system may also combine both top-down and bottom-up organizations.

## **1. Searle's Thesis and Critique of AI**

In our previous chapter the notion of a machine computer was discussed as given by Turing<sup>166</sup>. Also the definition of computer, following Crane, was given as a device which processes representations in a systematic way<sup>167</sup>. It was said that what is essential to a computer is processing of representations in a systematic way. It was also mentioned what representation and systematic means. Representation is that which represents. In case of computers this is done through abstract symbols, sequences of ones and zeroes. Systematic operation means to follow some procedures (effective procedures). Procedures are step-by-step instructions given to the computer to follow. These instructions then serve as rules which direct the computer. So then, a computer operates purely through symbol manipulation, where the manipulation is based on or done according to rules. According to strong AI then, this thing, which operates purely through symbol manipulation based on rules, can think. To put it in another way, this thing that has only

---

<sup>165</sup> For more on PDP network see chapter 1, section 6.

<sup>166</sup> See chapter 1, section 1 for the notion of machine computer as opposed to a human computer.

<sup>167</sup> See chapter 2, section 4.

syntax can think. Abstract symbols that have no semantic content, that is, no meaning attached to it and are manipulated according to some rules, can be considered as mere syntax. Searle says that a computer is defined solely in terms of its syntactical operations. As such it lacks semantics. Mind has syntax, but it also has semantics, whereby it is meaningful. Therefore, computers which only have syntax cannot have mind and thereby cannot think.

Searle illustrates this point through a thought experiment. A computer is provided with a program that enables it to simulate the understanding of Chinese. A question is given to it in Chinese, it matches the question against its database and produces appropriate answers, also in Chinese. If the answers are as good as a native Chinese speaker then according to Turing's test, this computer will have to be thought as having a mind. The interrogators will actually think that they are speaking to a Chinese person. This is then taken as enough to attribute understanding of Chinese to the computer. Searle now substitutes the computer with himself locked in a room. So Searle is in a room and in this room are several baskets of Chinese symbols. Searle does not understand Chinese at all. But there is a rule book in English given to him for manipulating the Chinese symbols. The rules specify the manipulations of the symbol purely in formal terms. So the rule might say, "take 我的名字叫 from basket number one and put it next to 约翰·瑟尔 from basket number two". Now some Chinese symbols are passed into the room and following the instructions from the rule book, Searle passes back Chinese symbols out of the room. Now again suppose that unknown to Searle, the symbols passed into the room are called 'questions' and the symbols passed out of the room are called 'answers to the questions'. So just by manipulating the symbols according to the rule book, Searle is said to have a conversation with the outsider interrogators. He is taken to be answering Chinese questions in Chinese language. Thus, Searle is considered as being able to understand Chinese. But this is absurd as it is known that Searle does not understand Chinese. The attribution of understanding Chinese to Searle in this case is simply misplaced.

From the above example, it becomes clear that simply symbol manipulation is not enough for someone to understand. If someone understands something, s/he does so in terms of both syntax and semantics. This may be explained. Searle asserts that one's internal mental states by their very definition "have certain sorts of contents"<sup>168</sup>. For instance, if one is thinking of

---

<sup>168</sup> Searle, 1984, p. 31.

drinking water or of going to bed or reading Turing's biography, in each case his/her mental state has a certain mental content in addition to whatever formal features it might have. Here the content may be water, bed or Turing's biography (perhaps a book about Turing's life) and along with them the context in which those contents are present. Suppose in case of drinking water, the context is that there is a water tap nearby and the person is thirsty. The content is that about which the mental states are. Even if thoughts occur in strings of symbols, they have to be about some content. The symbols by themselves, taken as abstracts, do not have any content or meaning. Syntax is just the formal structure or arrangement of the symbols. If the symbols themselves are not semantical, that is, if they do not have any content, then there can be no meaning to those symbols purely based on their formal structure or arrangement which is the syntax. Searle says, "If my thoughts are to be *about* anything, than the strings must have a *meaning* which makes the thoughts about those things. In a word, the mind has more than a syntax, it has a semantics."<sup>169</sup> It may be observed that while Searle-in-the-room knows the rules to manipulate the symbols, he does not know what the semantic content of these symbols are. He does not know what the symbols refer to in the outside world. He has no knowledge of the objects that are being represented by the symbols. Nor does he have any knowledge of the context in which those symbols are being used. As such, due to the absence of semantics, Searle-in-the-room cannot understand the meaning of the symbols being passed to and fro, as well as, the conversation that is going on unknown to him. A computer program, just like Searle-in-the-room, lacks any understanding of Chinese. Thus, a computer program which is specified only in terms of syntax can have no understanding whatsoever. And because it can have no understanding, it cannot be thought as having a mind. In other words, mind has both syntax and semantics. A computer program simply has no semantics. Thus, a computer program, in principle, cannot be either equal to a mind, or be a mind.

The Chinese room argument puts forward the point that to have only syntax is not enough to have a mind, as the mind also has semantics. But can syntax be sufficient to give meaningful content? Can syntax be sufficient for semantics? Searle asserts that it is not. This, he says, is "a conceptual truth"<sup>170</sup>. The proposition "syntax is not sufficient for semantics" in fact expresses the distinction between the notion of what is formal and what has content. It is in the concept of

---

<sup>169</sup> Ibid., p. 31.

<sup>170</sup> Ibid., p. 39.

syntax itself, perhaps as an entailment, that it may be understood as not sufficient for semantics. The formal structure by itself cannot give us the content. So, the claim is that from syntax alone one cannot arrive at semantics.

Searle takes up an objection that may be offered to his Chinese room argument<sup>171</sup>. If the Chinese understanding program is placed inside a robot and the robot could causally interact with the world, it would then be enough to guarantee that it understood Chinese. Searle responds to this that the distinction between syntax and semantics once again creates a problem for such cases. As long as the supposition is that the robot has only a computer as a brain, the Chinese room argument still applies. It could be imagined again that Searle is placed inside the robot's head. It is not known to him that some symbols come in through video cameras attached to the robot's head and some symbols go out to move the body of the robot. As long as he is unaware of the fact that the robot is having causal interactions with the world outside and all he has is a computer program, he would have no idea about the meaning of the symbols. Searle-in-the-robot lacks as much understanding of the causal interactions with the world as Searle-in-the-room lacks understanding of Chinese language. The causal interactions become relevant only if they are represented to some mind or the other. In this case they will be relevant only if they are represented to Searle-in-the-room's mind. This version of the Chinese room argument may be termed as Searle-in-the-robot argument.

Another thesis that Searle puts forward to reject strong AI is that mind is essentially biological. Mind is caused by the biological brain and at the same time is a feature of that same biological system. Searle justifies this with analogies from sciences. He describes the distinction often made in physics between micro and macro properties of a system<sup>172</sup>. For instance he takes up the case of solidity. An object that is solid, say a stone, is made up of micro-particles. The micro-particles can be molecules, atoms or even sub-atomic particles. The micro-properties may then be understood as the features exhibited by molecules or atoms or sub-atomic particles depending on the level at which the micro-particles are being considered. But the object, that is the stone, also has certain properties, solidity being one of them. In case of water and glass the properties may be liquidity and transparency, respectively. The properties being considered are

---

<sup>171</sup> Ibid., pp. 34, 35.

<sup>172</sup> Ibid., pp.20-23.

properties of the physical system considered at the surface level. Solidity, liquidity and other such properties, according to Searle, are higher level or surface features. This higher level or surface level is what Searle means by macro-level. Now, solidity and other properties at the macro-level can be explained causally in terms of micro-level structure or activity. The micro-level is also called the lower level. The solidity, say of a hammer, is caused by the behaviour of particles at the micro-level which constitutes the hammer. But solidity is not only caused by these micro particles, it is also realised in the very system consisting of those very same micro-particles. Similarly in case of mind, the mind is taken to be at a higher level which is caused by micro-level neuronal processes in the brain and at the same time it is also realised in that very system consisting of the neuronal processes, that is, the brain. Mind is then, both caused by and realised in the brain<sup>173</sup>. This, then, leads to the assertion that mind is caused by and also a feature of the brain, brain is biological and thereby mind can only be biological. This argument serves as another objection for Searle against Strong AI or computationalism as it serves as the underlying theory of AI<sup>174</sup>.

A further extension of Searle's macro-level and micro-level explanation of mind and brain may be given. According to Searle the mind-body problem is owing to the confusion regarding the nature of mind. The apparent difficulty with this problem is that any solution to it has to account for four features of mental states. These four features are consciousness, intentionality, subjectivity and mental causation<sup>175</sup>. For the purpose of this dissertation, only intentionality will be briefly taken up. For Searle, intentionality is a biological phenomenon. Intentionality may be understood as "the feature by which our mental states are directed at, or about, or refer to, or are of objects and states of affairs in the world other than themselves"<sup>176</sup>. Just as explained earlier how mind is biological, similarly, intentionality is understood to be biological. A simple argument for now can be given as: Mind is essentially and only biological, consciousness and intentionality is a feature of the mind, and therefore, intentionality can only be biological. Searle takes up the example of thirst. Certain kinds of thirst are caused by nerve

---

<sup>173</sup> Ibid., p. 22.

<sup>174</sup> Computationalism is the thesis that mind is an algorithm that is computable by a Turing machine. See chapter 2, section 4.

<sup>175</sup> Searle, 1984, pp. 15-17.

<sup>176</sup> Searle says regarding intentionality, "Intentionality'... doesn't just refer to intentions, but also to beliefs, desires, hopes... and all of those mental states (whether conscious or unconscious) that refer to, or are about, the world apart from the mind". See Searle, 1984, p. 16.

firings in a particular region of the brain called the hypothalamus. The nerve firings are in turn caused by the action of hormones known as angiotensin in the hypothalamus region. Angiotensin is synthesized by rennin which is secreted by the kidneys. Thirst is then said to be caused by a series of events in the central nervous system, principally the hypothalamus, and is at the same time realized in the hypothalamus. At the same time, to be thirsty is to have the desire to drink. Cases of being thirsty are therefore intentional. This is how intentionality is caused by biological processes and at the same time is realised in biological systems.<sup>177</sup> Searle's argument for intentionality being biological is in terms of similar analogies in science as discussed earlier where the relation between mind and body was juxtaposed with the relation between macro- and micro- level features.

Mind is biological, digital computers are not biological. Thereby, digital computers cannot have mind and therefore cannot think. Note that the opposition here is only against digital computer being able to think. If there was suppose another kind of thing (or brain or whatever that is found inside of say an alien, but not a computer as computers operate only in terms of syntax), may be made up of some different biological stuff or stuffs which has some other chemical and biochemical composition from a human brain, it is, in principle, still possible for it to have a mind. Searle's argument then is based on two things. Firstly, his arguments rest on the definition of computer as only systematic processing of symbols. Secondly his arguments rest on the rejection of something digital ever being capable of having mind on the basis of it being non-biological.

## **2. Criticism of Searle's Thesis**

Searle does not rest at rejection of computationalist assertions. He further goes on to assert that any study/science<sup>178</sup> accepting AI or combining it with some other studies/science is bound to be a failure in understanding mind or is to be discredited. The failure and discrediting is

---

<sup>177</sup> Ibid., p. 24.

<sup>178</sup> Searle's specific target is the field of cognitive science. He says that the need to fill the gap between mind and brain has led to the recent efforts on drawing analogy between human beings and digital computers. Cognitivism is one such recent attempt to fill the gap by deriving from works in cognitive psychology and artificial intelligence, forming the new discipline of cognitive science. However, Searle's criticism of this discipline is not taken up here as one of his main contentions with such an attempt is the association with computationalism or the analogy with digital computers. For more details and his criticism of cognitive science see Searle, 1984, pp. 42-56.



precisely owing to the rejection of computationalism as a theory underlying AI. This is the claim that Boden has an objection with.

Boden in her article “Escaping the Chinese Room”<sup>179</sup> responds to Searle’s arguments. She gives a preliminary objection to Searle’s arguments by criticizing his reason for confining intentionality only to biological processes and systems. It was mentioned earlier that Searle’s argument that intentionality is biological is based on analogies from the sciences. It is this argument from analogy that Boden takes up. The brain’s production of intentionality may be compared to, say, photosynthesis. But Boden asks, if they are really comparable<sup>180</sup>. On one hand we have chlorophyll and photosynthesis, of which we know everything right down to the subatomic processes. On the other hand we have brains and understanding, of which we know very little or at least not as much and not at the same level as its assumed counterpart in science. Our theory of what intentionality is does not bear comparison with our knowledge of say, carbohydrates. What intentionality *is* is still philosophically controversial. Boden also points out that there is very good reason to believe that neuroprotein supports intentionality, but we don’t have any idea how qua neuroprotein, it is able to do so. There is no reason to believe that only neuroproteins supports intentionality by virtue of its neuroprotein-ness or say by being biological. This Boden says is an empirical issue to find out if some other stuffs, like say the stuffs used on a digital computer, can support intentionality and just applying intuition won’t solve the issue<sup>181</sup>. Thus, Boden objects to the analogy given by Searle as being not sufficient to deny intentionality in something other than the biological neuroprotein.

This serves as the preliminary for her rebuttal of Searle’s claim that formal-computational theories cannot explain understanding. Boden does so in two parts. Firstly she takes up the Chinese room argument. In the Chinese room argument, Searle-in-the-room is not able to understand Chinese. As mentioned earlier (in section 1 of this chapter) a version of this argument is Searle-in-the-robot, where Searle is now placed inside the robots head which may be taken as a room in the Chinese room argument. Searle claims that even if this Searle-in-the-room was placed inside a robot’s head, just as computer programs are placed, the robot cannot be credited

---

<sup>179</sup> Boden, Margaret A., “Escaping from the Chinese Room”, *The Philosophy of Artificial Intelligence*, ed. Margaret A. Boden, Walton Street, Oxford: Oxford University Press, 1990, pp. 89-104.

<sup>180</sup> *Ibid.*, p. 92.

<sup>181</sup> *Ibid.*, p. 93.

with understanding of any worldly matters. This reply was in response to the objection against the Chinese room argument. The objection was that perhaps Searle-in-the-room could not understand, but if Searle was put inside the robot and the robot interacts with the world, it may then be taken to have understanding. It does so by being provided along with the Chinese language understanding program (like the rule book in Chinese room argument), also the visual programs and limbs capable of walking, picking things and so on. If the behaviour of such a robot were to be identical with that of a human, it would then be said of the robot that it has understanding. Searle's reply is that it is precisely what he is pointing out. That cognition, understanding and so on is not solely a matter of formal syntax but requires additional semantics or contents from the outside world. The causal interaction between the world and the robot becomes relevant only if it is represented to Searle-in-the-room. To be represented to Searle-in-the-room is to be represented to his mind, as Searle-in-the-room is characterized with a mind same as Searle in reality has a mind. However this representation of the causal interaction is absent in the example. So in the above Searle-in-the-robot example, Searle-in-the-robot is still as clueless about Chinese as Searle-in-the-room. For in both cases, Searle is just manipulating the symbols, even though the robot may seem to be interacting with the world. The symbols presented to Searle-in-the-robot are mere abstract symbols. He has no idea of the wider context. Consequently, it is argued by Searle that the robot cannot be credited with understanding of any worldly matters.

Boden finds Searle's reply to the Robot objection as unacceptable<sup>182</sup>. She objects that Searle's description of the system inside the robot's skull does not truly parallel what computationalists say about the brain. The computationalist or the computational psychologists know that the computer models of the mind are relatively stupid. They become more and more stupid as one moves to increasingly basic theoretical levels. It is claimed by Daniel C. Dennett that in understanding a system that requires pre-positing of a homunculi<sup>183</sup>, one may posit another stupid homunculi to explain the first homunculi and a more stupid homunculi to explain the stupid homunculi and so on<sup>184</sup>. Precisely because the psychologists wish to explain say human language, vision and so on that they posit underlying processes which lack the capacities.

---

<sup>182</sup> Ibid., p. 95.

<sup>183</sup> 'Homunculi' is the plural of homunculus which means a very small human or humanoid. For more on homunculi, see Dennett, Daniel C., *Consciousness Explained*, New York: Little, Brown and Company, 1991, pp. 259-62.

<sup>184</sup> For positing homunculi in a system and how they might be replaced by machines, see Dennett, 1991, pp. 87-91.

To posit underlying processes that has the same capabilities as the one that is being explained in terms of the underlying processes is just circularity and begs the question. In the Searle-in-the-robot example, Searle-in-the-robot is taken to be fully intentional as understanding English. Here, Searle-in-the-robot is taken as the brain or pseudo-brain<sup>185</sup> of the robot. Following Searle's argument that mind is both caused by and realized in the brain, it may be said that the pseudo-brain not only understands English, that is, realizes the phenomena of understanding English, but it also serves as the cause for that understanding. Understanding is both caused by and realized in this pseudo-brain. It may then be further implied for a number of mental features that the pseudo-brain has, for instance, intentionality or intelligence or consciousness which is caused by that very pseudo-brain. This Boden points out is a mistake by Searle for treating brain as the bearer as opposed to the causal basis of mind. To explain the features of mental states by referring to a brain that already has those same features is circularity. So brain cannot have understanding, or intentionality and so on same as the mind that it causes.

Just as one has to posit underlying processes lacking the very capacities that is to be explained, it may be said that the brain as the underlying cause of mind lack the features of mind that is supposed to be explained by its activities. Boden points out that to assert that intentionality cannot be ascribed to brain, and yet to refer to brain or the brain processes as the stupid underlying processes, might seem contradictory. This is so because to be stupid is to be intelligent, but not very. Boden says, "... stupidity is virtually a *species* of intelligence"<sup>186</sup>. Thereby, it seems that Boden is ascribing intelligence even to the underlying processes, that is, the brain. So how can it be that, on one hand, Boden says that the notion of intentionality is not grounded on any biological matter and thereby not grounded on the brain and yet, at the same time she says that the brain is stupid? If brains can be ascribed with stupidity, which is intelligence, it should also be capable of being ascribed intentionality. Boden seems to be contradicting herself. Boden explains that intentionality is in fact ascribed to the brain by computationalists too. But the intentionality ascribed here is not the same "full-blooded" intentionality as ascribed to mind<sup>187</sup>. An analogy may be drawn of intelligence and the stupid underlying processes. In case of the brain, the most basic theoretical level would be at the

---

<sup>185</sup> Boden, 1990, p. 96.

<sup>186</sup> Ibid., p. 96.

<sup>187</sup> Ibid., p. 97.

neuroscientific equivalent of the machine code, a level engineered by the evolution. The firing of a neuron can inhibit the firing of another one and this may be explained by the biochemistry of the brain. Notions such as stupidity may hardly be appropriate in explanations of such facts. However, the very basic information-processing functions that are performed by the firing and inhibition of neurons could be characterized as “very, very, very... stupid”<sup>188</sup>. Also Boden points out that when she says that the brain does not understand English, the understanding here is of the language English. She doesn’t mean that the brain doesn’t understand anything at all. The understanding of English by the brain may be explained by referring to a more simple or stupid form of understanding which occurs at the level of neuronal information-processing processes. However the limited understanding of the underlying processes cannot be at the same level of understanding as understanding of English. This is how, at a theoretical level, comparison of a brain and a computer may be made in terms of the underlying processes and insights may be gained into them. There by the analogy drawn by computationalists and the computational psychologists between mind and machine is not completely wrong. This comprises her first rebuttal.

The second rebuttal is termed English Reply by Boden. The English reply is basically the assertion that the instantiation of a computer program, no matter by man or by machines, does involve some kind of understanding<sup>189</sup>. The machine or computer in case of AI and Searle-in-the-room in case of the above example does have some understanding of at least the rule book. In the example Searle-in-the-room’s being able to understand English, the language in which the rule book was given, is critical to his manipulation of the symbol. Without this familiarity with English, the robot would not behave in a given context. It is not necessary that Searle-in-the-room understands the whole of English language, all that is required is he understands a limited subset of it. The only English he needs to grasp is whatever is necessary to interpret the rule-book. Similarly, the computer is not completely lacking of understanding. It understands the programming language just as Searle-in-the-room understands the rule book. This claim may seem contradictory. In fact Searle’s basic premise is that a computer program is wholly defined as purely formal in nature. The operations are specified purely syntactically and have no semantic content to be understood. If this is accepted then the above response by Boden, that is

---

<sup>188</sup> Ibid., p. 97.

<sup>189</sup> Ibid., p. 97.

the English reply, doesn't work. But if it is accepted computer programs are not only concerned with syntax then it works. Boden choose to accept it. She points out that a computer program is a program for a computer<sup>190</sup>. As such, when the program is run on a suitable hardware, the machine does obtain some results. A programming language is a medium not only for expressing representations but also for bringing about the representational activity of certain machines. Thus there is more to a computer than just formal symbolization.

A programmed instruction is a procedure specification that can cause procedure in question to be executed. This attacks one of Searle's basic premises that computers are mere symbol manipulation. Boden provides Smith's<sup>191</sup> argument that the familiar characterization of computer programs as all syntax and no semantics is mistaken. It is often the case that computer programmers draw a distinction between the knowledge representations and the procedures required to interpret it when a computer is run. In fact they are often written in two quite distinct programming languages having distinct formal structures. They may also be written in the same language and thereby under the same formal structure. However, Smith argues that it will be less confusing to adopt a unified theory of programming language that cover both the denotative and procedural aspects<sup>192</sup>. He argues so because the theoretical distinction between the denotative and the procedural aspects often lead many, even the computer programmers, to the assume that they are totally different and when considered in isolation they seem to be mere abstract symbol manipulations, just as Searle describes them to be. This Smith, and following him also Boden, think is a mistake. For instance consider the statement "J. Bachchan is the mother of A. Bachchan". This may be represented in various ways knowledge is represented in a computer program<sup>193</sup>. For the computer to make use of the knowledge representations there has to be an interpreter-program. So if the question is asked – who is the mother of A. Bachchan, the interpreter program will enable the computer to find the relevant knowledge in the database and give "J. Bachchan" as the answer to the question. No computer program has knowledge representations without a procedure given to interpret them. In the Searle-in-the-room example, Searle is given a rule book in English precisely because he understands English. If suppose he

---

<sup>190</sup> Ibid., p. 99.

<sup>191</sup> B. C. Smith is a computer scientist. See Smith, B. C., *Reflection and Semantics in a Procedural Language*, Cambridge, Massachusetts: MIT Ph. D. dissertation and Technical Report LCS/TR-272, 1982; also see Boden, 1990, pp. 100-3.

<sup>192</sup> Boden, 1990, p. 101.

<sup>193</sup> For knowledge representation see chapter 1, section 6.

were given the rule book in Japanese, he would not be able to manipulate the symbols. Similarly in the case of computer programs they can only make use of the rule book if they understand the rule in the first place. So computer programs consists of interpreting procedures which give them some hold on semantics, however limited that might be. And though the theoretically distinct descriptions of the programmes as knowledge-representations and interpreter-programs give the sense of distinct abstract formalisms, they are best described within the same formalism and having some form of causal relation. The causal relation here may be explained as the interpreter program causing the computer to locate the relevant knowledge in the database. Thus, the semantic content of a computer program may given by referring to knowledge representations which may denote some facts/variables (such as J. Bachchan and A. Bachchan which may be replaced by variables) or relationships between facts/variables (such as the relationship between J. Bachchan and A. Bachchan) that a programmer may specify or map onto the computer's database. The syntax may then be specifications to arrive at those semantic contents. Further, there are causal links which are formal specifications or instructions connecting some rules in the interpreter program with the relevant knowledge representations. These causal links between the interpreter program and the knowledge representation may also be referred to by the program itself, and hence, those causal links, then may serve as content for the programs. Thus, Boden infers, "It follows from Smith's argument that the characterization of computer programs as all syntax and no semantics is mistaken"<sup>194</sup>. Thereby, Searle is mistaken. A system's causal powers are not always only in terms of it referring to some external object or content. Sometimes it refers to some causal links which are purely internal computational processes.

Two things may be noted in the above discussion. Firstly, in a computer program syntax is related to semantics and that syntax cannot be just in isolation. Syntax cannot be in isolation because in a computer programme they are supposed to bring about the representational activity of the computer which requires that the symbols employed denote knowledge and relations mapped on to the program and also the causal links which brings about the representational activity of the programme. So computer programs are not purely syntactical. Secondly, semantics need not be realized only in a biological system. It was explained above that computer programs consists of some form of semantics, no matter how limited or restricted the sense of

---

<sup>194</sup> Boden, 1990, p. 102.

semantics here may be. Computers are then one system that is not biological and yet has semantics involved.

Thus, Boden rebuts both of Searle's premises – the premise that intentionality is merely biological and the premise that there is only formal symbol manipulation in computers. Of course the upshot of this is that computationalism escapes Searle's objection by asserting that computer is not only about formal syntax where Searle asserted that a computer is essentially formal symbol manipulation. Also it was shown how analogy may be drawn between brain processes and computer processes. Earlier, the most basic theoretical level in case of brain was taken to be the neuroscientific equivalent of the machine code, a level engineered by the evolution. Such analogies between the brain and the computational processes, thus, help in gaining insights into underlying processes of the brain. This leads to the rebuttal of Searle's assertion that computational psychology or any other such endeavour to understand mind which is either based on AI and its underlying theory of computationalism or takes them to be augmenting in its studies is bound to be failure or that it is to be discredited.

There seems to be a collateral implication of Searle's discussion of the mind-body problem, Chinese room argument and rejection of new sciences that are associated with AI and computationalism. It all culminates down to an assertion, the assertion that mind, and thereby mind-body relation may be understood and any confusion<sup>195</sup> on it may be resolved by appealing to the already existing scientific knowledge of the world that we have. Searle had earlier in section 1 of this chapter, argued that the concept of micro- and macro- features as described in physics may be employed to understand and explain the relation between the mind and the brain. According to him mind is basically biological as far as it is caused by brain and issues of it can be addressed within the biological framework. No new science or field of study is required. All the essential principles that we need to understand mind are already there in the current science. So all we need is a better scientific understanding of the relevant subject matter, may be some technological advancement to get access to scientific data (data on neurophysiological processes

---

<sup>195</sup> The confusion here refers to questions such as- “what is the relation of mind and the brain”, “how does mental features such as intentionality and consciousness fit in the overall physical picture of the world”, “how can we ascribe intentionality to something that is constituted of just atoms and voids” and so on. The confusion also refers to the confusion regarding the nature of mental states such as considering it to be a separate substance distinct from the brains, whether mind is to be considered only as those mental states which are conscious and so on which leads one to commit mistakes in finding answers to the above questions. However, further discussion of these confusions is not taken up here.

of brain) not available yet. Or wait till the time when a more detailed account of the biological and physical composition and mechanism of the brain are given, and all the confusions and the mystery of mind (in the sense that nature of mind and its place in the physical picture of the world may seem mysterious to some) will go away. Not only that but Searle claims the confusion, at least the confusions associated with mind body problem, already gets resolved in light of our current scientific understanding. Other novel fields of studies, such as computationalism and thereby AI (though it may have its own technological benefits) is not required to shed more light on the mind.

It is such a contention that Roger Penrose seems to object to. The next section then takes up Penrose's thesis on issues of mind.

### **3. Penrose's Thesis and Critique of AI**

Penrose objects to Searle's position that a computer can simulate a mind. This position is also known as Weak AI. But Penrose's own viewpoint shares with Searle's position in objecting to strong AI. Strong AI is basically the position that thinking is computational and thereby computers can do it as good as human brains. Software is compared to mind and hardware to the brain. There are then three distinct viewpoints – viewpoint of Strong AI, Searle's viewpoint and Penrose's viewpoint. Searle's viewpoint is that computers cannot think nor have mind. This, he says, is because aspects of thinking or of having mind such as consciousness, intentionality, subjectivity and mental causation are unique to only biological matters that brains (more specifically human ones) are made up of. But Searle in his objection to strong AI does allow for the simulation of mind by a computer, though this simulation is not equal to the real mind. Simulation is not equal to duplication. Furthermore Searle does allow that a mind can be considered as a computer and thereby can be understood as being computational, with only the distinction that it has consciousness and so on by virtue of its being biological.

Penrose shares Searle's view that computers cannot have minds. Till this point Searle's objection does in fact help Penrose's position. But Penrose further goes on to assert that a computer cannot even properly simulate mind where Searle has conceded that a computer can do so. Searle's objection to strong AI was a matter of the biological composition of the brain. Computer is defined solely as a symbol manipulation system and thereby consisting of only



syntax which by itself does not give semantics. In case of a human mind semantics is present owing to the neuroproteins of the brain. The nature of refutation given by Searle against strong AI does not depend at all on the current or future stage of technology. It depends on the very definition of a digital computer<sup>196</sup>. As such the refutation is in principle denying strong AI. But this objection still allows simulation of the mind. This allowance of simulation of the mind by a computer program in Searle's viewpoint is what Penrose has objection to. Penrose's objection to strong AI is also an argument that refutes AI in principle. Here 'in principle' again means that the argument does not have to do anything with the current or future stage of technology. Some may argue that it is only due to lack of sufficient level of current technology that computers cannot have minds yet. But the argument being provided by Penrose does not have to do anything with technology. His refutation is based on the very idea of computationalism. As such he attacks the very basis of Strong AI.

Computationalism may, perhaps, be said to have started with and has its foundation in Turing's thesis. So to attack computationalism at its core will be to contradict Turing's thesis. It should be noted that contradicting Turing's thesis is not to be understood as contradicting the 'Church-Turing' thesis. Both Church and Turing independently had proved that anything that could reasonably be called a mathematical process could be achieved within a particular scheme discovered by Church, known as the *lambda* calculus<sup>197</sup>. Turing, on his own, showed the same in terms of Turing machine actions. The contradiction is not of these assertions. The contradiction here is of the Turing thesis according to which, a machine can compute what a human computer can compute<sup>198</sup>. Turing, it seems, viewed physical actions, including cognitive actions of the human brain, to be reducible to some kind of Turing machine action. It is this assertion that Penrose intends to contradict.

Turing machine is basically a theoretical mathematical model of a machine and thereby has its foundation in mathematics. Penrose to contradict Turing's thesis thus feels that a powerful argument to show that mind cannot be properly simulated by a computer has to come from the mathematical domain. He uses the concept of mathematical understanding. He shows how

---

<sup>196</sup> Searle, 1984, p. 30.

<sup>197</sup> It is not of much importance to the thesis to explain what *lambda* calculus is. Its purpose here is just a matter of facts.

<sup>198</sup> See chapter 1, section 1.

mathematical understanding applied in mathematical computation is itself non-computational and thus cannot be reduced to computation. Penrose feels that most of human mental activities involve consciousness and understanding<sup>199</sup>. If he can show that human understanding cannot be computational, then there is at least something non-computational about human minds. Mathematical understanding, being as such a form of human understanding, is therefore taken up. Penrose then uses the Gödel argument to show that human understanding cannot be computational at all. Two questions do arise – why would showing that human understanding is non-computational achieve the end of providing some objection to Strong AI; and what is Gödel’s argument? To answer the first question, Penrose asserts that the general claim of strong AI is that a computer can think, that *artificial* intelligence is possible. But according to Penrose, intelligence requires some understanding. There can be genuine intelligence only when understanding is involved<sup>200</sup>. He does allow that some degree of simulation of genuine intelligence is possible without any actual understanding. For this he gives the example of how human individuals may often make us believe that s/he understands something when the converse is true. But, otherwise genuine intelligence necessarily requires genuine understanding. This is another reason as to why Penrose takes up understanding or more specifically mathematical understanding to give his argument against strong AI.

The second question is in regard to Gödel argument. The Gödel argument is derived from and is a simple form of Gödel’s theorem. In 1930, mathematician Kurt Gödel presented his theorem which was of fundamental importance to the foundation of mathematics. According to Penrose, this theorem also had major implication for the philosophy of mind. What the theorem established among many other things was that “... no *formal system* of sound mathematical rules of proof can ever suffice, even in principle, to establish all the true propositions of ordinary arithmetic.”<sup>201</sup> Penrose argues that based on this proposition, a case can be made for establishing “... that human understanding and insight cannot be reduced to any set of computational rules.” What this means is that in a mathematical system, there are mathematical truths which are given by human intuition or insight and it is this intuition or insight that cannot be reduced to any set of rules. This is the Gödel argument.

---

<sup>199</sup> Penrose, Roger, *Shadows of the Mind*, 1994, p. 51.

<sup>200</sup> *Ibid.*, p. 37.

<sup>201</sup> *Ibid.*, p. 64, 65.

A demonstration of Gödel's theorem may be given. This requires first to understand non-stopping computations. Computation has already been discussed in chapter 1. As an example of carrying out a computation Penrose demonstrates a simple task<sup>202</sup>: Find a number that is not the sum of three square numbers. A number here means a natural number, i.e., 0, 1, 2, 3, 4 and so on. So square of these numbers will be 0, 1, 4, 9, 16 and so on. The computation will then try each natural number in turn, starting from 0 and see whether it is a sum of three square numbers. So starting with 0, one would first check if it equals to the sum of three square numbers which are also calculated starting from 0. So in this case  $0^2+0^2+0^2=0$ . This checks out with 0, the number for which it was being checked whether or not it is the sum of three square numbers. Next the computation tries for 1. It will start with 0 again.  $0^2+0^2+0^2=0$ ,  $0^2+0^2+1^2=1$ . 1 is then a number that is the sum of three square numbers. Next the computer tries for 2.  $0^2+0^2+0^2=0$ ,  $0^2+0^2+1^2=1$ ,  $0^2+1^2+1^2=2$ . This goes on for 3, 4 and so on. So if one continues calculating the sum of square number then we get:  $0^2+0^2+0^2=0$ ,  $0^2+0^2+1^2=1$ ,  $0^2+1^2+1^2=2$ ,  $1^2+1^2+1^2=3$ ,  $0^2+0^2+2^2=4$ ,  $0^2+1^2+2^2=5$ ,  $1^2+1^2+2^2=6$ . But when we move forward with the computation it can be seen that the sum of three squares beyond what has been calculated till 6 give numbers greater than 8. The computation will thus stop at 7. It gives us 7 as the number that is not the sum of three square numbers. The computation does halt and a result is given.

Penrose gives another task: Find a number that is not the sum of four square numbers. So we can proceed with the above computation. Just adding a  $0^2$  with each step will show that till 6 the computation finds the numbers to be the sum of four square numbers. Also now,  $1^2+1^2+1^2+2^2=7$ . So the computation moves on to 8, which also checks out as  $8=0^2+0^2+2^2+2^2$ , then to 9 which is equal to  $0^2+0^2+0^2+3^2$  and so on. The computation goes on and on. It actually turns out that this computation never stops. Every number is the sum of four square numbers and there is no number that may be given as an output for this computation, as the task was to find out a number that is not the sum of four square numbers. This then is a computation that never terminates. Another computation may be: Find an odd number that is the sum of even number. It is obvious that this computation will not stop. There is no odd number which is the sum of even numbers. But how is it that one decides whether a computation stops or not? From the above two computations that do not terminate it may be observed that the latter is obvious while the former

---

<sup>202</sup> Ibid., p. 66.

is difficult to ascertain for sure. Thus, computations that do not terminate can be of both types. So what procedures do mathematicians use to demonstrate that certain computations do not terminate? Can it be that they themselves follow some algorithm in order to be sure that a particular computation stops?

The question boils down to this – are there any clear cut procedures or rules that is sufficient to establish the non-stopping nature of computations that do not stop? The answer to this is no. Penrose claims this as one of the implications of Gödel’s theorem. Any set of rules will be insufficient to demonstrate that a computation never stops. For instance each of the above computations can be termed as  $C_1$  for finding a number that is the sum of three square numbers,  $C_2$  for finding a number that is the sum of four square numbers and  $C_3$  for finding an odd number that is the sum two even numbers. Now suppose there is a set of procedures or computation  $A$  for determining whether computations such as  $C_1$ ,  $C_2$  and  $C_3$  do stop or not. From  $C_1$  it may be observed that the computation stops only when it finds a result, the result being 7.  $A$  will terminate only when it finds a computation that will in fact not stop. So in case of  $C_1$ ,  $A$  will not stop. Here it is not important that when  $A$  applied to  $C_1$  will never stop.  $A$  will have to be tested with a computation that never terminates as  $A$  is a procedure for determining computations that do not terminate. Suppose that a computation that does not terminate is fed to the  $A$ . Normally  $A$  should stop, but suppose that it does not stop. Even such cases do not show that  $A$  is an unsound computation to determine whether the computation is one that does not terminate. But the important point here is that whenever  $A$  stops then the computation which was fed to  $A$  has to be a computation that does not stop.  $A$  never gives a wrong output or answer. Thus, as long as  $A$  terminates when a non-stopping computation is fed to it, it is a procedure to determine whether a computation does not terminate. Now instead of  $C_1$ ,  $C_2$ , and  $C_3$ , a computation that depends on a natural number is considered. This may be represented as  $C(n)$ .  $C(n)$  then may be thought of as providing for a family of computations where there is a separate computation for each natural number 0, 1, 2, 3... and so on. The computations are  $C(0)$ ,  $C(1)$ ,  $C(2)$ ,  $C(3)$  and so on respectively, where the way in which the computation depends upon ‘ $n$ ’ is itself entirely computational. All these can easily be applied in terms of Turing machine.  $C(n)$  then is the action of some Turing machine on the number ‘ $n$ ’. The number ‘ $n$ ’ is fed to the machine as input and the machine then computes on its own from then own. What is of importance here is that whether the computer’s action ever stops or not, for each choice of ‘ $n$ ’.

Suppose again that  $A$  is a computational procedure which when terminates shows us that a computation such as  $C(n)$  actually does not ever stop.  $C(n)$  here is a computation for any natural number 'n'. It is not important that  $A$  can always decide that  $C(n)$  does not stop. What is important that  $A$ , whenever it stops, gives us the conclusion that  $C(n)$  does not stop. It does not give wrong answers. Thus,  $A$  may be said to be a sound procedure. Now all the different computations  $C(n)$  may be coded as  $C_1, C_2, C_3 \dots$  and so on for  $A$ . The  $q^{\text{th}}$  computation may be referred to as  $C_q$ . One can imagine that for a natural number 'n' there may be many computations.  $C_q$  here is the  $q$ th computation for a natural number. When such a computation is applied to a particular number 'n', we may list it as  $C_1(n), C_2(n), C_3(n)\dots$  and so on. It is important here that this listing is computable, that is, there is a single computation  $C_\bullet$  that acts on a pair of numbers 'q' followed by 'n' to give  $C_q(n)$ . So if  $q = 1$  and  $n$  is a natural number starting from 0 then  $C_q(n)$  will be  $C_1(0), C_1(1), C_1(2)$  and so on. Now  $A$  is the computation that when presented with the pair of numbers 'q' and 'n' tries to ascertain if the computation  $C_q(n)$  does not halt.  $A$  here is dependent of 'q' and 'n' and hence the computation performed by  $A$  may be given as  $A(q, n)$ . So if  $A(q, n)$  stops then  $C_q(n)$  does not stop. As the computation goes on for each number 'q' and 'n', there will a point where 'q' will be equal to 'n'. Then if  $A(n, n)$  stops, then  $C_n(n)$  does not stop. Here it is observed that  $A(n, n)$  now depends on one number and not two. It must be one of the computations  $C_0, C_1, C_2, C_3$  and so on, as applied to 'n', since this was supposed to be a listing of all the computations that can be performed on a single natural number 'n'. Suppose that this particular computation is  $C_k$ . So now  $A(n, n) = C_k(n)$ . If  $n=k$ , we have then  $A(k, k) = C_k(k)$ . Also earlier it was derived that if  $A(n, n)$  stops then  $C_n(n)$  stops. From this we have if  $A(k, k)$  stops then  $C_k(k)$  does not stop. From this it is deduced that the computation  $C_k(k)$  does not in fact stop. But then  $A(k, k)$  cannot stop either as it is the same as  $C_k(k)$ . Thus, there is a contradiction and  $A$  is incapable of demonstrating for sure whether a particular computation, in this case  $C_k(k)$ , is a non-stopping computation<sup>203</sup>.

We do know that  $A$  is sound, and if  $A$  is sound then,  $C_k(k)$  does not stop. Now coming back to mathematical understanding, in case of computations that do not terminate it is obvious to our mathematical understanding that a computation such as the one given earlier is a computation that does not stop. But now we know something that  $A$  is unable to ascertain. From

---

<sup>203</sup> Ibid., p. 75.

all these it then follows that “*A cannot encapsulate our understanding*”<sup>204</sup>. There is a computation  $C_k(k)$  that we know does not stop and yet the computational procedure *A* cannot ascertain that fact. This is the Gödel (Turing) theorem. An important thing is that *A* must be sound. Then we know that a sound set of computational rules can never be sufficient to determine that computations do not stop, since there are some computations that do not terminate that elude such rules. From this Penrose concludes – “Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth.”<sup>205</sup>

For Penrose this is a conclusion that is inescapable. The argument points out something very significant about the mental quality of understanding. The essential point to notice is that the algorithm *A* exhibits two different levels in the argument. At one level, it is the normal algorithm that is supposed to give us an outcome. But at the other level, *A* is treated itself as being an algorithm that will not stop. This argument does not only apply to algorithms but extends to conscious ‘understanding’ too. Penrose says, “... the interplay between the two different levels... considered – as a putative instance of conscious activity and as a computation itself – that allows us to arrive at a conclusion expressing a fundamental conflict between conscious activity and mere computation.” This then provides the basis for the claim that conscious mathematical understanding cannot be properly modeled at all in computational terms, whether it is the top-down approach of computation or bottom-up approach, or even a hybrid of the two.

What may be observed from the above is that *A* if sound, that is, consistent cannot be complete because earlier it was seen that *A* which was taken to be a sound procedure for ascertaining the non-termination of computations failed to do so as it showed itself as a computation that both stops and does not stop. What this amounts to is the claim that a formal system cannot be both consistent and complete at the same time. This was in fact one of Hilbert’s problem<sup>206</sup>, to search for a system that is both consistent and complete. Gödel showed that there cannot be a formal system that is both complete and what he originally termed as  $\omega$ -consistent.

---

<sup>204</sup> Ibid., p. 75.

<sup>205</sup> Ibid., p. 76.

<sup>206</sup> David Hilbert was a mathematician and a prime figure in a movement known as *formalism*. The aim of the movement was to establish mathematical reasoning on secure and precise foundations.

The original assertion was that: If  $F$  (a formal system) is  $\omega$ -consistent<sup>207</sup>, then it cannot be complete. For the purpose of Penrose, algorithms and formal systems are taken to be equivalent as procedures for evaluating mathematical truths. Another point of importance to be noted is that the soundness of a system implies its consistency. That means, any propositions within that system will be consistent with the basic set of rules or axioms that define that system. The axioms will prove the consistency of such statements. But what about those fundamental axioms themselves? What makes us believe that the soundness of the system guarantees the consistency of those axioms? The fact is that the consistency of those axioms which define the system has to lie outside the system. The system cannot prove itself to be consistent. This is what Gödel and later Rosser<sup>208</sup> showed, that the consistency of a formal system is something that lies outside the system itself to establish. The importance in this is that it shows us how to go beyond any given set of computational rules that we believe to be sound and obtain further rule(s), not contained within those rules that we believe to be sound. Those further rule(s) are namely the rule(s) asserting the consistency of the original rules. This same mathematical evidence, Penrose points out led Gödel and Turing to arrive at different conclusions. Gödel, on one hand, came to the conclusion that though physical brain must itself behave computationally, the mind is something that is beyond the brain, so that the mind's action is not constrained to behave according to some computational rules. On the other hand, from this very evidence, Turing comes to the conclusion and indeed demonstrates what is called the Universal Turing machine. Turing suggests a way of evading the same conclusion as Gödel. He suggests that the mathematician's algorithm would be 'technically' unsound and it certainly would not be knowably sound. But either way, Penrose claims that both Turing's and Gödel's conclusion does not seem to be inconsistent with his position, that there is something beyond the computational algorithms which is non-computational.

From the above, Penrose thus proves mathematically that there is something non-computational which in fact is beyond computation and must be inferred from the computation itself. Based on this then, Penrose argues that mathematical understanding is non-computational. Given that mathematical understanding is non-computational and mathematical understanding

---

<sup>207</sup> Ibid., p. 90.

<sup>208</sup> J. Barkley Rosser was an American logician. He obtained the version of Gödel's incomplete theorem which asserts that for a formal system  $F$ ,  $F$  cannot be both complete and consistent. In Gödel's original pronouncement, it was  $\omega$ -consistent.

being ‘understanding’, it may be deduced that there is something about ‘understanding’ that is non-computational. Thus, genuine understanding cannot be reduced to computation. Now, Penrose also does assert that genuine intelligence requires genuine understanding and hence, genuine intelligence cannot be simulated properly or wholly. This is then an objection to the viewpoint of strong AI. But in presenting his argument against strong AI, Penrose also points out why his viewpoint is not the same as Searle’s and in fact does provide it as an objection to Searle’s claims. His argument thus, provides for the rejection of two of Searle’s claims. First is that Searle claims that everything may be simulated in a computer. The universe, the human brain and thereby mind may also be simulated computationally. It is only the case that they are not the exact thing as mind as they are not instantiated in biological systems. This claim, that mind can be considered as computational, and thereby simulated is what Penrose rejects.

It is seen from the above discussion that Penrose’s argument serves for the wholesale rejection of any computational simulation of mind. Intelligence requires ‘understanding’. ‘Understanding’ requires awareness. Awareness is stated by Penrose as the passive aspect of consciousness<sup>209</sup>. The active state of consciousness then is taken to be free will. The perception of sound, color, pain and so on may be considered as the passive aspect. Active consciousness comprises of willing an action or deciding to desist from something. There may also be mental activities such as bringing to mind some early memory that may involve both active and passive aspects of consciousness. It seems necessary then that some kind of consciousness is present in the type of mental activity that is referred to as “understanding”. Now since ‘understanding’ requires consciousness, it then shares the passive and active aspects. In Searle’s case, according to Penrose, the Chinese room argument argues that though the output of the computational activities resembles the output of activities that requires understanding, there is no understanding involved in the computational output. According to Penrose, Searle’s argument against strong AI is concerned with the inward, passive or subjective aspect of ‘understanding’,<sup>210</sup>. But the argument does not deny the possibility of simulation of ‘understanding’ in its active, outward or objective aspects. For him brain might as well be a digital computer. In the Chinese room argument, the outside conversation in Chinese is taking place and hence the active aspects of understanding is being simulated (note, not duplicated but only simulated), though inside the

---

<sup>209</sup> Penrose, 1994, p. 39.

<sup>210</sup> Ibid., p. 41.



room Searle-in-the-room does not understand anything. Penrose points out, “Searle allows that a simulation of the output of the results of understanding could be possible... since he is prepared to admit that this could be achieved by a computer simulating every relevant physical action of a human brain (doing whatever it does) when its human owner actually understands something”,<sup>211</sup>. But Penrose’s argument asserts that even this outward aspect of ‘understanding’ cannot be simulated. This again is demonstrated by showing how mathematical understanding, which involves the outward, objective aspect of understanding is non-computational. Now ‘understanding’ in both its active and passive aspects is demonstrated to be non-computational. The very ‘understanding’ that underlies the computational rules is something that is itself beyond computation. Similarly by appealing to mathematical awareness and also to the above discussion, it may be said that consciousness is beyond computation. This then contradicts the Turing thesis, the assertion that computers can think same as the human mind or brain. This contradicts the thesis, for how can something computational in its nature and operation process something non-computational.

The second claim is that Searle asserts that the mystery of mind can be solved by our current day scientific knowledge. Mind is ultimately a biological phenomenon. And as that it readily fits in with our present day scientific picture of the universe. Penrose’s argument asserts that there is something non-computational in our mental activities. This then implies that the non-computational ingredient is to be accounted for by the present day scientific understanding as well. But the current scientific understanding fails to do so. Suppose that one accepts Searle’s assertion that mind is unique to biological stuffs, the brain to be more specific, and the non-computational ingredient is the biological stuffs. But this does not hold. As mentioned earlier, Searle allows for the mind to be simulated, at least in its output. But Penrose has argued that even the outward manifestation of mind has non-computational aspects to it as in the case of mathematical understanding and awareness. Thus, Searle’s assertion that mind can be simulated is rejected by Penrose. Note that Searle allows for the simulation of mind, but he asserts that this simulation is not equal to duplication. Penrose in demonstrating that there is a non-computational ingredient to the phenomena of mind also draws attention to what its implications are regarding our knowledge of mind in the scientific picture. If there is such an ingredient, and there is

---

<sup>211</sup> Ibid., p. 41.

according to Penrose, then the scientific picture has to explain it. The current day scientific picture is given by physics. But according to the language of physics and nowhere within the physics among its fundamental principles we find anything that may be used to account for this ingredient. One reason for this is that the whole of physical laws, principles and even the universe as understood within this physical picture could more or less be described computationally<sup>212</sup>. The thing to be accounted for here is something non-computational. As such, it is only reasonable then to conclude that a new science or radical changes to the paradigm of current science is required. It is not to say that the current science is irrelevant, but if science has to be broadened and extended to accommodate this ingredient then it will have to be at fundamental levels because if the whole of physics is unable to account for the non-computational aspect of mind (which we know for sure to be present), then there must surely be something missing at the fundamental levels on which rests the whole physical picture of the world. And the subsequent science that emerges will be revolutionary and will include principles that cannot be accommodated within the current scientific framework. This new framework is what, Penrose believes, will account for the non-computational nature that is very apparent in mental phenomena which are very much part of our universe.

It may seem that a collateral implication of Penrose's arguments is a fourth viewpoint. This viewpoint is the position that mind is something totally mystical and it is beyond what can be known or beyond scientific understanding. Penrose agrees that his arguments might lead towards some kind of mysticism or platonic world of ideas. But he rejects mysticism on the grounds that it asserts that mind cannot be accounted for within scientific paradigm, that its nature is mystical and therefore cannot be known or requires some mystical power to grasp it. Penrose finds Platonism somehow consistent with his position. The platonic idea of a world of forms as distinct from the physical world implies that there is something beyond the physical world. The world of forms is what makes the physical world intelligible in the first place. While the physical world may be described computationally, there is still the world of forms that is not accounted for through the computational description of the physical world. It implies that there is something beyond it which is in fact only accessible through our mind, as is often accessed

---

<sup>212</sup> This stems from Penrose's comment that, "The precision and scope of physical laws, as presently appreciated, is extraordinary, yet they contain no hint of any action that cannot be simulated computationally." He also says, "it appears that neither classical nor quantum physics, as presently understood, allows room for non-computable behaviour of the type required". See Penrose, 1994, pp. 214-16.

through mathematical understanding for instance. As such Platonism possesses no threat to Penrose's position.

For Penrose, mind has to be accounted for scientifically, though it may require a new kind of science. "Whatever it is that ... describes mind must ... be an integral part of the same grand scheme which governs all the material attributes of our universe."<sup>213</sup> Thus, Penrose argues that mind has to be accounted for according to the scientific picture of the world but to do that the scientific understanding of the world might itself require some changes or new forms of scientific method might be required that the classical physics lacks. Additional support is given to this claim by Penrose through analogies. For instance it is claimed by some that to view mind as something different, as is the case with dualism, is a categorical mistake. Such a claim is made by Ryle against Cartesian dualism. He asserts that Descartes had committed a categorical mistake when he concluded from his method of doubt that mind is separate from body and thereby is a distinct separate substance<sup>214</sup>. Penrose though rejects dualism for a scientific explanation of mind, precisely for the reason mentioned above, he does also object to the criticism that in considering mind as different from the matter a categorical mistake has been committed. He says that merely pointing to a possible categorical error in case of mind and body does not actually solve the puzzle. The puzzle itself is genuine no doubt, and Ryle's accusation of a mistake does not solve it. Penrose shows that even in current day physics there are certain concepts which can be equated with one another, even though it appears to be a categorical mistake<sup>215</sup>. He gives the example of  $E = mc^2$ . The equation in fact effectively equates energy with mass. It might look like a categorical mistake as mass is a measure of actual substance where as energy is the potentiality for doing work. Yet this equation has been experimentally proved and serves as a cornerstone of modern physics. Thus, the mind-body puzzle is not mere 'category mistake' and might in fact be accounted for by a broader scientific understanding which deals with the non-computationality of mind but in terms of the physical, where the concept of physical itself has become broader. Additional support is given by Penrose to his claim that the phenomenon of mind requires fundamental changes in the current scientific picture. He draws an analogy with the transformation of science from Newtonian physics to

---

<sup>213</sup> Ibid., p. 213.

<sup>214</sup> For more on this one may refer to chapter 2, section 3.

<sup>215</sup> Penrose, 1994, p. 214.

Einstein revolution. It was pretty much accepted since the time of Newton that gravity and its mathematical description served as the model for the description of other physical processes. It served as the foundation of physics as most of the things could be explained more or less in terms of Newtonian gravitational theory. Einstein examined the very basis of Newton's theory and in 1915 came up with his 'General Theory of Relativity'. This was a revolutionary theory and provided a radically different picture of the universe. Gravity was no longer a force that acted between bodies. It was represented as a kind of curvature in the space-time fabric in which all other particles were housed<sup>216</sup>. Penrose believes that such fundamental re-examination of modern physics is required so as to accommodate the non-computational aspects of mind. This will lead to a better, broader scientific understanding of the world where the current principles of physics may be considered to be obsolete just as Newton's theory became obsolete after the Einstein revolution.

#### **4. Warwick's Thesis**

The discussion till now may be considered in general, to have put forward, four positions. First is the strong AI position which asserts the possibility of machine/computer being able to think or be intelligent same as a human being. Next is the weak AI position which may be attributed to Searle's position. This position argues that simulation of intelligence/mind/thinking by a computer/machine is possible but not the exact thing. The third position is that of Penrose who argues that proper simulation is not possible at all by a computer of the human mind, yet explanation of human mind in terms of understanding and consciousness should lie within the framework of some new or extended physics. Last position is basically the position that human mind and therefore, consciousness, intelligence and so on, is something which cannot be properly simulated computationally nor can be explained by science. Penrose points out that such a position might be that of the 'mentalists'. Mentalists are those who hold that both the physical and mental phenomena are ultimately explicable in terms of the mental. This is known as mentalism. Alternatively, this last viewpoint might also be held by those who consider mind as explicable only through mystical explanations. It may be observed that in each of these positions, human mind or intelligence seems to be the standard according to which AI's possibility and capability are discussed. Kevin Warwick picks up this issue and charges most of the discussion

---

<sup>216</sup> Ibid., p. 217-20.

on AI to be human-biased. This serves to be the main underlying contention of Warwick against most of the viewpoints regarding AI.

Warwick points to the apparent biasness in the whole discussion on AI. He starts with the Turing test. The claim was that if a machine, playing the imitation game, can make a human interrogator come to the conclusion that it is a human, then the machine will be considered as intelligent as a human. Warwick asserts that this kind of test is unfair for the machine to examine its intelligence. He says, “If, during the Turing Test, the human participant is not fooled by the machine, it is *not* correct to conclude that the machine is less intelligent than the human.”<sup>217</sup> Imagine for example that a human took the place of the machine and cats took the place of humans. If the human failed to make the cat think that the other being on the end is a cat then will it mean that the human is any less intelligent than the cat? It is most likely that the human will fail this test at convincing other cats that it is a cat. But this will not mean that the human is any less intelligent. Thus, test for intelligence, according to Warwick should be free of any such biases. He feels the “need for a viewpoint on AI that is much less anthropomorphic than the classical AI”<sup>218</sup>.

Turing’s claim that a machine could imitate humans seems to be the position of weak AI. Marvin Minsky defined AI as “... the science of making machines do things that would require intelligence if done by men.”<sup>219</sup> This definition, which is also what classical AI subscribes to, clearly is what weak AI claims. As such the beginning of AI itself is laced with human bias. Warwick feels that such biased understanding of AI’s aim is outdated. This biasness in AI is in turn owing to a very biased understanding of ‘intelligence’. It is only human beings who are considered to be intelligent or at least to be superior in intelligence compared to others. It is ironical that such considerations are given by human beings themselves. And now that a machine has to be considered intelligent, it is being done so by the terms and conditions dictated by humans. Warwick feels that a broader and inclusive understanding of intelligence is required which accounts for not only the differences in intelligence between person to person but also culture to culture, human groups/society to human groups/society and species to species.

---

<sup>217</sup> Warwick, Kevin, *In the Mind of the Machine: The Breakthrough In Artificial Intelligence*, London: Arrow Books Limited, 1998, p. 33.

<sup>218</sup> Warwick, Kevin, *Artificial Intelligence: The Basics*, Abingdon, Oxon: Routledge, 2012, p. 69.

<sup>219</sup> *Ibid.*, p. 31-32.

Warwick says, “What is needed now is an up-to-date viewpoint that ....., but also encapsulates the different forms of intelligence witnessed in life in its broadest sense.”<sup>220</sup> Machine is then to be considered just as another species and thus included within this broader view of intelligence. Warwick next suggests a standpoint different from the other standpoints mentioned earlier. This standpoint then asserts that an artefact can be intelligent and think in *its own right* and *its own way*<sup>221</sup>. This is termed by Warwick as ‘Rational AI’. Rational AI then is the claim not only that a machine/computer is or can be intelligent, but also that it is or can be ‘truly’ intelligent in a way that is subjectively considered to be intelligent within the machine species. Whether this intelligence turns out to be in any sense similar to human intelligence or not is kept open. This concept of AI therefore still retains weak AI and strong AI, but in a limited sense. The intelligence may or not may not be similar to human intelligence but it will be doing things that require intelligence and hence a sense of weak AI is included. On the other hand, rational AI claims that the machine/computer will have genuine intelligence, intelligence now understood in a broader sense, and hence has a sense of strong AI. This position then takes in the essential claim regarding intelligence from both weak and strong AI but without the human bias.

Warwick applies this claim of human biasness in his argument against both Searle’s and Penrose’s viewpoints. Searle’s claim was that machines can simulate human mind and therefore simulate intelligence but the simulation is not the same as genuine intelligence. This is due to the fact that factors such as consciousness, intentionality and so on could be present only in the biological brain. Digital computers do not have the capability to have consciousness and other aspects as they are not made up of the same biological stuff. Thus, the intelligence exhibited by a computer will be only a simulated one. This is apparently a case of weak AI. As such it is biased in favour of humans. Hence, a shift in the concept of intelligence, from being possible just in biological stuff to a more inclusive concept, will overcome this boundary set by Searle. Warwick’s accusation that Searle works with such a narrow, human bias definition of intelligence can be seen in his argument against the Chinese room argument. The Chinese room argument supposes that Searle is in a room, and tiles containing symbols are being passed to him. He in turn, following some rules from an English rule book given to him, returns appropriate symbols. To a Chinese speaker the person inside the room seems to be a Chinese person or at

---

<sup>220</sup> Ibid., p. 68.

<sup>221</sup> Ibid., p. 69.

least a person good at understanding and speaking Chinese. But Searle is not a Chinese person and he for sure does not understand anything of the language. He just knows English. From this, he argues that a computer in the same position as Searle-in-the-room does not understand what it is simulating. Thus the computer is not intelligent and does not possess understanding of the assigned task. The Chinese room argument uses the same setting as the Turing test. So in a way the Chinese room argument directly argues against the Turing test. But earlier it was shown that Warwick considers such a test to be extremely human biased. Thus, the same can be said of Searle's argument. Warwick says of Searle's argument, "The argument is human centric"<sup>222</sup>. Warwick in fact turns the Chinese room argument on its head. He suggests that suppose again that a human is in a room and he is given a rule book that he can understand. But instead of tiles containing Chinese symbols being passed in and out of the room, it is tiles containing machine codes that are being passed in and out of the room. Also the entities outside that are sending and receiving tiles and processing them are machines. It is apparent that the human inside the room does not understand anything of the conversation that is taking place but the machines outside understand it very well. But just because the human does not understand the machine language, it does not follow that he is not intelligent or is any less intelligent than a machine. It is the same in the Chinese room argument given by Searle. Just because the machine does not understand Chinese, which is a human language, it does not mean that it is not intelligent.

A similar argument<sup>223</sup> is made by Warwick against Penrose's claim. Penrose has argued that intelligence requires understanding. This follows from the fact that genuine intelligence requires genuine understanding which in turn requires awareness. Awareness is then taken to be the passive aspect of consciousness and free will to be the active aspect. Penrose then explains passive aspect as perception of colours, sensation of pain and so on. Active aspect is explained as desisting from a decision, willing to act and so on. It is pointed out then that in these explanations, the aspects that are being explained are human aspects and in terms of human sensations and actions. It is then obvious that what is meant by genuine understanding and therefore genuine intelligence is a very narrow human oriented form of understanding and intelligence. This narrow attitude towards such concepts is to be changed. Awareness in a human, according to Warwick, is due to the sensory systems (which give awareness of the

---

<sup>222</sup> Ibid., p. 74.

<sup>223</sup> Ibid., p. 62-64.

external state<sup>224</sup>) and the feelings (which is awareness of the internal state). Just because a computer does not have any sensors with which it can be aware, does not mean that one should conclude that it is not capable of being aware<sup>225</sup>. It may be seen that even with machines such sensations and actions can be achieved but in a subjectively machine way and not the same as humans. Warwick explains that human intelligence is partly owing to the genetic make-up of the species (and individuals, when it comes to determining the level) and partly due to the surrounding or environment of the species (or individual again). Education, culture and any other external factors will be included as part of the environment of the individual or species. Warwick calls these factors as 'nature' and 'nurture'<sup>226</sup>. Nature refers to the internal make-up of the organism and nurture refers to the external factors of the organism. It is then in case of humans that their understanding and intelligence are shaped by these two factors. The same thing, claims Warwick, can be said in the case of machines. The internal makeup which consists of the machine's/computer's program may be called the nature aspect and the surrounding with which the machine interacts and the education it is provided with can be called nurture. For instance, a robot may have a body of its own. It may be endowed with infrared sensors instead of eyes that humans have. Wheels might be given for the purpose of legs in a human. Other limbs as required may also be given. This will lead to machines capable of interaction with the world and learning from it. As a result it will be intelligent in its own way. So the machine would actually be aware of the world and thus have awareness of some kind. Hence, Penrose's argument against machines being intelligent, according to Warwick does not hold.

It may be observed that the AI mentioned here as learning is no longer the AI in the sense of classical AI. The classical AI mostly comprised of the top-down approach which considers an overview of the task at hand (simulating human intelligence) and then specifies what it may be constituted of. But the new and updated version of AI, are based on bottom-up approach<sup>227</sup>(may also be combination of both top-down and bottom-up) where bottom units are first specified and the desired system is gradually built up on or emerges from the inter-relations of those units. AI based on this approach can be made to not only learn but also adapt. This form of AI then is

---

<sup>224</sup> External state here refers to both the state of an individual's external physical body and the state of affairs in the world.

<sup>225</sup> Warwick, *In the Mind of the Machine: The Breakthrough In Artificial Intelligence*, 1998, p. 95.

<sup>226</sup> Warwick, *Artificial Intelligence: The Basics*, 2012, p. 23-27.

<sup>227</sup> For more on top-down and bottom-up approach, see the introduction of this chapter.



more in line with the assertion of *rational* AI whereby, the AI would have intelligence of its own kind. The thrust of Warwick's argument against both Searle and Penrose is based on the concept of intelligence. All one needs to accept that machine can be intelligent is to first change one's concept of intelligence by making it free of any biases. In considering that intelligence can be subjective one can then conduct studies to compare among different intelligences in an objective way, without any biases, which scientific methods usually do. Thus, the objective attitude of science has to be employed in understanding intelligence. What Turing and Searle did was to give tasks to decide whether the machine had intelligence. Warwick agrees that performing a task is perhaps the right approach in determining intelligence as it is objective. But such tasks should not be biased towards any one species, especially human beings. And this biasness is what Turing and Searle seems to have committed in their arguments and understanding of machine intelligence. In case of Penrose, Warwick says that defining intelligence in terms of features, that too human features, is not a good way to go about it. First of all intelligence has to be defined in objective terms and the desired performance in specific. Hence, Warwick defines intelligence broadly as, "the variety of information-processing processes that collectively enable a being to autonomously pursue its survival."<sup>228</sup> The task here is then taken to be survival as every species is understood to pursue its own survival and well-being. This definition then includes the intelligence of animals, humans and machines. It also gives a perspective that sees human intelligence as only one kind or variety of intelligence. The definition is taken as an improvement over other definitions of intelligence.

Having made a case of machines being capable of their own intelligence, Warwick also makes the case that human brain can be artificially simulated. Artificial brain can be modelled on human brain using artificial neural networks. Searle had claimed that duplication of human mind is not possible as human brain being biological accounts for the features of mind such as consciousness, understanding and so on. Also our current scientific understanding of physics, or biological phenomenon specifically, already presents us with the necessary ground to explain the emergence of mind. On the other hand, Penrose had said that physics in general had more or less ignored mental phenomena. As such phenomena such as consciousness and understanding still remain left out of the overall scientific picture of the world. Thus, there is nothing within the

---

<sup>228</sup> Warwick, *Artificial Intelligence: The Basics*, 2012, p. 17.

present physical picture of the world that addresses these issues. Penrose concludes that there is an ingredient missing in our scientific knowledge. This ingredient is the non-computational aspect of mind. Once this underlying non-computational aspect is included within what will be the new or extended science then everything including consciousness will be accounted for. In this debate of whether or not the current scientific knowledge is enough to explain issues such as mind and the mind-body problem, Warwick takes a position similar to Searle. According to Warwick, not only the current scientific understanding is enough to simulate but also enough to model the human brain and thereby the mind. So in a way Warwick's assertion is stronger than that of Searle. Warwick at the same time feels that maybe Penrose is right and there might be some aspects of the world (the non-computational aspect exhibited by our mind) that is not accounted for yet. But the science today is already at a level where it can start modelling brain and to an extent the mind. He argues that to model anything one needs to start at an appropriate level<sup>229</sup>. If one aims too high, that is, aims for a too complex level then some features might become ungraspable. This is what he feels that Penrose is suggesting. And though Warwick partly feels that Penrose may be right<sup>230</sup>, he still thinks that not all the necessary basic elements of the brain has been defined yet which would be needed in order to achieve a machine brain that operates almost exactly the same way as a human brain. Not all the necessary basic elements but, Warwick feels, sufficient basic modelling blocks is already known. As such an appropriate level has already been reached to do the modelling. How close to exact will it be? Warwick feels that it comes down to two positions –

- 1) One may accept that the 'almost' is not enough and even minute differences matter or in fact the little things are the most important ones; or
- 2) One may accept that the minor differences are so minute as to not matter.

Warwick subscribes to the second position. A supporter of the first position may say that even a small change or damage in the neural system leads to huge effect on the mental states or capabilities of an individual. As such even minute differences are important. Another response in support of the position may be given with regards to *qualia*. Qualia is the quality of first-person subjective experiences of an individual. It is usually considered to be very subtle in its

---

<sup>229</sup> Warwick, *In the Mind of the Machine: The Breakthrough In Artificial Intelligence*, 1998, p. 108

<sup>230</sup> *Ibid.*, p 111

manifestation. One may then say that features such as qualia which are very subtle in our mental events and may be looked upon as a minor feature, are in fact the most important and distinct feature of the mental. A supporter of the second position may reply that qualia and such minor features are very subtle and minor features. As such attempts to model artificial brains and thereby minds need not consider them important. As to the question what those basic blocks are, Warwick thinks 'neurons' is the answer. The neuronal level gives us an appropriate level to model the human brain. This can be seen also as an argument against classical AI<sup>231</sup>. So Warwick not only argues against Searle and Penrose, but also the classical approach of AI.

---

<sup>231</sup> Ibid., p. 112.

## Chapter 4

### Analysis of the Relevant Debates and Concepts in AI

We have seen in the previous chapter that both Searle and Warwick give arguments against the view that computers/machines are capable of thinking based on the classical approach to AI. Warwick's argument against classical AI is that it is anthropomorphic. The top-down approach doesn't give much space to the machine's own subjective ways of understanding things. As such he favours the bottom-up approach<sup>232</sup> whereby a machine can learn in terms of its own understanding and awareness of the world around it. Searle, on the other, hand argues that a computer solely defined by its symbols can never think, or in other words, have a mind. Searle's argument is an "in principle" argument. It may be extended even to those AI having bottom-up approaches or for that matter to any form of AI which is based on computers (more specifically digital computers) defined in terms of formal symbol manipulation. Even bottom-up systems may be considered as consisting of mere symbol manipulations for the distinction between top-down and bottom-up approaches are just distinctions in structures and the methods employed. The system is still a computational system for which algorithms are specified. Therefore, the argument may be put up against AI based on such an approach.

What is important to Searle's argument is the understanding or definition of computer. Searle explicitly mentions what definition of computer he is using in his arguments – "It is essential to our conception of a digital computer that its operations can be specified purely formally; that is, we specify the steps in the operation of the computer in terms of abstract symbols"<sup>233</sup>. Whereas in case of Warwick's arguments, he doesn't explicitly give a definition as such and yet, it may be taken as an implication from his arguments, that for him the definition of a computer is such that it allows for the computer to have an "understanding" in its own way. This is because for Warwick, a computer may, given that it has the necessary sensory systems, be aware of external states in the world and therefore, may have some "understanding". But what is the concept of "understanding" that Warwick uses? One thing that Warwick allows for in his concept of "understanding" is "awareness". Just as human understanding involves awareness according to Penrose, likewise for Warwick machines can have awareness too. Sensations can be

---

<sup>232</sup> For top-down and bottom-up approaches see introduction of chapter 3

<sup>233</sup> Searle, John, *Minds, Brains and Science*, Cambridge, Massachusetts: Harvard University Press, 1984, p. 30

generated using sensory devices such as infrared sensors and then the machine can have its own subjective awareness. Based on this awareness the machine can have some kind of understanding of the world around it. To go back to Searle's argument briefly, he says that the mind has both semantics and syntax. The semantic content in a mental state is given by the state of the external world. Searle-in-the-room has to know what the symbols refer to in the outside world to understand or attach meaning to the symbols (see chapter 3, section 1). Now if Warwick accepts Searle's explanation of how "understanding" is dependent on semantics and at the same time makes the claim that machines can have "understanding", it will then imply that a machine or a computer can have more than just syntax. The fact that it understands the world around it implies that it can have semantics. But this contradicts Searle's assertion that a computer can have only syntax. It was mentioned earlier that Searle points out the proposition – syntax is not sufficient for semantics – as a conceptual truth. Further, he considers the essence of a computer to be only syntactical. And thus, for Searle, a computer cannot have semantics by the virtue of having syntax. A comparison may be done between how Searle and Warwick understand computers. It then seems that both Searle and Warwick are working with two different notions of computer. It is known that a computer may have syntax by virtue of computing being an algorithmic procedure. Following the earlier implication that a computer can also have semantics, the question then comes up – how can a computer have any semantics?

## **1. Understanding Computers and Programs**

Searle's argument is based on the definition of a computer as a device which is essentially syntactical. A computer's operations involve mere abstract symbols and rules according to which those symbols are manipulated. Boden provides an alternative understanding of computers whereby they are not concerned only with syntax. In considering a program it is assumed that the program of a computer will not only express representations but also bring about the representational activity of that computer (see Boden's discussion in chapter 3, section 2). Here the definition of a computer, as a device which processes representations in a systematic way, is taken to implicitly contain, within the definition itself, an understanding of its representational activity. It seems that Boden's assertion implies that it is in the understanding of computers (or programs for computers), rather than the understanding of representations, where this tacit implication of bringing about this activity lies.

The above discussion may be slightly re-framed. Following Crane, a computer is a device that processes representations in a systematic way (see chapter 2, section 4). For Searle, these processing of representations are mere symbol manipulation. Representations are taken by him as mere abstract symbols. For Warwick, the processing of representations involves not only syntax but also semantics. If “understanding” requires “awareness” and “awareness” is always awareness of “something” then awareness gives the semantic content. A machine can have awareness and therefore, understanding and for sure, semantics. For Boden, the processing of representations contains within it the notion of bringing about representational activities (see chapter 3, section 2). If one were to accept this understanding of computers rather than the one given by Searle then his argument doesn’t hold. The task then is to make people shift their understanding of computer from the kind that Searle asserts to the one that Boden herself puts forward. For this she appeals to what has been said by Smith that supports her kind of notion of computers (again see chapter 3, section 2 for discussion of Boden and Smith). These are then what have been understood by each of them respectively from the definition of a computer.

A point I would like to make here is that something more can be understood from this definition of a computer that states – a computer is a device that processes representations in a systematic way. The definition contains the term “representation”, and it is in the understanding of this term that the definition implies something more than what has been understood from it till now as mentioned above. One may start by asking the question – What is representation? A simple answer to this may be given as – *a representation is that which represents something*. This is exactly what Tim Crane says the idea of representation is when he says “a representation is something that represents something”<sup>234</sup>. This he says is stating the obvious. He goes on to assert that a representation may even represent itself, but the normal case being, the representation represents another thing. As representing another thing the notion of representation has in it this idea of being about something, that something being the object of representation. And as such, it follows that – *the notion of representation itself has within it this tacit reference to something, something of a kind which is a content or content-like*. This is what distinguishes representations from mere symbols, which does not represent anything but may be used as representations. So coming back to the definition of a computer – a device that processes

---

<sup>234</sup> Crane, Tim, *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*, first published in 1995, second edition, London: Routledge, 2003, p. 11.

representations systematically, it may be understood as processing of not mere abstract symbols as Searle asserts, but processing of symbols that represent something. The symbols may be abstract but as long as they represent, they are representations. This basic understanding of the nature of representation is what I *claim* makes this definition of computer not only about symbol manipulation but also something more. Also if Boden is right in claiming that programs are written such that it is understood as bringing about some representational activity in a computer, then, the point I would like to make is – *this power to bring about representational activity is owing to the nature of representations*. One may think that may be this is what Boden’s claim implies too. However, this is not the case, as she specifically says, “A programmed instruction ... is a procedure specification that ... can cause the procedure in question to be executed”<sup>235</sup>, and also “a fundamental theory of *programs*, and of *computation*, should acknowledge that an essential function of a computer program is to make things happen.”<sup>236</sup> She is taking the very function of a computer program to execute some activities, which includes representational activity as well as representation manipulation. My claim is that *representations by its very own definition represents something and therefore is about something. And it is due to the power*<sup>237</sup> *of representations themselves that this representational activity can be brought about*. So this claim underlies what Boden claims in the understanding of computers that she puts forward.

Earlier we asked – how can a computer have any semantics? Perhaps, it may not be too bold to assert that an answer to this may be given in light of my above claim. Following Crane it was understood that a representation represents something. As such it gives the semantic content. But can this really be considered the same as the semantic that is involved in human understanding? May be or maybe not. Searle would probably deny that mental content can be brought into the computational domain. Perhaps, just having this power to represent does not amount to being equal to actually representing something, the content. As such perhaps it is not at the same level as the semantic content of our mental representations (that is, if mind can be said to be mental representations<sup>238</sup>). This does not contradict or affect the claim that a computer

---

<sup>235</sup> Boden, Margaret A., *The Philosophy of AI*, ed. Margaret A. Boden, Walton Street, Oxford: Oxford University Press, 1990, p. 99.

<sup>236</sup> *Ibid.*, p. 102.

<sup>237</sup> The power refers to the power of representing something.

<sup>238</sup> By a “state of mind” or “mental state”, Crane means something like belief, desire and so on. For him beliefs, desires and so on represent the world, that is they are directed at something; see Crane, *The Mechanical Mind*, 1995, pp. 22-23.

is more than just syntax, as to be more than syntax doesn't mean semantics. What is important is that *representations are not mere abstract symbols*. And as such *processing of representations cannot be equal to mere symbol manipulation*.

One may wonder as to how from the concept of representation does symbols in an actual computer gain this power of representation. This is done because when symbols are used in programs they are done so as representations. There is an underlying assumption that the symbols do represent something. As such the rules are written based on this assumption. In fact the assumption of what a symbol represents or is meant to represent, to some extent, determines the nature of the rules to be written. An example may be given of a program written in C++, which is a programming language.

```
#include <iostream>

int main()
{
int x, y, z;
char a;
string l(50);
cout << "Enter the first letter of your first name.";
cin >> a;
cout << "Enter your last name.";
cin >> l;
cout << "Enter your age and your son's/daughter's age.";
cin >> x >> y;
z = x - y;
cout << a << "." << l << "was" << z << "years old when his/her son/daughter was
born.";
}
```

In the above program, "int", "char" and "string" gives the type of content the respective variable may accept as its value. "int" then may be understood as integer and hence accepts only integers.



“char” stands for character, in this case alphabets, and accepts any one alphabet as its value. “string” stands for a string of letters (may also include numbers) and accepts words of any length. In the above program the length of the string is fixed at 50 letters by writing it within “()” next to “l”. “cout” (to be pronounced as “see-out”) may be understood as “print out on the display/screen”. “cin” (to be pronounced as “see-in”) may be understood as “take in the value entered for a variable”. “x”, “y” and “z” are variables of “int” type, “a” is a variable of “char” type and “l” is a variable of “string” type. Now consider a digital computer with monitors to display and keyboards to take in inputs and run this program. First it will display on the monitor screen “Enter the first letter of your first name.”, to which a user types in “A”. The computer takes the input as the value for variable “a”. Next it displays “Enter your last name.”, to which the user types in “Bachchan” and the computer accepts it as the value for the variable “l”. Next “Enter your age and your son’s/daughter’s age” is displayed on the screen. The user enters “75” and “42”, which are assigned as values to “x” and “y”. The computer then takes the difference of “x” and “y”, that is 75 and 42 in this case, and assigns the value to “z”. The value of “z” is then 33. The computer then displays “A. Bachchan was 33 years old when his/her son/daughter was born”. This program then gives us the age of people when their son/daughter was born.

So in a program symbols such as “x”, “y”, “z”, “a” and “l” does not have any particular content but the type of content they can represent are already fixed by the rules. For instance if one were to enter “Bachhan” or even an integer when the computer asked for the value of “a”, it would not accept the input values as the value of “a” since the values are not of the type that “a” is supposed to represent. So the rules specify what the representations are supposed to represent and how these representations are to be manipulated. It is to be noted that though it seems that the rules determine the representations, it is actually the representation that partly determines the rules. This is because a programmer when writing algorithms already assumes the symbols to be representing something and given this assumption he/she then goes on to specify the procedures to be followed by the computer. For instance, for a given task, say finding a number that is not the sum of three squares of a number, the programmer knows that in this case the symbols are to represent things of a particular type. As such the rules are specified accordingly, which will actually lead to a computer processing it and giving an output of the type required. If the rules didn’t take account of the representation and what it is suppose to be representing then the program may fail to give any relevant outputs. The representations are then in turn based on

actual world objects and things that are required to be represented for a given task. This is where Warwick's notion of "awareness" may come in. The sensory system will provide for what objects are to be represented. But all a programmer needs to know while assigning symbols for representations is what type of inputs a particular representation will represent. And based on the type specified the symbols represent things in the world. Just considered by themselves, these symbols are just inert symbols. But in a program they are no longer inert symbols and therefore are taken to be representations. So a computer is then a representational system whose representations are not mere symbols.

One may disagree with the primacy of representations over rules and also how the symbols get their representative power, but it is not of much importance here to find out what exactly goes on. My point here is simply that if one accepts the definition of a computer as a device that processes representations, then the representations in this definition cannot be mere inert symbols and the manipulations cannot be only of abstract symbols which have no representative power. Also it is not necessary for this assertion that there is actually any semantics involved as pointed out earlier. Having a formal structure of representations might not be sufficient for semantics (see the discussion on the question – how can a computer have semantics – and its preceding paragraphs). Perhaps there might be some semantics involved in the sense that Boden asserts. Even the machine has to understand the rules just as Searle-in-the-room understands the rule book which is in English. My assertion is compatible with both or rather it is a position between two poles. One pole is the assertion that there is no semantics and the other pole is the assertion that there are semantics. So my claim here does not contradict the fact that mere syntax does not give semantics, which Searle asserts. Rather the claim here only deals with the definition of computer as processing representations. It contradicts Searle's notion that a computer is only formal manipulation of abstract symbols. The weight of this contradiction lies on the "abstract symbols". For what is being claimed here is *that the symbols in a computer are not only of the abstract inert kind but actually involves representations which mean more than inert symbols.*

But what does this representation exactly represent? This may lead to another discussion on "representations" which is not the focus of this dissertation. But just to give an idea of what one may further look into so as not leave the answer to the above question hanging, a brief

discussion may be made. Also given that a representation represents, what it may represent can be various things. Its content may be variable. As such the answer may not necessarily hold and is supposed to be considered only as a hypothetical posit, one of the many possibilities. A rejection of it does not reject the earlier claims of this dissertation nor will any arguments be provided in support of it. The important point here is that the question does provide for a problem to further study and investigate. The answer discussed below only serves as one of the various ways to look at it. So, without digressing much some speculation may be made. It is not required that the speculation holds or necessarily follows from earlier claims made in this dissertation. Earlier it had come up that the representations in a computer program represent something that is the content or content-like. Also it had come up that representations represent the type of things that is to be represented. If one were to look among the various philosophical concepts discussed in this dissertation till now, a suitable candidate that fits, up to some extent, the condition of being content-like and to be represented by representations, is that of *forms*. So, one may claim that the representations in computer programs represent “forms”. And it is through forms that they reach out to the particular objects in the world. Form is a predicate that is predicated or said of its subject that is the underlying matter<sup>239</sup>. It is the property which is common to many things. In the Aristotelian notion of form, the form is always as form of some matter. Thus, the representations in being about the forms of things, represents those things through their form. If one were to invoke the Platonic notion of forms as ideas then representations represent ideas of things. The representation then can be manipulated only in ways that those ideas can be manipulated. Particular objects in the world are then just copies of these ideas. Hence, representations can be about ideas and thereby about things in the world. But these are just possible ways to conceptualize representations and their content or object of representation so that a representation in a computer program may have its representational power.

It may be obvious now that an abstract inert symbol considered in isolation does not represent anything. So how can a symbol by itself, as a representation bring about the representational power of itself? The thing about symbols in a computer program is that they are not present as inert or as a structure of symbols in isolation, everything is specified. For instance

---

<sup>239</sup> One may refer back to chapter 2, section 2.

in the earlier mentioned program the value of “x”, “a” and “l” may be taken by the computer in terms of 1s and 0s. And even though the value of “x” which was 75 and value of “a” which is “A”, both may be in the form of 1s and 0s (the most basic level of machine-code), they cannot just be processed together as values of the same type, for they are of different types. If one were to try and give the values of “x” and “a” as inputs to an addition function, the function won’t be carried out even though both are in terms of 1s and 0s. As such the rules take as much part in determining what a symbol represents. That’s why when rules are being written, *the nature of the object of representation has to be pre-assumed*. This assumption is then what a programmer takes the basis for setting the “type” of a symbol that is to represent. The example used was of the classical approach kind. But as far as representations are concerned even the bottom-up approach kind such as the artificial neural network consists of representations. The representations may be not as symbols but as states of a neural net, and therefore the same can be said of such systems.

But does the above example of a C++ program actually counter the Chinese room argument? An analogy between the program and the Chinese room example may be readily drawn. The Chinese room example involved operations such as taking in tiles of symbols, operating on them by following the actions specified by the rules and then handing out tiles as specified by the rules. In the above program too, inputs are taken in (which is done by the “cin” command), inputs are operated on according to the relevant rules and then outputs are displayed (which is done by the “cout” command) as specified by the rules. As such both have the basic Turing machine settings. What then may be observed is that in the Chinese room argument Searle does not consider the symbols as representations, but as mere abstract symbols. On the contrary, in the C++ program each symbol is specified according to the type of values they may represent. This is precisely because those symbols are being considered as representations and not just as abstract and inert. Hence, Searle does not do justice to the definition of a computer, but more specifically to representations by reducing them to mere symbols.

So how do the claims made earlier along with the above comparison between C++ program example and the Chinese room argument affect Searle’s thesis? Searle’s thesis can be put together as follows<sup>240</sup>:

---

<sup>240</sup> Searle, *Minds, Brains And Science*, 1984, pp. 38-41.

1. Brain causes minds (the mental processes that constitute a mind are entirely caused by processes going on inside the brain).
2. Syntax is not sufficient for semantics. This is a conceptual truth<sup>241</sup>.
3. Computer programs are entirely defined by their formal or syntactical structure.
4. Minds have mental contents, specifically, they have semantic contents. One's thoughts and beliefs are about something, or they refer to something, or they concern states of affairs in the world. The contents direct them at these states of affairs in the world.

From the above premises the following conclusions are drawn:

1. No computer program by itself is sufficient to give a system a mind. (based on premises 2, 3 and 4)<sup>242</sup>
2. The way that brain functions cause minds cannot be solely in virtue of running a computer program. (this follows when premise 1 is conjoined with conclusion 1)
3. Anything else that caused minds would have to have causal powers at least equivalent to those of the brain. (based on premise 1, because if brains cause minds, then, anything else that may cause mind should at least have causal powers equal to the brain. This is a trivial consequence and hence of not much import here.)
4. For any artefact that might be built which had mental states equivalent to human mental states, the implementation of a computer program would not by itself be sufficient. (follows from conclusion 1 and conclusion 3)

My earlier claim, that the definition of a computer cannot be understood as manipulation of mere inert symbols but rather as representations, then affects only premise 3 of Searle's argument. Searle may be partly right, that a computer may be defined by its formal or syntactical structure. However, I have argued for and shown earlier that the structure and the rules are themselves partly determined by the representations involved within that structure. So a computer may be defined by its formal or syntactical structure but that very formal/syntactical structure is itself partly determined by the representations involved within that structure. But this is neither to claim that computers have semantics involved. For then it will mean that syntax is sufficient for semantics which contradicts the conceptual truth given in premise 2, that syntax is not sufficient

---

<sup>241</sup> See chapter 3, section 1; see Searle, 1984, p. 39.

<sup>242</sup> Searle, 1984, p. 39-40.

for semantics. Since the contradiction just stated above has not been made, conclusion 1 which follows from premise 2, 3 and 4 may still hold with a slight adjustment to premise 3. It may then be stated as, “Computers are fundamentally or essentially defined in terms of representations which does not necessarily give semantics and certainly not in case of computers.” So Searle’s conclusions are more or less intact. What then the above discussions manage to show is that – *if one were to accept Searle’s premises and conclusions then one will also have to accept that the directedness towards something cannot be only due to semantics. There can also be directedness with no “full-fledged” or “full-blooded” (if one were to borrow Boden’s words<sup>243</sup>) semantics involved.* But if one were to accept that Searle’s view of computer programs as all syntax and no semantics is a mistaken view<sup>244</sup>, then, the representations in a computer program, as being about something, can be considered to have semantics. Hence, my claims are compatible with Boden’s view.

## **2. What is the notion of “Understanding” in Humans and Machines?**

It has been discussed above that representations represent something. But is there any understanding involved? Searle denies any understanding to a digital computer. He does that on two fronts. First, understanding requires semantics. Syntax is not sufficient for semantics. A digital computer is all about syntax. Therefore, it has no semantics and hence, no understanding. Second, human brain processes have semantics. For Searle, brain processes both cause and realize mental states and processes. The causation and realisation of mind are explained as biological phenomena caused by and realized in a biological entity, the brain. Mental states and processes have contents, specifically semantic contents as thoughts and beliefs are about or refer to something or they are concerned with the states of affairs in the world. Thus, semantic contents of the mind are due to the biological processes and as such may be considered as biological. “Understanding” itself as a feature of mind has to be considered as caused by and realised in the brain and therefore biological. So if “understanding” involves semantic content then it has to be in terms of the biological phenomenon that “understanding” is. This then leads to the conclusion – semantics and thereby “understanding” are both biological features. Computers are not biological. Therefore, computers don’t have understanding. Warwick, on the

---

<sup>243</sup> Boden, “Escaping from the Chinese Room”, *The Philosophy of Artificial Intelligence*, 1990, p. 97.

<sup>244</sup> *Ibid.*, p. 102.

other hand, argues that computers can understand. He says, "... a computer may well understand the machine code"<sup>245</sup>. He says this in the context of Chinese room argument where machine is taken to be not capable of understanding. However, he shows the biasness in the Chinese room argument and argues back that a human placed in the room and given machine codes might also not understand the symbols just as how machines don't understand Chinese (for more on this see chapter 3, section 4). Machine can at the least understand machine codes. Dogs can understand, cats can understand, basically every other species can understand, so why not machines too. Of course, one has to first understand "understanding" as a subjective faculty. In humans, they understand things around them through awareness. This awareness boils down to sensations, human specific sensations. In cats, dogs, bats and so on, awareness occurs according to their sensory specifications. A bat is aware of its surrounding through its sonar or echolocation system. A machine then can be given its own sonar or infrared systems and be made aware of its surroundings. The input from the sensors may be converted into machine codes and the machine will be aware of the surrounding. It will then have its own subjective understanding of the world around it just as how a bat has its own subjective understanding. Warwick in asserting that computers have "understanding" and interact with the world through sensory systems, would surely attribute semantics to the computers. But a question may come up – Why so? Why attributing "understanding" leads to attributing semantics? Is there a necessary relation between "understanding" and semantics? One may say that if we attribute semantics then one has to necessarily attribute "understanding". But is it the case that if we attribute understanding then it necessary follows that there is some semantic content? This is an issue that may require further research but for the present purpose a brief reply may be given. Warwick draws from Penrose's argument that "understanding" requires "awareness". So according to Warwick "understanding" requires "awareness" and "awareness" may be based on sensory systems. Now, a sensory system represents the world, it provides for the content of the "awareness" and hence it provides for the semantic content. So Warwick's claim that a computer has "understanding", surely and strongly seems to suggest that it has semantic content too.

For Penrose it is a different story. There is something about genuine understanding that makes it beyond computation (that is, non-computational). So there might be awareness (in a

---

<sup>245</sup> Warwick, Kevin, *Artificial Intelligence: The Basics*, Abingdon, Oxon: Routledge, 2012, p. 74.

computer or a species) through sensory systems, but awareness has no necessary causal connection with “understanding”. Conceptually “understanding” pre-assumes awareness and hence, if there is “understanding”, there will then necessarily be awareness. But this is an asymmetric relation. The converse is not necessarily true. It may be the case that if there is awareness, there need not be “understanding”. So a computer may have a sensory system and thereby some content about the world, but yet it still may not have “understanding”. While awareness in this simple sense may be attributed by Warwick to a computer, Penrose will refuse even that. He is in fact of the viewpoint that “Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally”,<sup>246</sup>. Through Gödel’s argument he shows that “understanding”, “awareness” and consequently “consciousness” itself is non-computational. So one may say all that has been said regarding understanding, awareness, semantics, sensory system and so on for human cognition. But according to Penrose, it simply doesn’t happen for computers. The computer may have sensory inputs, it may receive them and operate on them, but this whole process in itself is of a different type, the type that is not actually “awareness” in the sense it is used for humans, birds and so on. The sensory apparatus or system is built based on the scientific knowledge of world. So, for instance, infrared devices work based on some scientific understanding of light. But it has been pointed out earlier in chapter 2 (section 3) in the example as given by Frank Jackson, that one may have all the scientific knowledge of say colour ‘red’ but upon actually seeing red, she comes to know something new that she hadn’t known earlier despite having all the physical knowledge of that particular spectrum of light. So, if one accepts that the current scientific picture of the world has a missing ingredient (to use Penrose’s word), namely, that which explains consciousness and like phenomena, then building a sensory input system based on that knowledge will surely be lacking something too. Hence, awareness of this kind will not be the kind that involves consciousness. But if it is not consciousness then will it be called “awareness” at all? This may be another point of debate perhaps, but one that will be avoided here lest it may lead to digression. The initial question was – is there any understanding involved in computers defined as processing of representations? For Penrose it is a clear denial.

---

<sup>246</sup> Penrose, Roger. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press, 1994, p. 12.



My claim was that representations in computers, by virtue of being representations represented something. But this didn't amount to giving a semantic content. So "understanding" that arises due to semantic content will surely be not possible in a computer. What about "understanding" that doesn't require semantics, if there is such a kind? If there is, and I think there is (will be discussed shortly), then such a kind of "understanding" must be attributable to computers. Thomas Nagel gives an account of the subjective character of experience for an organism<sup>247</sup>. This subjective character of experience is species specific. Thus, members of radically different species would never understand what it is like to be a member of another species. A human would never understand what it is like to *be* a bat. This is because, Nagel points out, the subjective character of experiences is not part of a common reality that is shared inter-species. Warwick considers machines as a species (in a broader sense)<sup>248</sup>. So, for instance, it may be said that a computer understands what it is like to be a computer, just as humans understand what it is like to be a human but a human would not understand what it is like to be a computer and vice-versa as the subjective experiences of both species does not share a common reality. In fact this is what Warwick seems to suggest, though he does more than that. He suggests that a computer (based on bottom-up approach) understands the world in its own subjective way.

So here is why I think there is a kind of "understanding" that doesn't require semantics. The subjective character of experience of an organism is not something external in the world and thus, an understanding of it is not about something external in the world either. Nor is it about any particular experience which can serve as the object of understanding, rather the understanding is present in all subjective experiences of an organism as subjective experiences for that organism. Hence, such an understanding has no semantic content. Perhaps then, this is the sort of understanding that a computer may have, if semantics is denied to it. But to assert this there is a problem. Nagel had pointed out that members of radically different species will never know what it is like to *be* a member of another species. This is due to the fact that the subjective characters of experiences of different species do not share a common reality. But there is a pre-assumption here. The bats, the humans and most of other species are already assumed to have

---

<sup>247</sup> Nagel, Thomas, "What is it like to Be a Bat?", *The Philosophical Review*, vol. 83 no. 4 (1974), p. 435-450.

<sup>248</sup> Warwick, 2012, p. 19; also for a similar position by Warwick that comparison between abilities of individuals from different species is meaningless, see Warwick, 2012, pp. 28-29.

subjective experiences which are to be accounted for. In fact it is the case that the science of today, does not provide a proper account of it precisely because the kind of reduction they make, walks away from the very thing they attempted to explain in the first place. Now computationalism as being a form of functionalism reduces mental states to functions which rip them off its subjective character. Further, if these functions are then taken and put (as algorithm) in a system, of which it is already known that it has no consciousness or mental states, then it is doubtful that these functions will in this new system get back that subjective character that was ripped off of them. Perhaps, it may be argued that in this case the subjective character is already in the new system. The system being whatever it *is* is unique in its own way and has its own subjectivity. One may argue that no two objects occupy the same space at the same time. An object is unique in its position and perspective from all other objects. This is a statement about physical position of the object. If something can be unique in its physical parameters, then it must surely be unique in its subjective parameters. So a system might already have a subjective character of its own. Will then the function in this new system get back the subjective character that was ripped off of them? One may argue that it will. I argue that it will not. Does the subjective character of this system appear so to the system itself? Does the system realize the subjective character of its experience? This basically amounts to pre-assuming consciousness in the system. So while it may be possible that the functions, now as algorithm, when put into a conscious system may re-gain a subjective character from the new system, it seems not at all possible that functions when put in a non-conscious system will have any such subjective character by virtue of being functions. Thus, a computer defined as processing of representations cannot be attributed the kind of “understanding” that has no semantic content. In fact such an understanding has to precede the understanding that consists of semantic content.

What has been argued in this section is that a computer does not have any “understanding”, genuine or not. But what about intelligence, can a computer be intelligent despite what has been asserted till now?

### **3. Understanding Intelligence as a Notion**

Can a computer ever be intelligent? For Warwick, a computer can be intelligent, that too in its own way, just as it can have understanding and awareness. Searle would perhaps allow the computer to be intelligent to the extent that it is clear that this intelligence is merely a simulation.

For Penrose, intelligence requires “understanding”. Since “understanding” is non-computational and computers, therefore, can’t have understanding, they can neither be truly intelligent. Penrose points out that some AI supporters might claim that intelligence can be possible without a computer actually understanding anything. This, Penrose says is owing to the inaccurate use of the term intelligence. So what then is “intelligence”?

Newell and Simon described intelligence as a composite, for no single thing accounts for intelligence in all its manifestation. There simply is no intelligence principle. One requirement though for intelligence, according to them, is the ability to store and manipulate symbols. Thus, they proposed that a physical-symbol system<sup>249</sup> has the necessary and sufficient means for general intelligent actions. Computers being such systems therefore can be intelligent. For them if a system is intelligent then it necessarily is a physical symbol system, even humans. Searle’s criticism is precisely against this kind of position. For a human brain might be considered as a computer (a biological one), but a computer will never be considered equal to a human brain owing to the different kinds of stuff they are made up of. Searle still allows that a computer can simulate being intelligent, but not “truly” in the sense that humans are. For instance, Searle says,

“The question is: ‘Can a digital computer think?’... From a mathematical point of view anything whatever can be described as if it were a digital computer. And that’s because it can be described as instantiating or implementing a computer program. In an utterly trivial sense, the pen that is on the desk in front of me can be described as a digital computer. It just happens to have a very boring computer program... Now since in this sense, anything whatever can be described as implementing a computer program, then once again, our question gets a trivial answer. Of course our brains are digital computers, since they implement any number of computer programs. And of course our brains can think. So once again there is a trivial answer to the question.”<sup>250</sup>

For Searle, brains can think and as far as brains are considered as digital computers, a computer (as defined earlier by Searle) can think and thereby be intelligent. But this thinking or being intelligent is not the same as what humans do or what humans are. Computers cannot be truly intelligent in the sense that humans are. This is because they lack consciousness, intentionality

---

<sup>249</sup> For more on physical-symbols system refer to chapter 1, section 2.

<sup>250</sup> Searle, 1984, p. 36.

and so on which in a human brain is due to its bio-chemical composition and processes. Searle says,

“Conclusion 2: *The way that brain functions cause minds cannot be solely in virtue of running a computer program...* But the importance of this conclusion is that the computational properties of the brain are simply not enough to explain its functioning to produce mental states... all it does is remind us of the fact that brains are biological engines; their biology matters.”<sup>251</sup>

He also says,

“The upshot of this discussion I believe is to remind us of something that we have known all along: namely mental states are biological phenomena. Consciousness, intentionality, subjectivity and mental causation are all a part of our biological life history, along with growth, reproduction, the secretion of bile and digestion.”<sup>252</sup>

This implies that to be truly intelligent, a physical-symbol system needs to be made up of some biological stuff or at least some chemical or biochemical constituents that have causal powers equal to the human brain. If this is the case then the essential ingredients/constituents/components/recipe for true intelligence would be consciousness, intentionality, and so on which are further constituted of biological processes and realised in a biological system made up of biological stuffs.

Warwick, on the other hand, says intelligence is not dependent on the brain alone. It depends also on how the brain senses and activates things in the world. As such intelligence is also dependent on an individual’s senses and actuators such as limbs or muscles<sup>253</sup>. Given that intelligence depends on these factors, it is also subjective as individuals from different species have different senses and different ways of movement and activating things in the world. The make-up of their brain might also differ. Besides these factors another factor that plays a role in determining or affecting intelligence is one’s environment. An individual educated in a particular way might be intelligent in a different way from an individual given another kind of education.

---

<sup>251</sup> Searle, 1984, p. 40.

<sup>252</sup> Ibid., p. 41.

<sup>253</sup> Warwick, 2012, p. 16.

Also an individual brought up in a particular culture might be intelligent in a different way from an individual belonging to a different culture. This is because different groups, different cultures, different educational systems might consider different qualities as being intelligent. Also among different species, different actions may be considered intelligent based on species specific needs to survive. Thus, on one hand, is the natural make-up of individuals such as brain, sensory organs and muscles, while on the other hand, is the nurture aspects which include experience, education, species-specific requirements and so on. But how much of an individual's intelligence does depend on nature and how much on nurture? What percentage is inherited and what percentage is due to environmental effects and experience or learning? What is the recipe for intelligence? Warwick asserts that based on these factors a computer can also be intelligent. A computer (of the bottom-up approach kind) may be provided with some essential programs, sensory systems and perhaps a robot body to interact with the world. Besides these, the computer can then learn from its interaction with the world and even education may be provided to it. This way a machine can be intelligent in its own subjective way as its initial program, sensory system, actuators, and experience are different from other species such as humans, bats, mouse and so on.

It may be noted that in the above discussion intelligence is considered as some kind of a composite mental state that is a product rather than a quality or an attribute. This mental state is then considered to be constituted of raw materials (may be other mental states) such as understanding, awareness in case of Penrose, consciousness, intentionality and so on in case of Searle<sup>254</sup> and brain, sensory systems, experience and so on in case of Warwick. While it is taken only as implied in case of Penrose and Searle, it is explicitly mentioned in case of Warwick. Warwick in fact draws an analogy of intelligence with a cake. He says,

“The question is perhaps similar to asking: when baking a cake, how much of the quality of the cake is down to the original mix of ingredients and how much is down to how it is cooked? For the cake we would see both aspects as being important and the overall quality as being a subtle mixture of the two. Sometimes the cake might come out really well despite it not being left in the oven for an ‘optimal’ time, whereas sometimes it can

---

<sup>254</sup> It is to be noted that these statements regarding Penrose and Searle does not mean that they don't think experience and other factors do not play any role in an individual's intelligence. Given that they have said that these are necessarily required for intelligence, I have mentioned them in such order.

turn out badly even when the recipe had been followed to the letter. What people have been looking for, for thousands of years (certainly back to Plato in third-century BC Greece), is the recipe for intelligence.”<sup>255</sup>

One may as well say, “While creating intelligence, how much of the quality of intelligence depends on...?” It may immediately be noted in this question the awkwardness of the term “quality of intelligence”. Can there be any quality of intelligence? Can there be, for instance, a “bad” quality intelligence just as there can be a “bad” quality cake? Or if one says “intelligent action”, can it be also said “bad intelligent action”? Here “bad” is not supposed to mean bad in an ethical or a moral sense. Perhaps, it makes sense for an intelligence officer (or law enforcement agents) to say that s/he was provided with bad “intel”<sup>256</sup> or bad intelligence. But here intelligence merely means information. To actually say “bad intelligence” in the normal sense creates some kind of confusion. This is precisely because intelligence itself is a quality attributed to other things, namely, actions and behaviours. Both “bad” and “intelligent” are adjectives. As such even if they are used together, they only affect the noun or verb they are connected to. So to say “bad intelligent action” may actually mean “bad action” and “intelligent action” where both are about the same action, and not that the quality of “intelligence” is bad. For if there was lacking in the action, then the quality will be of “lesser intelligence” and not “bad intelligence”.

To say human beings are intelligent is to attribute them the quality of intelligence. What does this attribution mean? One answer to this may be that when humans are attributed intelligence then it just means that they have the ability to acquire and apply knowledge. But this ability is merely a cognitive capability and as such means that a human can act in particular ways, that is, act so as to gather, analyse, infer and apply knowledge. This attribution refers to some actions (cognitive acts) which are deemed to be intelligent and are not considered as intelligence itself. It may be observed that the above answer correlates knowledge and intelligence. In fact many seem to think that this correlation with knowledge is essential to the definition of intelligence. George Butterworth points out three criteria that seem essential to the definition of intelligence<sup>257</sup>. Out of the three, the first criterion states – “Intelligence cannot be

---

<sup>255</sup> Warwick, 2012, p. 24

<sup>256</sup> “Intel” is short for intelligence. The term is often used in spy-movies.

<sup>257</sup> Out of three only one is stated as the other two are not important to the purpose here.

well understood without reference to the internal representation of knowledge. Using a knowledge base to decide what information is relevant and what irrelevant is a feature of intelligent behaviour. One way to understand intelligence therefore may be by reference to the psychological processes which give rise to knowledge.”<sup>258</sup> This is, perhaps, what Warwick thinks intelligence is and that’s why he favours AI capable of learning rather than the classical ones which only had pre-specified principles and pre-stored knowledge on which to act upon.

Richard Gregory describes descriptions of intelligence as beset with paradox. “Thus intelligence is attributed to those who have to think because they do not know a lot, and to those who know a lot and so do not have to think.”<sup>259</sup> He resolves this paradox by positing two kinds of intelligence<sup>260</sup>: intelligence of stored knowledge; and intelligence of processing, for problem solving. These he then terms potential intelligence and kinetic intelligence, respectively, and though they are distinct concepts, they are often found intermixed together. The two concepts are termed so to suggest an analogy to potential and kinetic energy. Potential intelligence may be found in brains or minds, tools, books and so on. Tools, for instance, Scissors were developed as a result of generations of kinetic problem-solving intelligence to produce the potential intelligence which is built into the design of the scissors. As such, one no longer needs to apply kinetic intelligence to solve the problem of cutting a piece of cloth into a particular design. One may instead use the potential intelligence already in the form of scissors to solve this particular problem. Here also the essential reference is to knowledge. Stored knowledge is considered as potential intelligence; while intelligence of the processes, which is supposed to supply new knowledge or use the old knowledge in new problems and new ways, is considered as kinetic intelligence. The processes themselves are not intelligence, but rather intelligence is the result of processes<sup>261</sup> just as potential energy is the result of kinetic energy. So to say humans are intelligent will mean that humans both stores knowledge and processes knowledge (while gaining or applying knowledge).

---

<sup>258</sup> Butterworth, George, “Infant Intelligence”, *What is Intelligence?* ed. Jean Khalfa, Cambridge: Cambridge University Press, 1994, p. 50.

<sup>259</sup> Gregory, Richard, “Seeing intelligence”, *What is Intelligence?* ed. Jean Khalfa, Cambridge: Cambridge University Press, 1994, p. 13.

<sup>260</sup> *Ibid.*, p. 14.

<sup>261</sup> *Ibid.*, p. 15.

Roger Schank and Lawrence Birnbaum put this kind of view in very simple terms. “What makes someone intelligent is what he knows. What is needed to make intelligent computers is to endow them with knowledge.”<sup>262</sup> For them intelligence simply cannot be about something innate like consciousness as posited by Searle. If this was the case then the whole point of education is useless. Intelligence is additive and therefore, education aims at enhancing intelligence. If Searle is right in saying that intelligence is owing to consciousness and so on which further depends on the biological makeup of the brain, then it will imply that no matter how much education is given to an individual, he/she will not be made more intelligent. Also it doesn’t make sense to say that one will become more intelligent by becoming more conscious.

In the above discussion, two kinds of definition of intelligence may be observed. One definition takes intelligence just to be something in the mind or brain. For instance, Searle and Penrose, who correlates intelligence with consciousness, brain processes and so on, may be taken as providing that kind of understanding of intelligence. On the other hand, there are those who define intelligence as a product of processes in correlation to knowledge. Intelligence as a product is therefore taken to be an entity or perhaps a more appropriate term will be quantitative property, just like energy. Warwick’s definition of intelligence takes specific processes as constituting intelligence. He defines intelligence as: “the variety of information-processing processes that collectively enable a being to autonomously pursue its survival.”<sup>263</sup> These processes may be in any domain such as communication, problem-solving but in Warwick’s case, specifically survival. Perhaps, all else can be taken as a result of adaptation to survive. In both kinds of definitions intelligence is taken to be an entity and a complex, composite entity at that. In the second kind of definitions, intelligence might be considered an entity only as an abstraction but they surely do consider intelligence to be a composite. My argument against such definitions is two-fold. Firstly, intelligence is not an entity and secondly, it is not a composite. Both will be discussed below. Together, as a conjunction, my assertion is that intelligence is not a composite entity.

Dan Sperber says, “... intelligence, like beauty, is a property rather than a thing. There is no area of a brain that might properly be called its intelligence. On the other hand, a variety of

---

<sup>262</sup> Schank, Roger and Lawrence Birnbaum, “Enhancing Intelligence”, *What is Intelligence?*, ed. Jean Khalfa, Cambridge: Cambridge University Press, 1994, p. 82.

<sup>263</sup> Warwick, 2012, p. 17.



doings and doers can be called intelligent.”<sup>264</sup> My claim, like Sperber, is that intelligence is a property (a qualitative one) rather than a thing. I agree (because I have myself claimed so in the discussion of Warwick’s analogy between intelligence and cake) with his statement that doings and doers can be intelligent. But unlike him, my claim will be that intelligence, though a property is more analogous with something else rather than beauty. This is not to say that I am arguing against using beauty as an example by Sperber. The analogy is fine, as it makes the claim that intelligence is a property like beauty. What I am saying here is that there can be another analogy with another concept which still holds the claim that intelligence is a property and also analysable in the same way as that concept. This will be discussed below. The fact that intelligence is not an entity is quite commonly understood among psychologists, who in fact often uses tests to investigate intelligence by treating it as an entity. But they are clear that this is only an abstraction and intelligence is not considered a real entity. W.B. Dockrell writes in the introduction of *On Intelligence*: “The different concepts of intelligence held by the participants in the symposium minimize the danger of accepting any one point of view about intelligence as correct. There remains the danger of unconsciously reifying the concept of intelligence and treating it as though it were an entity and not merely ‘a convenient manner of speech’.”<sup>265</sup> It was pointed out earlier in the analogy of cake and intelligence, that to use the concept of intelligence as an entity like cake rather than a quality does not give much sense or is a misnomer. The discussion on this analogy with cake had shown that intelligence taken as an entity does give rise to some problem as pointed out in the discussion. Thus, the notion of intelligence cannot be as that of being an entity. But why should it be a quality either? Does considering it a quality solve the problem that comes up with its notion as an entity? Given the discussion earlier, I would say yes. But, should there be any apprehension regarding this, a further analysis may be carried out to show how intelligence as a quality makes more sense than intelligence as an entity.

Intelligence is a quality that is attributed to doings and doers, that is, actions and actors of those actions. The doers or the actors are attributed the quality precisely because they perform actions that are considered to be intelligent. To say that an action is intelligent, is not to draw an identity relation between the action and intelligent/intelligence. It is not to say that action =

---

<sup>264</sup> Sperber, Dan, “Understanding Verbal Understanding”, *What is Intelligence*, ed. Jean Khalifa, Cambridge: Cambridge University Press, 1994, p. 179.

<sup>265</sup> Dockrell, W. B., *On Intelligence: The Toronto Symposium On Intelligence, 1969*, ed. W. B. Dockrell, London: Methuen & Co Ltd, 1970, p. 5

intelligent. Similarly, in case of doers/actors, when it is uttered, “She is intelligent”, it doesn’t mean “she (the doer)” = intelligent/intelligence. It is only attributing a quality to a person. So is the case with information-processing processes (from Warwick’s definition of intelligence) and the result of any processes (as asserted by Gregory). Intelligence can only be a quality of doings, doers, processes and the results of processes. At this point an analogy may be drawn between the concept of intelligence and the concept of good.

Another concept that applies to actions is that of “good” or “goodness”. The term good is a quality. One may say, “That is a good mango” or “Out of the goodness of one’s heart...” or “You played good”. In all these cases “good” or “goodness” is a quality. Like “intelligence”, the concept of “good” is that of being a quality and as such an analogy may be drawn between them. The term “good”, in ethics, is applied to actions. Ethics deals with the study of what one ought to do. So to say one ought to do what is good, is to attribute the quality of good to an action and say one ought to do this particular action. The action has the quality of good and thereby the one performing it has as his/her quality goodness. Good, then may be said, is a quality that is attributable to both doings and doers. Similarly, it has been argued earlier, that intelligence is attributed to actions, where actions may include thinking, processing and so on. My claim here then is that the concept of intelligence is similar to the concept of good. I would further claim that due to the analogy between them, a similar analysis as that of the concept of good is also possible for the concept of intelligence. As such, my next claim is that, the analysis of the concept of intelligence may be carried out just like the analysis of the concept of good given by G. E. Moore. Moore says,

“‘Good’, then if we mean by it that quality which ... belong to a thing, when we say that the thing is good, is incapable of any definition, .... The ... sense of ‘definition’ is that in which a definition states what are the parts which invariably compose a certain whole... ‘good’ has no definition because it is simple and has no parts”<sup>266</sup>.

It is of importance here the sense of ‘definition’ that Moore uses. Definition often is taken in the sense that when one defines, say a cow, the definition enumerates what a cow is composed of and in what manner – two horns, four legs, a tail and so on, all of them arranged in some relation

---

<sup>266</sup> Moore, G. E., *Principia Ethica*, Cambridge: Cambridge University Press, 1903, p. 9

to each other. So defining then means decomposing the object to its parts. But such decomposition can only be done if there are parts, that is, the object is a complex whole. It is not to claim here that there are no other senses of definition. Moore does consider other senses of the term “definition”. He does so when he says,

“When we say, as Webster says, ‘The definition of horse is “a hoofed quadruped of the genus Equus,”’ we may, in fact, mean three different things. (1) We may mean merely ‘When I say “horse,” you are to understand that I am talking about a hoofed quadruped of the genus Equus.’ This might be called arbitrary verbal definition: and I do not mean good is indefinable in that sense. (2) We may mean, as Webster ought to mean: ‘When most English people say “horse,” they mean a hoofed quadruped of the genus Equus.’ This may be called the verbal definition proper, and I do not say that good is indefinable in this sense either; for it is certainly possible to discover how people use a word: otherwise, we could never have known that ‘good’ may be translated by ‘gut’ in German and by ‘bon’ in French. But (3) we may, when we define horse, mean something much more important. We may mean that a certain object, which we all of us know, is composed in a certain manner: that is has four legs... a liver, etc., etc., all of them arranged in definite relations to one another. It is in this sense that I deny good to be definable. I say that it is not composed of any parts, which we can substitute for it in our minds when we are thinking of it.”<sup>267</sup>

It is in this sense, following Moore, that I take “definition” to mean the decomposition of a complex whole to its parts. But one may as well wonder why that particular definition? Once again a reference to Moore might help answer this question. Moore explains,

“What, then, is good? How is good to be defined?... A definition does indeed often mean the expressing of one word’s meaning in other words. But this is not the sort of definition I am asking for. Such definition can never be of ultimate importance to any study except lexicography. If I wanted that kind of definition I should have to consider in the first place how people generally used the word ‘good’; but my business is not with its proper usage, as established by custom... I shall, therefore, use the word in the sense in which I

---

<sup>267</sup> Ibid., p. 8

think it is ordinarily used; but at the same time I am not anxious to discuss whether I am right in thinking it is so used. My business is solely with that object or idea, which I hold, rightly or wrongly, that the word is generally used to stand for. What I want to discover is the nature of that object or idea, and about this I am extremely anxious to arrive at an agreement.”<sup>268</sup>

Following Moore, my intention too, when the question is raised – what is intelligence or how is intelligence? – is only analyse the idea of intelligence. The reason why I do so may be best expressed in G. Thomson’s words which he uses in the context of ‘mental energy’ (it is not necessary here to explain what is meant by ‘mental energy’, one may just read the quotes in the context of intelligence). He says, “With the bulk of those studying science there exists always the danger that this may be taken too literally... the danger of “reifying” such terms... is however, very great...”<sup>269</sup> Dockrell does, in fact, take this quote in the context of intelligence. The earlier quote by Dockrell was made by him following Thomson.

Now to proceed with the analogy, Moore takes good/goodness to be a simple notion and hence, not definable. But how is good a simple? Moore explains that propositions about “the good” are all synthetic statements and never analytic. “The good” is the thing which has good/goodness as its quality. So in the statement, “Pleasure is the only good”, “the only good” is the thing that has its quality goodness and pleasure is asserted to be that thing. The statement, according to Moore, is not analytic, the statement does not give the meaning of the word “good”. Analytic statements are those whose truth depends upon the meanings of its constituent terms, for instance, “Bachelors are unmarried”. As such analytic statements are those where the relation between the subject and predicate is drawn based upon the meanings of the constituent terms. Synthetic statements are those whose truth depends upon the facts about the world, which is known through experience. As such, the relation between the constituent terms is not based on their meanings, but on facts or state of affairs in the world. So, synthetic statements will be of the kind – “Pleasure is good”, “Happiness is good”, “The mango is good” and so on. This implies that good has to be understood only as “good” and not as some other terms. Also Moore explains that simple terms are “... simply something which you think of or perceive, and to any one who

---

<sup>268</sup> Ibid., p. 6

<sup>269</sup> Thomson, G., *The Factorial Analysis of Human Ability*, London: University of London Press, 1950, p. 251.

cannot think of or perceive them, you can never, by definition, make their nature known.”<sup>270</sup>  
Good then is a simple notion and hence cannot be defined.

A similar case can be made for intelligence. The reason why many feel intelligence is a complex, multifaceted thing or entity is precisely because they feel that its meaning can be understood in terms of some constituent parts, such as communicating, thinking, processing, desiring, intending, being conscious and various other actions. But then it may be realized that statements such as “He is the intelligent”, which will be referred to as statements of “the intelligent”, like statements such as “Pleasure is good”, which Moore refers to as statements of “the good”, are all synthetic. “The intelligent” here again means the thing which has intelligence as its quality (more on this below). Also intelligence is such a term which, to anyone who cannot think of it or perceive it, can never be made to know its nature, simply by definition. Imagine for instance that one has to explain an AI computer what intelligence is. How will one do it? It seems almost impossible to define it for the computer given that computers follow strict rules, and telling it “This is intelligent” while pointing at a game of chess or rules of logic, will only lead the computer to equate the game of chess or rules of logic to the term “intelligent”. Or if one were to say instead “This is the concept of intelligent/intelligence” while pointing again at the game of chess or rules of logic, the computer will only equate the game of chess or rules of logic with the concept of intelligent/intelligence and when asked “what is intelligent or intelligence?”, the computer might give the answer, “The game of chess is intelligent/intelligence” or “The rules of logic are intelligent/intelligence”. One might also say that a computer (of the bottom-up approach kind) can extract features from cases which are marked as intelligent and then have a kind of learning about intelligent acts. But how will it know what the features are being marked as? Suppose in the game of chess, it may learn which moves to make when faced with a given situation. But how will it know those features are intelligent? Such marking just give rise to synthetic relations and it does nothing to explain what intelligence is to the computer. It is precisely why if that very same computer was to be given a set of rules of logic (supposing rules of logical are intelligent rules), it won’t be able to identify them as intelligent rules based upon its previous learning from the game of chess. Explaining the quality of intelligence to the computer might perhaps be too much to ask. One might even ask the question: “Do Chimpanzees

---

<sup>270</sup> Moore, 1903, p. 7.

ever think that humans are more intelligent than them?” or “Do they ever wonder if monkeys can be as intelligent as them?” As such, my claim then is that intelligence is a simple notion and cannot be defined.

Confusion may arise regarding to the term “the intelligent”. One may point out that earlier I had claimed that intelligence is a quality and not an entity or thing. How can I then again claim that “the intelligent” is a thing? Again Moore’s explanation of “the good” might be considered. According to him “good” is an adjective. “The good”, he claims, is that which is good, in the sense that “the good” has the quality of good, and therefore is the substantive to which the adjective good will apply. He explains “But if it is that to which the adjective will apply, it must be something different from the adjective itself; and the whole of that something different, whatever it is, will be our definition of *the* good.”<sup>271</sup> He also points out that “the good” which is different from the adjective “good” will have other adjectives besides “good”. Similarly in case of “the intelligent”, it is not the quality intelligent/intelligence, but rather it is the substantive to which the adjective intelligent will apply. It is then different from the adjective intelligent. And as such there will also be some other adjectives of that something which is “the intelligent”. Suppose if one says, “Obama is intelligent”. Here the “intelligent” is the quality, which is attributed to Obama. Obama then is “the intelligent”, the substantive which has as its adjective “intelligent”. Besides “intelligent”, Obama, that is the substantive, will have other adjectives too. One may then say, “Obama is tall”, “Obama is handsome”, “Obama is lean”, “Obama is good” and so on. There are many adjectives which may apply to the substantive, Obama, other than the adjective “intelligent”. Now to have all these adjectives applied to the same substantive does not mean that the adjectives have any relation among themselves except that of being the adjectives of the same substantive.

So my claim then amounts to: “Intelligence is a simple quality”. This then contradicts, the implication which was drawn earlier from the discussion on various notions of intelligence: “Intelligence is a complex entity”. It is then more sensible to attribute that simple quality of intelligence to individual actions and processes, rather than try and define intelligence in terms of those actions and processes. Those actions and processes can at best be defined as “the intelligent” but that will only again lead, after decomposition, to the simple quality of

---

<sup>271</sup> Ibid., p. 9.

intelligence which we already understand as a simple. Perhaps, this difficulty of explaining what constitutes intelligence is what made Turing put a human interrogator in the midst of the Turing-test. A human already understands what intelligence as a quality is and as such if he is led to believe that the entity on the other end is a human, then he is obviously attributing the quality of intelligence to it, for he sure knows human beings are intelligent. Also it may be noticed that in attempts to build or develop AI, the focus is mostly on the AI being able to carry out specific actions and not in giving rules that embody within them the principle or definition of intelligence itself. The latter is simply not possible as there is no definition of intelligence as it is a simple notion and hence, incapable of any definition. Meanwhile, the actions are selected precisely because they are attributed intelligence as their quality.

An objection here may be raised that the claims – intelligence is simple, only humans can identify cases of intelligence and so on – show that the whole argument is biased in favour of human. The only implication of these claims is that only those who are familiar with the concept of intelligence be it human or some alien (as long as they have a concept that is the same or similar to the idea of intelligence that human beings have), can understand what intelligence is and identify intelligent actions and behaviours. It is not to claim that animals cannot be intelligent. In fact, if the claims are taken to be valid, then, those claims imply that intelligence is a quality that applies to actions, behaviours, processes and so on. And the ones to perform those actions, behaviours or possess those processes will be intelligent irrespective of whether they are human or other animals. But perhaps, it may be taken as another implication of the claims made above, that only humans (until the time when other creatures with a similar concept is found) can identify those actions and behaviours and processes as intelligent.

#### **4. Intelligence in AI**

What about AI? Can AI computers be intelligent? My claims have shown that intelligence as a quality is disjunct from any mental or neural entity. It is also shown that intelligence is a simple notion and has no further components such as consciousness, intentionality and so on. As such, the implication from the claims regarding the above question will be that AI computers can be intelligent. But a point of importance to be noted is the fact that intelligence as a quality is a quality of something substantive. This something may be an action, behaviour or a process. And it is only if AI computers can be thought as doers and perform

actions that the quality of intelligence will be attributed to them. If what they perform is not considered as an action, then intelligence as a quality would not be attributed to them. For the purpose of simplicity, if behaviours, processes and actions are all included in the term action, then it may be said that actions are the substantives for the quality of intelligence. As such there can be two positions. First, one may consider only certain kinds of movements (for simplicity, only movements are considered, so this will include even the act of thinking or having thoughts) as actions. Movements involving consciousness, intentionality and so on only may be considered as actions. Second, any purposeful movements, whether conscious or not, intentional or not, and so on may be considered as actions. The term “purposeful” has been used here as computers, which may be taken as lacking consciousness, understanding (as argued earlier) etc., are still capable of movements that are purposeful, the purpose being solving a problem or being a solution to a problem. Even the manipulation of symbols might be considered as movements. One may argue that purposefulness itself pre-assumes consciousness, intentionality and understanding. Well it may be true but in case of computers this purpose is given by humans who are conscious, intentional and so on. So it may be taken as “extended” purposefulness. Whether to use the term “purposeful” or not, is not important. One may choose to call such movements as X, but whatever X is it should include the computer’s operations. So if we consider actions as involving only the first kind of movements (those involving consciousness, intentionality and so on), then surely, intelligence would not be attributed to AI. But if actions are considered as involving second kind of movements (those which are purposeful), then AI computers may be attributed intelligence. This creates a problem as to where to draw the line for considering some movements as actions or not. On one hand are those who consider only those movements as actions which have *at* or *as* its source consciousness, intentionality, understanding and so on. One such holder of this position is Searle. He says<sup>272</sup>,

*“Principle 1: Actions characteristically consist of two components, a mental component and a physical component.”*

And also,

---

<sup>272</sup> Searle, 1984, 63



“*Principle 2: The mental component is an intention. It has intentionality – it is about something.*”

Those, like Searle, supporting this position, that only those movements are to be considered as actions which have intentionality, consciousness and so on, will surely not accept the movements of computers as actions. As such they won't attribute intelligence to computers. But, on the other hand, are those who would consider the movements of a computer (which has no consciousness or intentionality at its source, and might only be purposeful in the sense that it is serving a purpose), say in playing a game of 'Go'<sup>273</sup>, as actions. If those movements are to be accepted as actions then computers may as well be attributed intelligence. But, of course, others like Searle, will deny this.

An alternative is also possible. One may just classify actions as *natural* and *non-natural* or *artificial*. Natural actions are the ones performed by organisms found in nature. “Organisms from nature” here, may simply mean organisms that have gone through natural evolution or are the result of natural evolution. Therefore, natural actions may include actions by plants too. It is to be noted that this notion of action is different from Searle's notion of action. For Searle, movements without the mental components even if the movement belongs to a body of human, fully conscious, does not classify as action. He says, “At first, it is tempting to think that types of actions or behaviour can be identified with types of bodily movements. But that is obviously wrong.”<sup>274</sup> Also the actions here are not being classified as natural actions on the basis of whether or not there is any consciousness or intentionality at their source or as their components. Non-natural or artificial actions are then those actions performed by entities/organisms which are created or made by human-kind. Whether they have consciousness or not is a point of debate and may depend upon what kind of organisms/entities humans are creating. But my claim here will be that they don't in case of AI computers. Based on these actions, then intelligence may be classified as – “natural” intelligence and “non-natural/artificial” intelligence. “Artificial” intelligence is then what AI computers are capable of. This seems to be stating the obvious.

What is to be understood in the above discussion of actions is that one may choose to accept which of the above alternatives s/he accepts as the idea of action. Based on this then one

---

<sup>273</sup> 'Go' is an ancient Chinese board game.

<sup>274</sup> Ibid., p. 57

may choose to call an AI computer as intelligent or not. It is not to be mistaken that, I am drawing any specific relation, other than the one I have discussed, between actions and intelligence. The only relation between them is that of substantive and adjective as argued earlier following an analogy with Moore's analysis of "good". Actions are substantive and intelligence is the adjective applied to the actions. So whichever entity, one may think are capable of actions, one may also then attribute intelligence. I have already argued earlier for the notion of intelligence being a simple notion and thereby not composed of consciousness, intentionality and so on. So if it were to be attributed to anything or anyone, it wouldn't be on the basis of considering consciousness, intentionality and so on as constituting intelligence. But rather it will be on the basis of what is to be considered as actions. One may as well bring in consciousness, intentionality and so on via the idea of actions. My only claim here will be that the notion of intelligence will share as much relation with them as the notion of good does when one utters such expressions as "Do good". Also it may be a point to be considered that what computers do is not the same as bodily movements that Searle denies are actions. Earlier it was mentioned that purposefulness may be taken as extended from humans as they are the ones creating computer for specific purposes or general purposes. As such what they do is different from say, the falling of stone due to gravity. So if one were to have the same position as Searle, they would need to show why, they would not accept what computers do as actions. After all what computers do is also the product or the result of human consciousness, intentionality and so on, as they are the ones who create computers. So what keeps Searle and others from extending the concept of actions to computers? This question surely requires further study and research.

This then shows that one might believe that AI can never have understanding, consciousness and so on, and yet, may still consider (if they consider what computer does as actions) them as intelligent, precisely because s/he can identify the quality of intelligence in its actions. This would, perhaps, be considered as the weak AI viewpoint. Penrose's viewpoint was that intelligence by virtue of necessarily requiring "understanding", which cannot be computational, is beyond AI. But my assertions would imply that this viewpoint, which has at its core the relation between intelligence and "understanding", is due to a mistaken idea of intelligence as something complex. Once this mistake is corrected, perhaps, intelligence will no longer be seen as necessarily connected to "understanding". But Penrose's assertion that "understanding" is non-computational still holds. Nothing in my assertions contradicts this. What

my claim amounts to here is that intelligence has no necessary connection to “understanding” and therefore, the effect of “understanding” being non-computational on intelligence is mitigated. As such Penrose’s claim that there is something missing in our present scientific picture of the world still holds. Physics might still need a reworking of its framework to allow such a thing as consciousness.

In this chapter I have put forward some claims. First was the claim that computers as processing representations are not only about formal symbol manipulation. Based on this claim I have further discussed that “understanding” involving semantics may or may not be possible in a computer. If one considers computers to have semantics (like Boden and Smith) then they may sure seem to have “understanding”. If one considers computers to not have semantics (like Searle), then they cannot have “understanding” that involves semantics. My own position was that representations by their very notion represent something and hence are about something. It is a matter of whether or not representations in computers may be said to have semantics. But at the same time I also have discussed “understanding” without semantics and it is this “understanding” that computers cannot be said to have. Also this “understanding” may be necessary to have the further “understanding” involving semantics, for an individual (be it of any species) cannot be thought to have any understanding without the subjective understanding (“understanding” without semantics) that underlies all his understanding. As such my position and claim will be that computers have no understanding at all. In this, my claim is as that of Penrose, who had shown, through Gödel’s theorem that understanding is non-computational and as such computers cannot have understanding. However, Penrose also connects intelligence with “understanding”. For him intelligence necessarily requires understanding. And since computers cannot have “understanding”, they cannot be intelligent either. I have argued and claimed that intelligence is a quality. As a quality it is a simple notion. Therefore, it can neither be an entity nor a complex. As such, it simply is a quality that humans recognize. Further this recognition of intelligence as a quality is often done as an adjective of a substantive. The substantive for intelligence is usually an action or an actor/doer. So if one has to attribute intelligence it has to be done on the basis of actions. Then, one has to consider what an action is or what can be actions. This of course requires furthers study and research. Perhaps one such area of research will be that of philosophy of action.

## Conclusion

In this dissertation we have first laid out the development and progress in AI, albeit, in a historical framework, to give an idea of the development of concepts that would, perhaps, help in understanding concepts and claims discussed later on. This forms the first chapter and is titled “Historical Antecedents of AI”. Chapter 1 then describes the early developments of AI as may be found in the works of McCulloch and Pitts (1943) and Alan Turing (1950). The idea of artificial neurons and Turing test are discussed. Further discussions are on works of Newell and Simon (1960s), some AI relevant concepts such as expert system, knowledge representation, artificial neural network and some future prospects of AI development. In the next chapter, that is chapter 2, I have discussed the philosophical concepts and theories on mind which further helps locate the discussion to be carried out later. Chapter 2 then is titled “Philosophical Discourse on the nature of Mind”. The discussion of mind has often been carried out in terms of soul. The philosophical concepts have been described along the lines of ancient notions of mind/soul in western philosophy, dualism and materialism. The ancient notions of mind are comprised of the ideas of mind in Hebrew, Homeric and pre-Socrates discourses. It briefly describes the usage of soul-terms such as *nepesh*, *ruach* and *leb* as found in *Old Testament*. It points out that the notion of soul was not considered as opposed to the notion of body as one may find in Cartesian dualism. This notion of soul runs common to both Homeric and pre-Socrates discussions of soul. However, it is also pointed out that the turn towards duality of soul and body is to be located in the pre-Socrates philosophy. Plato (4<sup>th</sup> century BC) is the first one to give arguments in favour of dualism. His notion of soul and body are discussed with critical comments by K. T. Maslin (2001). Platonic dualism is then contrasted with Aristotle’s (384 – 322 BC) notion of soul and body. In both Plato’s and Aristotle’s notions of soul and body, the idea of *forms* and *matter* play important roles. The discussion then proceeds on to Cartesian Dualism. The modern notion of mind-body dualism is often attributed to Descartes’ (1596 – 1650) notion of mind and body as two distinct and separate substances. Dualism as such gives rise to questions such as – what is the relation between mind and body or how does the causal interaction between the mind and the body take place? Such questions then pose the mind-body problem. Critical comments on Cartesian dualism by Maslin are taken up. Further discussions are on identity theories, forms of behaviourism and functionalism. The last section of chapter 2 is on the computational theory of

mind. The section describes how philosophical theories and concepts, that have been discussed till now, play important roles in the discourse on AI. Chapter 1 and 2 of this dissertation are then descriptive accounts of the relevant theories and concepts as found in AI and in the philosophy of mind. The purpose, of these two chapters, is to supplement the discussion that is to be carried out in subsequent chapters. Chapters 1 and 2, thus, help to build a backdrop for the philosophical debate on AI.

The third chapter takes up the philosophical debate on AI. The debate may perhaps be characterized as centred on the question posed by Turing, “Can a machine think?” The debate is based on the arguments by John Searle, Margaret Boden, Roger Penrose and Kevin Warwick. Searle’s position is that digital computers can only simulate thinking, and thereby intelligence. However, this simulation does not amount to duplication or attributing mental states to digital computers in the sense that humans are attributed. He does so in his famous Chinese Room Argument. He claims that mental states are *about* something or refers to something and as such have a semantic content. Digital computers are essentially the formal manipulation of abstract symbols. They are only syntactically defined and hence, do not have semantics. As such they cannot be thought to have mental states. He also provides an explanation to explain the mind-body relation. He claims that mind is both caused by and realized in the brain. In doing so he draws support from physical sciences by drawing an analogy of mind and brain with macro- and micro properties. Searle further feels that studies on mind conducted in association with computationalism or AI are not of much help in providing insights into the nature of mental states. Boden criticizes Searle on all fronts. She first argues that Searle’s analogy of mind and brain with its counterpart in physical sciences does not hold as our knowledge of mind and brain is very limited compared to our knowledge of those counterparts. She criticizes Searle’s Chinese room argument by pointing out that it doesn’t truly reflect what computationalists or computational psychologists say about the brain. It is often the case that as one tries to explain a feature of the mind, s/he does so in terms of underlying theoretical models which are relatively stupid or basic than the feature that is to be explained. Also, she has pointed out that Searle-in-the-room (another version of Chinese Room argument) at least has an understanding of the rule book and hence, a computer may understand the rules of the program. An important point highlighted by Boden is the notion of computer programs as bringing about representational activities of the program. She, following Smith, further makes the claim that computers have

semantics too. Penrose, like Searle, argues against the claim that computers can be intelligent. He provides his own arguments based on Gödel's theorem, to show that computers cannot even simulate "understanding", let alone have "actual" understanding as in the case of humans. This is where Penrose differs from Searle as Searle allows the simulation of mental states by a computer. Penrose uses the notion of "mathematical understanding" to demonstrate that the understanding that mathematicians employ cannot be described computationally. As such understanding is non-computational. Also Penrose feels that intelligence requires understanding. If "understanding" is non-computational then so is intelligence. Warwick brings in the charge that arguments as given by Searle and Penrose are human biased. Their notions of intelligence, understanding, and awareness don't do justice to the non-human species which includes birds, animals and machines. He, thus, appeals for a more inclusive and broader idea of the respective notions. The discussion and analysis of the debate on AI carried out in this chapter highlight some key concepts that are important to this debate. The next chapter then discusses those concepts.

The question at the centre of debate on AI is "Can a machine think?" or "Can a machine be intelligent?" A study and analysis of the debate, based on the arguments as given by Searle, Boden, Penrose and Warwick, puts forward some key concepts as important to this debate. These concepts are the notion of computer, the notion of understanding and the notion of intelligence. Chapter 4 deals with these concepts. Searle considers computers as just formal manipulation of abstract symbols and hence, syntactically defined. Following Tim Crane, a computer can be understood as a device processing representations systematically. Here the notion of computer includes representations. As such the notion of representation plays a central role in the notion of computer. We then arrive at the claim that computers as processing representations are not merely the manipulation of abstract symbols. The symbols as representations are not inert and represent something. As representing something, they are about something. The claim is that they represent at least the type of the content that the system computes. A further claim is that it is due to the pre-assumed notion of representations that the representational activity as described by Boden is brought about by a computer program. The discussion of semantics within the notion of computers provides for the analysis of the notion of "understanding". The notion of "understanding" is further analysed as "understanding" with semantics and "understanding" without semantics. Both the notions of "understanding" is then analysed in terms of computers.

The claim arrived at is that computers cannot have “understanding”. Next the concept of intelligence is taken up. Various notions of intelligence are considered. It was observed that intelligence is often mistakenly thought of as an entity that is present in the brain. Further, there is also the notion of intelligence as a complex notion. Intelligence is thought of as constituted by further components such as learning, knowledge application, consciousness, “understanding” and so on. It seems then that the notion of intelligence is that of a complex entity. It is then claimed that, this is not the case. Intelligence is a simple notion and at the same time it not an entity but a quality. So intelligence is understood as a simple quality. This claim is based on the analogy drawn between the notion of intelligence and the notion of good as given by Moore. Intelligence as a quality is shown to apply to actions and the doers (of the action). The notion of intelligence is also shown to be disjunct from consciousness and intentionality. With these notions of intelligence it is then discussed whether AI computers can be called intelligent. A claim then is made that the answer to this question lies on whether or not computers may be said to perform actions. Can what a computer do be included under the notion of actions? It is then claimed that if what a computer does is considered as actions then one may attribute intelligence to them if their actions are intelligent. However, if what a computer does is not considered as an action then they cannot be considered as intelligent either. A way of classifying actions is described. One view, as Searle explains, understands actions as having both mental and physical components to it. Another view, classifies actions as those movements which are purposeful, whether or not there is any mental component to it. An alternative pair of distinction is also given in terms of the doers. On one hand then, if doers are natural organisms and entities then what they do may be considered as actions. On the other hand, if the term “doer” includes even those entities created by humans, whether conscious or not, then again what they do may be considered as actions. And, if this view is accepted then AI computers will be considered as intelligent.

In this dissertation, the conclusion consists of specific claims as described above, mostly in chapter 4. The claims are supported by analysis carried out in chapter 3 and further developed in chapter 4. It is not to claim that they are the last and final word on the discussion of the notions considered here. In fact, not making any concrete assertions with regard to questions such as “Can a computer be intelligent?” and “What to classify as actions?” or “Whether to classify what a computer does as action or not?” show the discussion to be still open.

## **Limitation of the study**

This dissertation carries out a study of the philosophical critique of AI. Based on the study a few conclusions have been posited. However, the study in no sense claims to be a perfect study and without limitations. It may be pointed out that the debate on AI is much more vast and varied than what has been considered here. The initial limitation has been set by considering mostly the arguments by only four individuals, namely, John Searle, Margaret Boden, Roger Penrose and Kevin Warwick. As such the discussions have been limited only to the debates carried out by them. The discussion within the dissertation had also often encountered concepts and notions that have been left out of further discussion in order to avoid digression. However, a more detailed study and analysis of such concepts would always provide for interesting areas of research. One such notion is the notion of representations. Representations by itself provides for a whole philosophical study. For the purpose of this dissertation, the notion of representation was taken to be as “that which represents something”. A very brief explanation was given and references were instead made to more detailed analysis of the notion of representations by others. This then sets another limitation to the dissertation. A limitation of this study may be that it does not give any concrete and conclusive remark. For instance on the question – does a computer have semantics? – the answer depends on whether one considers computer as only syntactically specified or not, keeping the question wide open. The claim provided here is that as a device that processes representations and representations being about something, a computer may or may not have semantics based on whether or not one accepts that just being about something is enough to be a content or at least content-like. This claim is inconclusive. The claim may be taken either way – as not having semantics or as having semantics. Another question that remains unanswered is – what is to be classified as actions? Based on this then the question “can a computer be intelligent?” needs to be attended to. These are some of the limitations of this study. Also it may be pointed out that in drawing an analogy with the concept of good as given Moore, it is not to claim that Moore’s analysis is absolute and thereby the analysis of the notion of intelligence in a similar vein is absolute too. The dissertation does not make any such claims. It only points out one way of analysing the notion of intelligence, that is, by drawing an analogy with the notion of good. Alternative analysis may be carried out too. In making this claim it is to accept that there are more methods of analysis and that this study does not claim to have



exhausted all the methods of analysis. It is to accept that this study is limited in its analysis of the notions.

### **Prospects for Future Research**

A broader conclusion may be made on the basis of this dissertation. It may be observed: first, that AI does provide for a philosophical debate and second, that philosophy has to discuss AI. The claims in AI do affect philosophy and philosophy in turn has to take up AI and deal with it philosophically. What this dissertation ultimately highlights is the inter-affectations that philosophy and AI may have on each other. The question then is – How, we as a student of philosophy and as philosophers confront the field of AI to study and analyse the various concepts employed there in as well as the possibilities that AI promises?

It has been mentioned above that there are some limitations to the study carried out in this dissertation. Some such limitation are due to not taking up in detail some concepts and theories. As such they provide for issues to be further looked into. Further research may then be carried out on those lines. One interesting topic of research is that of representations. It may be pointed out that questions such as – what is the nature of representations; are representations physical; or are they something else; can mental states as representations fit within the physical picture of the world; if mental states are representation, can they be compared with computers as another system processing representations etc. pose some research problems. Representation, thus, provides for an interesting topic of research. Another area of research, as pointed out earlier in chapter 4 was that of philosophy of action. To bring AI into the philosophical discussion on action provides for an interesting inter-disciplinary topic of research. The discussed analogy of the notion of intelligence with the notion of good also provides for further lines of research. One such line of research can be in terms of whether other analyses of the notion of good provides for any insight when applied to the notion of intelligence. Also a line of research can be what other kinds of analysis may be applied to intelligence. A further question can be – does such an analysis, at all, capture the notion of intelligence?

## BIBLIOGRAPHY

### Primary sources

#### Books

Boden, Margaret A. *AI: Its nature and future*. Great Clarendon Street, Oxford: Oxford University Press, 2016.

Boden, Margaret A. *The Philosophy of Artificial Intelligence*, ed. Margaret A. Boden. Walton Street, Oxford: Oxford University Press, 1990.

Crane, Tim. *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representations*, first published in 1995, second edition. London: Routledge, 2003.

Dreyfus, Hubert L. *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row, Publishers, Inc., 1972.

Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, Massachusetts: MIT Press, 1992.

Kripke, Saul A. *Naming and Necessity*. Cambridge, Massachusetts: Harvard University Press, 1980.

Maslin, K. T. *An Introduction to the Philosophy of Mind*. Cambridge: Polity Press, 2001.

Moore, G. E. *Principia Ethica*. Cambridge: Cambridge University Press, 1903

Penrose, Roger. *The Emperor's New Mind*. London: Oxford University Press, 1990.

Penrose, Roger. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press, 1994.

Searle, John. *Minds, Brains And Science*. Cambridge, Massachusetts: Harvard University Press, 1984.

Walmsley, Joel. *Mind and Machine*. Hampshire, England: Palgrave Macmillan, 2012.

Warwick, Kevin. *Artificial Intelligence: The Basics*. Abingdon, Oxon: Routledge, 2012.

Warwick, Kevin. *In The Mind Of The Machine: The Breakthrough In Artificial Intelligence*. London: Arrow Books Limited, 1998.

## **Articles**

Turing, Alan M. "On Computable Numbers, With An Application To The Entscheidungsproblem". *Proceedings of the London Mathematical Society*, series 2, Vol. 42, originally published by the London Mathematical Society, 1936-37, pp. 230-65.

## **Secondary Sources**

### **Books**

Annas, Julia. *Hellenistic Philosophy of Mind*. Berkeley: University of California Press, 1992.

Baker, Gordon & Morris, Katherine. *Descartes' Dualism*. London: Routledge, 1996.

Bermúdez, José Luis. *Thinking without Words*. New York: Oxford University Press, Inc., 2003.

Block, Ned. "Troubles with Functionalism" in *Readings in the Philosophy of Psychology*, vol. 1, ed. Ned Block. London: Methuen, 1980.

Boden, Margaret A. *Artificial Intelligence: Handbook of Perception and Cognition*, 2<sup>nd</sup> ed. Sand Diego, California: Academic Press, Inc., 1996.

Boden, Margaret A. *The Creative Mind: myths and mechanisms*. London: Routledge, 2004.

Boden, Margaret A. *Mind as Machine: A History of Cognitive Science, Volume 1&2*. Great Clarendon Street, Oxford: Oxford University Press, 2006.

Boden, Margaret A. *Creativity and Art: Three Roads to Surprise*. Great Clarendon Street, Oxford: Oxford University Press, 2011.

Braddon-Mitchel, David and Frank Jackson, *Philosophy of Mind and Cognition*, UK: Blackwell Publishing Ltd, 1996.

Bremmer, Jan. *The Early Greek Concept of Soul*. Princeton, NJ: Princeton University Press, 1983.

Chalmers, David J. *Philosophy of Mind: Classical and Contemporary Readings*. New York: Oxford University Press, 2002.

Churchland, Paul M. *Plato's Camera: How the Physical Brain Captures a landscape of Abstract Universals*. Cambridge, Massachusetts: MIT Press, 2012.

*Collected Papers on Mathematics, Logic and Philosophy*. Translated by M. Black, V. Dudman, P. Geach, H. Kaal, E.-H.W. Kluge, Brian McGuinness and R.H. Stoothoff. New York: Basil Blackwell, 1984. Originally published as *Kleine Schriften* (Hildesheim: Georg Olms, 1967).

Dennett, Daniel C. *Consciousness Explained*. New York: Little, Brown and Company, 1991.

Dennett, Daniel C. *The Intentional Stance*. Cambridge, Massachusetts: MIT Press, 1987.

Fillard, Jean-Pierre. *Brain vs Computer: The Challenge of the Century*. New Jersey: World Scientific, 2016.

Foster, John. *The Immaterial Self: A Defence of the Cartesian Conception of the Mind*. London: Routledge, 1991

Geach, P. and Anscombe, G. E. M. (eds.). *Descartes: Philosophical Writings*. Nelson University Paperbacks for The Open University, 1970.

Granger, Herbert. *Aristotle's Idea of the Soul*. Dordrecht: Kluwer Academic, 1996.

Grayling, A. C. *Berkeley: The Central Arguments*. London: Duckworth, 1986.

Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge, Massachusetts: MIT Press, 1985.

Hobbes, Thomas. *Leviathan*, ed. Richard Tuck. Cambridge: Cambridge University Press, 1991.

Horst, Steven. *Beyond Reduction: Philosophy of Mind and Post-Reductionist Philosophy of Science*. New York: Oxford University Press, 2007

Jackson, Frank. "What Mary Didn't Know", *The Nature of Mind*, ed. D. Rosenthal. Oxford: Oxford University Press, 1991.

Warwick, Kevin and Shah, Huma. *Turing's Imitation Game: Conversation with the Unknown*. UK: Cambridge University Press, 2016.

Khalifa, Jean. *What is Intelligence?* ed. Jean Khalifa. Cambridge: Cambridge University Press, 1994.

Klement, Kevin C. *Frege and the Logic of Sense and Reference*. London: Routledge, 2002.

Leibniz, G. W. *The Monadology: An Edition for Students*, ed. Nicholas Rescher. London: Routledge, 1991.

MacDonald, Paul S. *History of the Concept of Mind*. England: Ashgate Publishing Limited, 2003.

Mintz, Samuel. *The Hunting of Leviathan*. Cambridge: Cambridge University Press, 1970.

Netton, Ian Richard. *Al-Farabi and His School*. London: Routledge, 1992.

Nussbaum, Martha & Rorty, A. O. (eds) *Essays on Aristotle's De Anima*. Oxford: Clarendon Press, 1992.

*On Intelligence: The Toronto Symposium on Intelligence, 1969*, edited by W. B. Dockrell. London: Methuen & Co Ltd, 1970.

Onions, R. B. *The Origins of European Thoughts*. UK: Cambridge University Press, 1951.

Plato. *Phaedo*. 80a-b. in E. Hamilton and H. Cairns (eds), *The Collected Dialogues of Plato*. Princeton: Princeton University Press, 1961.

Reus, Bernhard. *Limits of Computation: From a Programming Perspective*. Switzerland: Springer Nature, 2016.

Russell, Bertrand. *History of Western Philosophy*, Special Indian Edition, First Indian Reprint, 2013, New York: Routledge, 2004.

Ryle, Gilbert. *The Concept of Mind*. London: Hutchinson, 1949.

Smith, B. C. *Reflection and Semantics in a Procedural Language*. Cambridge, Massachusetts: MIT Ph. D. Dissertation and Technical Report LCS/TR-272, 1982.

Spinoza, Baruch. *Ethics*. In E. Curley (ed.), *The Collected Works of Spinoza*. Princeton: Princeton University Press, 1985.

Thomson, G. *The Factorial Analysis of Human Ability*. London: University of London Press, 1950.

Yolton, John. *Thinking Matter: Materialism in Eighteenth-Century Britain*. Minneapolis, MN: University of Minnesota Press, 1983.

## Articles

Clarke, J. J. "Turing Machines and the Mind-Body Problem." *The British Journal for the Philosophy of Science*, Vol. 23, No. 1 (Feb., 1972), pp. 1-12.

Copeland, B. Jack. "Turing's O-Machine, Searle, Penrose and the Brain" *Analysis*, Vol. 58, No. 2 (April., 1998), pp. 128-138. Oxford University Press on behalf of The Analysis Committee.

Dennett, Daniel C. "Illusionism as the Obvious Default Theory of Consciousness." *Journal of Consciousness Studies*, 23, No. 11-12, 2016, pp. 65-72.

Frege, Gottlob. "On Sense and Meaning". In *Collected Papers on Mathematics, Logic and Philosophy*, ed. Brian McGuinness, New York: Basil Blackwell, 1984, 157-77. Originally published as "Über Sinn und Bedeutung". *Zeitschrift für Philosophie und philosophische Kritik* 100, 1892, pp. 25-50.

Jarrett, Charles. 'Spinoza's Denial of Mind-Body Interaction and the Explanation of Human Action', in *Southwest J. Phil.* 29, 1991, pp. 465-85.

Koistinen, Olli. 'Causality, Intensionality and Identity: Mind-Body Interaction in Spinoza', in *Ratio*, 9, 1996, pp. 23-38.

Lys, Daniel. *Ruach: Le soufflé dans l'Ancien Testament*. EHPHR, annual issue, 1962.

Pitts and McCulloch, "How We Know Universals: The Perception of Auditory and Visual Forms", *Bulletin of Mathematical Biophysics*, 9, reprinted in S. Papert (ed.), *Embodiments of Mind*. Cambridge, MA: MIT Press, 1965, pp. 127-47.

Turing, A. M. "Computing Machinery and Intelligence". *Mind* 49: 433-460. 1950.

Yandell, David. 'What Descartes Really Told Elizabeth: Mind-Body Union as Primitive Notion', in *Brit. J. Hist. Phil.* 5, 1997, pp. 249-73.

### Online Sources

Bostrom, Nick. "Superintelligence: Answer to the 2009 EDGE QUESTION: "WHAT WILL CHANGE EVERYTHING?"" <https://nickbostrom.com/views/superintelligence.pdf> Accessed on 10/10/2017.

Dreyfus, Hubert L. "Artificial Intelligence." *The Annals of the American Academy of Political and Social Science*, vol. 412, 1974, pp. 21–33. *JSTOR*, [www.jstor.org/stable/1040396](http://www.jstor.org/stable/1040396). as accessed on 21/07/2018.

Jackson, Frank. "Epiphenomenal Qualia." *The Philosophical Quarterly (1950-)*, vol. 32, no. 127, 1982, pp. 127–136. *JSTOR*, [www.jstor.org/stable/2960077](http://www.jstor.org/stable/2960077).

McCarthy, John. "What has AI in Common with Philosophy?" <http://www-formal.stanford.edu/jmc/aiphil.pdf> Accessed on 09/10/2017.

Stoljar, Daniel. "Physicalism". *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition). Edward N. Zalta (ed.). URL = <https://plato.stanford.edu/archives/spr2016/entries/physicalism/>. Accessed on 16Moore, G. E., *Principia Ethica*, Cambridge University Press, 1903.