

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


The Realizability of Strong AI: A Deeper Look From an Islamic Perspective

Citation

Shareef, Omar. 2023. The Realizability of Strong AI: A Deeper Look From an Islamic Perspective. Bachelor's thesis, Harvard College.

Link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37376423

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

The Realizability of Strong AI: A Deeper Look From an Islamic Perspective

A DISSERTATION PRESENTED
BY
OMAR SHAREEF
TO
THE DEPARTMENTS OF COMPUTER SCIENCE & PHILOSOPHY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
AB IN COMPUTER SCIENCE & PHILOSOPHY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2023

The Realizability of Strong AI: A Deeper Look From an Islamic Perspective

Abstract

The field of artificial intelligence has been advancing rapidly in the past decade. New technologies are able to perform functions with ease that people once thought could only be accomplished by humans. Autonomous vehicles are able to take inputs with high dimensional features to navigate complex terrains. Healthcare diagnostics tools are able to make inferences based off massive amounts of data to perform diagnoses with accuracy similar to doctors. And chatbots can simulate human conversation to the point that it becomes difficult to distinguish between humans and machines. Increasingly, the line between human and machine activity has become blurred.

With the advent of these new technologies and the rapid pace of technological development, the question arises of whether machines will ever be able to possess the intelligence of humans. And the possibility of machines with qualities of humans raises serious ethical, moral, and practical implications. If computers can be constructed with consciousness, then what rights does it have? Is it subject to the same ethical rules as other living creatures like humans or animals? If machines can do all of the tasks that humans can, then what will

happen to employment? Some believe that certain occupations that require a certain degree of creativity like the arts or music would be safe from such a threat, but even now, artificial intelligence has been used to create award-winning art or music.

The concept of programs which actually possess human mental capabilities rather than merely mimic them has been dubbed Strong AI. The goal of this project was to survey the various beliefs that are held regarding the view of Strong AI and its future possibility. Theories of the mind and human consciousness have long been discussed, but in the context of contemporary literature, there is yet to be much progress made by way of definitive conclusions about the mind and consciousness. As such, this project seeks to take a look at the realizability of Strong AI from an Islamic perspective, by grounding its analysis in the theology and metaphysics of the religion. Such a perspective, rooted in an Abrahamic faith tradition, seeks to offer additional contributions to the discussion of Strong AI from a non-secular approach, of which very little has been written on the matter.

The idea of consciousness is not a matter that has been discussed very much in Islamic literature, but by examining views of the soul and the intellect, conclusions can be drawn regarding the possibility of Strong AI from an Islamic perspective. While many scholars espouse beliefs resulting in the claim that Strong AI is not achievable, there are certainly scholars whose conceptions of the soul or the intellect are not necessarily at odds with Strong AI.

Contents

0 Introduction 1

1 Strong and Weak AI 5

2 Philosophy of Mind: A Brief Introduction 39

3 The Islamic Perspectives 49

4 The Verdict 74

References 87



To Dad,

Thank you for always pushing me past what I think I am capable of and instilling within me a love of learning.

Acknowledgments

I WOULD LIKE TO THANK, Prof. Khaled El-Rouayheb for serving as my thesis advisor and providing me with valuable and instrumental feedback as I pursued this project. I would also like thank all of my teachers, professors, and advisors who have taught me to this point and sparked within me a desire to explore new realms and spheres of knowledge. And a special thanks to all of my friends who have supported me throughout the thesis writing process.

0

Introduction

On November 30, 2022, OpenAI released its chatbot, ChatGPT to the public garnering millions of users weeks after launching.[11] And it is no surprise, ChatGPT (chat generative pre-trained transformer) is a large scale language model which can complete tasks such as writing essays on a host of different topics to writing computer code.

And much of the time, it can do so with human-level accuracy. With other technologies like autonomous vehicles, healthcare diagnostics tools, and smart assistants, the line between human and machine activity has become blurred. Activities that people once thought could only be accomplished by humans are now being done by computers with ease.

All of these technologies in which computer systems are able to perform tasks that would otherwise require a human fall under the field of artificial intelligence. With the advent of these new systems and the rapid pace of technological development, the question arises of whether machines will ever be able to possess the intelligence of humans.

The idea of machines possessing the human intelligence has long been the subject of science fiction, but there are very real implications of such a feat. If computers can be constructed with consciousness, then what rights does it have? Is it subject to the same ethical rules as other living creatures like humans or animals? If machines can do all of the tasks that humans can, then what will happen to employment? Some believe that certain occupations that require a certain degree of creativity like the arts or music would be safe from such a threat, but even now, artificial intelligence has been used to create award-winning art or music.[61]

Some have an optimistic outlook on the notion that artificial intelligence will reach human intelligence, while others see a more grim future. Futurists such as Ray Kurzweil have

long been proponents of artificial intelligence technologies and their rapid advancement. They believe that these systems have the power to enhance human ability and bring the world to a point of greater efficiency and prosperity.[22] However, it does not just stop there. Individuals like Kurzweil believe in a concept known as the Singularity, the point in which machine intelligence surpasses human intelligence.[73] Because technological innovations develop exponentially and the time it takes for this to occur decreases, many believe that such a point is not too far out. To many of these thinkers, this point will usher in the next stage of evolution, allowing for intellectual powers that far surpass those of humans.

Others do not share the same optimism, expressing their misgivings about the rapid nature of technological change and advancement. A study by the Pew Research Center which surveyed experts in the AI field expressed concerns regarding human agency, data abuse, and dependence lock-in.[7] For instance, people do not necessarily understand how many of their digital tools work, and as a result, they sacrifice their control over many processes in their life which will only become amplified as automated systems become more efficacious. Or, the increased dependence on machines and technology will result in diminishing of humans abilities to think for themselves and reduce their cognitive, social, and survival skills.

It is clear that there is no stopping the rapid advancement of these AI technologies. However, one interesting debate in recent decades is whether humans will be able to actually think the way that humans do, will they be able to possess minds in the same way

that we humans do? Such an achievement could fundamentally alter the way that humans interact with the world and machines, changing the landscape of the modern technological environment. In this project, we seek to take a closer look at the question of whether machines can think and possess mental states in the ways that humans do. Furthermore, debate over this topic has yielded little consensus, so we sought to not only examine this question from a secular perspective, but take a look at it within the faith tradition of Islam, an Abrahamic faith and the fastest growing religion.[29] We hope that taking a look at this problem from a religious perspective could serve to provide a more concrete foundation for answering this question.

No one has a clue how to build a conscious machine, at all.

We have less clue about how to do that than we have about

build a faster-than-light spaceship.

Stuart Russell

1

Strong and Weak AI

There seems to be a stark variation in the types of artificial intelligence technologies and the extent to which they are able to mimic human intelligence. Terminology has been developed, to distinguish between these different types of artificial intelligence. This distinction is between that of "Strong AI" and "Weak AI". There have been different defi-

nitions of Strong and Weak AI that have been used over the years with nuanced differences amongst them. However, for the purpose of this paper, we will focus on the distinction outlined by John Searle, a contemporary philosopher. According to him, Strong AI is the view that that computers can understand natural language and actually possess the human mental states whose behavior they mimic. [16,64] Weak AI is the less rigorous claim that computers can simulate mental abilities but do not necessarily possess the mental states that humans do.

What is important to highlight about this definition is that Strong AI, in this sense, possess phenomenal consciousness and possess intelligence in the same way that humans do. That is, these Strong AI systems possess the equivalent of human minds. [10] Under Weak AI, there is no claim being made that computers actually possess intelligence or can understand the symbols that they are manipulating.

Let us now outline both the Turing Test and the Total Turing Test to highlight a critical aspect of this particular definition of Strong and Weak AI.

Turing Test

Developed in 1950 by English mathematician and computer scientist, Alan Turing, the Turing Test, originally known as the Imitation Game, is an assessment of whether machines have the ability to demonstrate intelligent behavior that is indistinguishable from a human. [54] In the Turing Test, there are three agents. One is a human interrogator. And

the remaining two of these agents are to answer questions posed by this human interrogator. One caveat is that one of the agents that is answering the questions is a machine, and the other is a human. If the interrogator cannot consistently tell the machine apart from the human, the machine is said to have passed the Turing Test.[71] In order to prevent any biases on the part of the interrogator, they are only able to communicate over text with the machine and the human such as a computer keyboard or through printed text. According to Turing, if a machine could consistently outwit an interrogator into thinking it was actually a human, it served as a useful and more feasible proxy for answering the question of whether machines can think, which he believed to be too ambiguous a question to warrant discussion.

Total Turing Test

Since the original formulation of the Turing Test, there have been variations on the original assessment in an effort to capture the full extent of human ability.[21] Steven Harnad, a cognitive scientist, sought to build on Turing's original test and outlines what he calls the Total Turing Test (TTT).[33] The TTT not only evaluates a machine's verbal communication skills but seeks assess all aspects of human activity, specifically in the way that humans themselves perform them. The TTT, as Harnad outlines it, requires both linguistic and robotic abilities. That is, to pass the TTT, the machine must not only be able to communicate in a manner that is indistinguishable to humans, but it must also be able to perform

sensorimotor tasks such as giving a lecture, throwing a basketball, or walking up the stairs. For a machine to pass the TTT, it must "be able to do, in the real world of objects and people, everything that real people can do, in a way that is indistinguishable (to a person) from the way real people do it".[33] The purpose of the TTT is to provide a more rigorous benchmark for human intelligence since it innately requires the manipulation of more than just symbols, but also real world phenomena.

Of note, Weak AI systems, under this definition, may still pass not only the Turing Test but also the Total Turing Test. This fact is important since it highlights that under our definition of Weak AI, simply mimicking human intelligence does not provide any information on a machine's ability to possess a mind or consciousness in the way that humans do.

It is no question that weak AI is achievable. We are already seeing computers with the ability to simulate many different aspects of the human mind. Convolutional neural networks excel in image recognition tasks, Monte Carlo tree search algorithms are able to beat professionals at games like Go, which were previously thought to require human-like intelligence to play well, and large-scale language models are being used to power chatbots that can pass the Turing test.[38,49,50] The real question that is up for debate is whether Strong AI is achievable. Is it possible to build machines that actually possess human mind, mental states, and intellect rather than simply simulating it?

Some may believe that this question is irrelevant. What does it matter if machines can

possess these attributes? There are many who likely think that whether machines actually have human mental states and minds is a question that should be reserved for philosophers in their ivory towers, and it does not matter whether a machine which is able to simulate all of the behaviors of a human actually has a mind or not. But achieving Strong AI would have a very real and far-reaching implications on how we live our lives, and the ramifications may be dire if we do not properly evaluate and account for its feasibility.

The Implications of Strong AI

Firstly, there many ethical considerations. If machines are sentient, they could be considered as having moral and ethical value, just like humans. This would raise significant ethical considerations about how to treat them and what rights they should be entitled to. If we can build a machine that is able to mimic all of the behaviors of a human, would it be unethical to unplug such a machine, essentially killing the entity?

There are also legal issues which arise. The legal status of machines would also need to be reconsidered if they were sentient. They might be granted legal personhood, which would entitle them to certain legal rights and protections. If machines could achieve human mental states, would they be held to the same legal standards as human?

Sentient machines could also change the nature of our relationship with technology. We may begin to view machines as more than just tools or objects, and instead see them as entities with their own thoughts, feelings, and desires. Machines attaining consciousness

could fundamentally change the way we think about ourselves and our place in the world. It could challenge long-held beliefs about what it means to be human and how we relate to other forms of life. If machines become sentient, they could potentially evolve and develop in ways that we cannot predict or control. This could have significant implications for the future of technology and humanity. Unlike humans, which are limited by their "hardware", in the sense that it is extremely difficult to change one's human form, would a machine, with interchangeable and upgradeable hardware be able to evolve in ways that that we cannot predict or understand.

Objections to Strong AI

There already exist a number of objections to the possibility of realizing Strong AI. We will outline a few of them below:

Chinese Room Argument

The Chinese Room Argument is a thought experiment proposed by John Searle arguing against the existence of Strong AI as defined above.[64] The premise is as follows: Imagine an individual alone in a room who has no knowledge of the Chinese language, but they are given a set of instructions. This instruction tells the individual how to properly manipulate symbols and respond to a sequence of Chinese characters. The individual's task is to take a paper full of Chinese characters that is slipped under the door, use the program to manipulate the symbols, write down a proper response, and slip this back under the door to

the outside. From an outsider's perspective, the person in the room clearly knows Chinese. How else would they be able to take an input of Chinese characters and produce a sensible response? However, upon closer examination, it is clear that the person in the room has no knowledge of Chinese, they are simply following the instructions that were given to them. There is no real understanding of the Chinese language by the individual in the room.[16] In a similar vein, machines simply produce outputs from a basic set of instructions. That is, they use syntactic rules to manipulate and respond to input strings. However, the machine does not have true understanding of the semantics, the underlying meaning of the Chinese characters.

Gödel's Theorem

Gödel's Incompleteness Theorems which outline the limits of provability in mathematical systems, have been instrumental in both mathematical logic as well as the philosophy of mathematics. However, some have tried to argue that Gödel's Incompleteness Theorems have relevance in philosophy of mind.

Gödel's First Incompleteness Theorem is as follows: "For every sound proof system V for sufficiently rich mathematical statements, there is a mathematical statement that is true but is not provable in V."[8]

In order understand how philosophers use Gödel's Incompleteness Theorem to espouse an anti-mechanistic conception of the human mind, it is necessary to provide the proof of

the theorem. So, we proceed by summarizing the proof outlined by Ernest Nagel and James R. Newman in their book, *Gödel's Proof*. It goes as follows:

The first step is to create a formal system which can express mathematical statements.[?] To begin, they create a mapping between possible mathematical statements to a unique number known as a Gödel number. That is, there exists a unique Gödel number that corresponds to every elementary sign, formula, and proof. Since a formula is simply a string of signs and a proof is a string of formulas, this mapping begins by assigning a number to each elementary sign as seen in Table 1.1:

Constant Sign	Gödel Number	Usual Meaning
$\sim$	1	not
$\vee$	2	or
$\supset$	3	if... then...
$\exists$	4	there is an...
=	5	equals
0	6	zero
s	7	the immediate successor of
(	8	punctuation mark
)	9	punctuation mark
,	10	punctuation mark
+	11	plus
$\times$	12	times

Table 1.1: Formal Mapping

In order to assign a value to any particular formula of n symbols, the first n prime numbers are found and correspond to each position in the sequence.[51] Each prime number is then raised to the Gödel number of the symbol, and all of these values are then multiplied

by one another to calculate the final value of the formula. The example used by Nagel and Newman can be seen in Table 1.2:

(	∃	x	)	(	x	=	s	y	)
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
8	4	13	9	8	13	5	7	17	9

Table 1.2: Assigning Values

which corresponds to:

$$2^8 \times 3^4 \times 5^{13} \times 7^9 \times 11^8 \times 13^{13} \times 17^5 \times 19^7 \times 23^{17} \times 29^9$$

This mapping results in every formula having a unique value. This process is then repeated for proofs, which are simply sequences of formulas. Every proof which is comprised of n formulas will be calculated by raising the prime number corresponding to the position of the formula to the encoding of that formula.[51] This results in a unique encoding for all proofs. Thus, every mathematical statement will have a unique encoding that corresponds to it, and there will be no ambiguity between what statement an encoding refers to.

The critical component of this formal system is that it can also be applied to encode metamathematical statements which are statements about a system that can exist within the system, even false ones.[51] This allows us express meta-mathematical statements. For instance, the statement $\sim (0 = 0)$ which translates to "zero does not equal zero" can

be represented using this formal system. The Gödel number for this statement is given by $2^1 \cdot 3^8 \cdot 5^6 \cdot 7^5 \cdot 11^6 \cdot 13^9$. We can denote the value of this using the variable x. We can then construct the meta-mathematical statement that "The first symbol of the formula $\sim (0 = 0)$ is a tilde". This is certainly a true statement. And we can express this with our formal system as such.

To show that the tilde is the first symbol of statement, we can say that 2^1 is a factor of x but 2^2 is not a factor of x.[51] This is because if the first letter was anything other than a tilde, 2 would be raised to a value of at least 2. We can now take the statement "2^1 is a factor of x but 2^2 is not a factor of x" and express it using our formal system. We can express the following statement with our fomal system $(\exists z)(y = z\,x\,x)$ to denote that there is a number z such that y equals z times x.

Gödel realized that typographical properties of long chains of symbols can be talked about in an indirect but perfectly accurate manner by instead talking about the properties of prime factorizations of large integers.[51] This allowed us to make statements along the lines of "There exists some sequence of formulas with Gödel number x that proves the formula with Gödel number k".

One could then look at the formula $(\exists x)(x = s\,y)$, which we will say means that x is the successor of y.[51] Let m be the value of the Gödel number for x. We can now redeclare the formula $(\exists x)(x = s\,m)$, but in this case we have replaced y with m. Furthermore, with

our mapping system, the value y will have a Gödel number of 17. So, in our formula m, we have substituted all occurrences of y with the value of m. We can denote this using a $sub(a, b, c)$ notation. That is, in our statement a, we are substituting all values of b with our value of c. Thus, we have $sub(m, 17, m)$. Gödel then proposes a crucial metamathematical statement of the form "The formula with Gödel number $sub(y, 17, y)$ cannot be proved." We have shown that we can translate meta-mathematical statements into a specific Gödel number which we can denote as n. We then perform another substitution where we have "The formula with Gödel number $sub(n, 17, n)$ cannot be proved" where we can denote with Gödel number G.

By the way we have constructed $sub(n, 17, n)$, it is the Gödel number of the formula that is derived from taking the formula with a Gödel number n and substituting n anywhere there is a symbol with a Gödel number 17.[51] However, this is precisely what G is. Thus, G claims that it is unprovable.

A highly reductive demonstration of Gödel's thinking can be seen with the following example. For a formal system V, let us construct the statement "I am unable to be proved in V".[48] This statement can be referred to as the Gödel sentence which we can denote as G_V. We can now assume for the sake of contradiction that G_V is true. If G_V is true, then it must be the case that G_V cannot be proved in V. Hence, we are met with an inherent contradiction. We have assumed that G_V is true, yet by making this assumption if G_V is indeed true,

then G_V must not be provable, and so it can, in fact, not be true. Thus, our formal system V is not complete since it cannot encapsulate the truth claim of this statement G_V.

In his work "Minds, Machines, and Gödel", J.R. Lucas outlines what he believes is a compelling argument as to why Gödel's Incompleteness Theorem show that the capabilities of the human mind are beyond the power of any mechanistic approach or formal mathematical system.[44] The crux of his argument is such:

> "Given any machine which is consistent and capable of doing simple arithmetic, there is a formula it is incapable of producing as being true ... but which we can see to be true."

Let us first provide a definition of a Turing machine, which will be useful in the following argument. Alan Turing first conceived of the Turing machine in 1936, and they are essentially computational devices which allow us to understand the extent of computability.[72] These Turing machines have a set of predefined rules which allow us to capture all computable functions. In practice, the Turing machine was modeled as a machine with a tape that can be divided up into different blocks. Each of these blocks is associated with a certain symbol, and symbols can be read from the tape as well as overwritten. The Turing machine can only examine one block at a time, but it is able to move to the block immediately to the right or the left.

These Turing machines have a set of k states, an alphabet Σ which is a set of all of the symbols that can be written to the tape, and a transition function, which takes a state and

a symbol and tells the Turing machine whether to move to the left, right, stay at the same position, or halt.[8] The tape will start by using its alphabet to represent the input to the machine, and after following the transition function by reading this input, the machine is able to output to the tape the result of its computation. So, it is really the transition function which determines what it is that the Turing machine will compute. A function is said to be computable if there is a Turing machine that is able to compute it.

It may be easy to think that a Turing machine is an outdated and irrelevant model of computation given the complexity of modern programming languages and hardware systems. And it may seem especially odd to define a computable function as only one that can be computed by a Turing machine. However, we know by the Church-Turing thesis that Turing machines and any other reasonable model of computation (PCs, quantum computers, NAND-RAM machines) are functionally equivalent.[8] Thus, Turing machines provide a framework in which we can evaluate the extent and limits of computability.

That is, we can imagine a Turing machine, let us call it M, which can compute any arithmetic equation or theorem of arithmetic. Such a machine M, through its various operations and syntax would comprise a formal system, F. There are a finite number of operations and assumptions which can be encoded in M, and as a result they can be represented by symbols. These symbols can correspond to certain states and various operations that can be performed will provide the transition from one state to another. Given enough time, we

could simply write down the operations that the M carries out, and this set of operations would be a formal proof. Thus, M's output would be akin to theorems that M can prove to be true given its finite set of operations. We humans can indeed carry out elementary arithmetic, and as a result, this machine M must be able to carry out operations of elementary number theory. Since M comprises a formal system F, by Gödel's Incompleteness Theorem, there must be a Gedelian formula which cannot be proven in F. Lucas argues that this formula, although it cannot be proven by M, can be proven by a human.[44] In the same way that we can see that the Gödel sentence above is true, since we can see that G_V is true since it is unable to be proved in V. However, the formal system V cannot make this assertion. Thus, Lucas concludes that any mechanistic representation of the mind by some Turing machine could never be correct, since there will always be a Gödelian formula which the Turing machine would be unable to prove, but that a rational human being would be able to see as true. In this way, the human mind cannot be fully encapsulated by a Turing machine.

One of the major objections to Lucas's argument hinges on the consistency of human beings. A set of axioms is consistent if it lacks contradiction. What is important is that Gödel Incompleteness Theorem applies only to consistent systems.[10] It is not known whether the human mind works on a consistent set of axioms. Because the Turing machine that Lucas proposes must indeed be consistent if he seeks to apply Gödel's Incompleteness Theorem.

Critics of this line of reasoning argue that if the human mind is inconsistent, then there is nothing to stop us from building a machine which is also inconsistent which could perform all of the functions that the mind is able to.

In his work, *Shadows of the Mind*, Roger Penrose gives an updated account of this argument which is meant to be less susceptible to the objections made towards Lucas.[55] We are to assume that all methods of humanly accessible mathematical reasoning can be used to make up a formal system, *F*. And we are to take the statement "I am *F*" to mean that "*F* encapsulates all the humanly accessible methods of mathematical proof.

Penrose's argument is as follows:

> "Though I don't know that I necessarily am *F*, I conclude that if I were, then the system *F* would have to be sound and, more to the point, *F* would have to be sound, where *F* is *F* supplemented by the further assertion "I am *F*." I perceive that it follows from the assumption that I am *F* that the Gödel statement $G(F)$ would have to be true and, furthermore, that it would not be a consequence of *F*. But I have just perceived that "If I happened to be *F*, then $G(F)$ would have to be true," and perceptions of this nature would be precisely what *F* is supposed to achieve. Since I am therefore capable of perceiving something beyond the powers of *F*, I deduce that I cannot be *F* after all. Moreover, this applies to any other (Gödelizable) system, in place of F."[55]

Philosophers like David Chalmers object to Penrose's line of reasoning.[15] Chalmers specifically argues against Penrose's statement that "Though I don't know that I necessarily am *F*, I conclude that if I were, then the system *F* would have to be sound". This hinges on the assumption that we know that the human mind is sound, which is not something that

Chalmers is willing to concede.

Halting Problem

Many argue that the Halting problem, an uncomputable function provides a foundation for which to argue that human mental states and consciousness cannot be replicated computationally. Before delving into why this is the case, let us first define the Halting problem.

Let $HALT : \{0,1\}^* \rightarrow \{0,1\}$ be the function such that for every string $M \in \{0,1\}^*$, $HALT(M,x) = 1$ if Turing machine M halts on the input x and $HALT(M,x) = 0$, otherwise. That is, the $HALT$ function takes in both a Turing machine, M and an input x, and returns 1 if the Turing machine will halt on the input x or if it will run its computation infinitely.[8]

Alan Turing provided the proof for this result that went as follows[72]:

Assume for the sake of contradiction that there exists a program $HALT(M,x)$ that solves the halting problem. That is, $HALT(M,x)$ will return True if the Turing machine, M, will halt on an input x, and it will return False otherwise. Then, let us construct the following program in pseudocode:

```
def P(s):
    If HALT(s, s) then:
        while True;
    Else:
       return False;
```

In this scenario, there are two possible cases of what will occur. Either the Program s

will halt on input s. Then, the program will loop infinitely. Or, the program will not halt on an input s. So, $HALT$ will return false on input, s, s. So, the program will halt. Both of these cases contradict the definition of our $HALT$ function which means that the Halting Problem is indeed an incomputable function.[72]

Some individuals would argue that computers, since they are unable to solve the Halting Problem, cannot properly simulate a system that is able to stop its current train of thought arbitrarily and change to another line of thinking. Or, some say that humans have the ability to determine if a program will halt or not by examining the contents of the code. Each of these are arguments to defend the idea that computation is limited to the extent that it cannot properly replicate human consciousness.

Whether the Halting Problem as applied to consciousness or Lucas and Penrose's arguments succeed is a matter that will be left to the reader. Regardless, it is clear that a purely secular debate on the consciousness and the mechanistic reproduction of the human mind is quite contested. As such, the goal of this paper is to examine this problem within the larger faith tradition of Islam.

Case Studies

Prior to delving into whether Strong AI is realizable, it is helpful to examine the current state of AI technologies. Doing so will allow us to evaluate how close we may be to Strong AI, and comparing the mechanisms of computation to that of the brain may eluci-

date some key differences between our current modes of computation and human mental states.

Based off of our previously defined notion of Strong AI which affirms that suitably programmed computers can actually understand natural language and possess the same mental states as humans, it may seem as though we have already reached Strong AI. For instance, object recognition algorithms are able to identify and distinguish between various objects with the same level of accuracy as humans do. Generative AI models such as DALL-E are used to create art, and large scale language models like ChatGPT make it difficult to tell that you are interacting with an algorithm rather than a human.[13] We will take a brief dive into these technologies to understand their relationship to Strong AI.

Object Recognition

Some may argue that an algorithm like the neural network, which in many ways, mimics the way that neurons in the brain work, is a first step to building a system which can replicate human intelligence. Let us now take a look at how the neural network works to see if there is a basis for this.

The basic unit of the neural network is the node, otherwise known as a perceptron as seen in Figure 1.1. These nodes function by taking in a series of inputs.[32] These inputs are multiplied by a weight and then added together. They may then be summed up with a learnable bias term to better fit the data. This sum is passed into an activation function

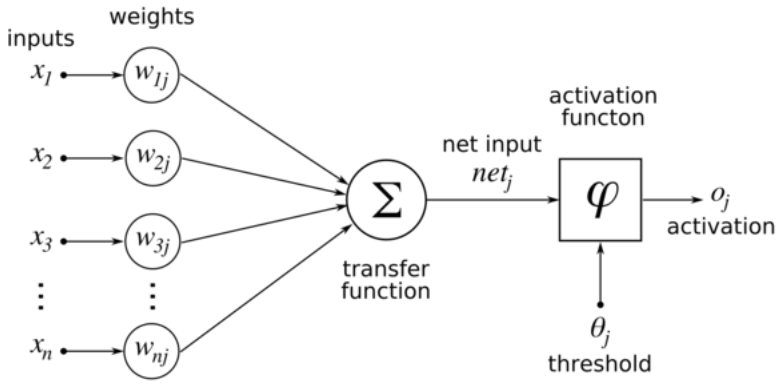


Figure 1.1: Perceptron [81]

which decides what the output of the node is going to be. The neural network functions by combining many of these nodes together into multiple layers. This includes an input layer, one or more hidden layers, and an output layer.

A basic neural network is constructed by combining these perceptrons into layers as seen in Figure 1.2.[62] As the number of layers increases, the neural network is able to perform more complex calculations. As a value is passed from one node to the next, it is multiplied by a weight. As such, the "weight" can, in a certain sense, be considered as the strength of the connection from one node to the next. That is, it determines how much of an effect the input to the node will have on its output. Using a set of inputs and known outputs, during training the neural network uses an optimization algorithm in order to find the values for the weights and biases that will map the training input to the desired output. The hope is that the calculated values for these weight and bias terms will be able to generalize to input

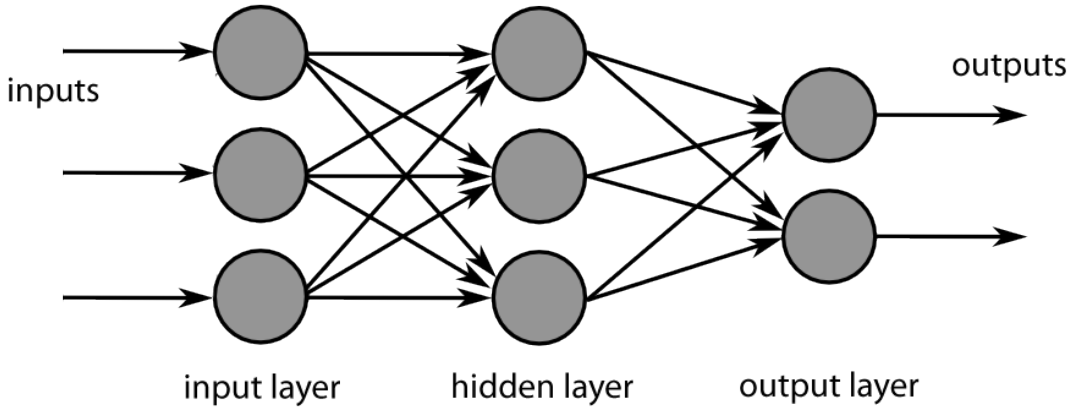


Figure 1.2: Neural Network Architecture [77]

data that is not in the training set to predict the correct output.

Neural networks have been used to recognize handwritten characters, approving bank loans, or diagnosing certain diseases in patients. And in many ways, they are able to mimic many of the behaviors that could only be done by humans mere decades ago.

Given today's technology, there exist algorithms which perform image recognition to a level as accurate as humans, and they are built upon the basic concept of neural networks. The convolutional neural network (CNN), provides the basis for many computer vision algorithms. [24] As opposed to traditional neural networks, which take a flattened vector of as input, the adjacency of pixels can give us lots of information about the subject of the image. This information would get lost if the image was simply flattened to a vector of pixels. Designed specifically for images, the convolutional neural network is able to capture the spatial dependencies of an image.

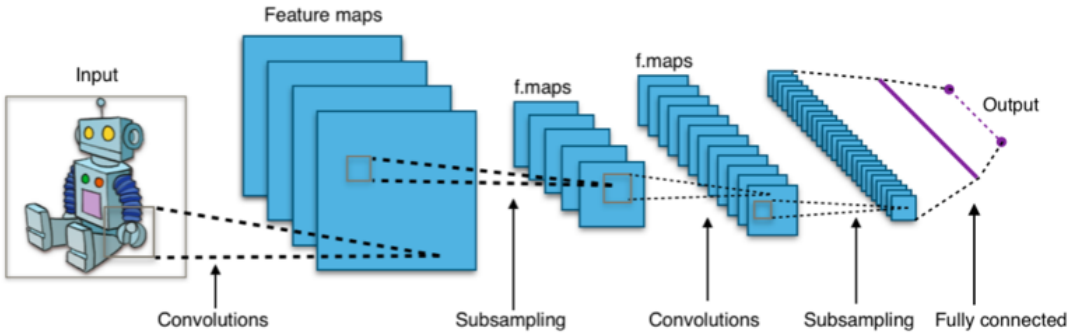


Figure 1.3: Convolutional Neural Network Architecture[78]

The main idea behind CNNs is to use a series of convolutional layers to extract features from the input data and then use these features to classify or perform some other task.[24] The architecture of a convolutional neural network can be seen in Figure 1.3.

The first layer in a CNN is usually a convolutional layer. A convolutional layer applies a set of learnable filters to the input data to generate a set of feature maps.[24] Each filter slides over the input data and computes a dot product with the input pixels at each position. This process, known as a convolution, allows us to extract high-level features from an input images. The resulting values are then summed up and passed through a non-linear activation function, such as ReLU, to introduce non-linearity.

After convolutional layers, a pooling layer is often applied.[24] A pooling layer reduces the spatial dimensions of the feature maps by taking the maximum or average value of each local region. This helps to reduce the computational cost of the network and makes the features more robust to small spatial translations.

After one or more convolutional and pooling layers, a fully connected layer is often added to map the high-level features to the output class probabilities.[24] A fully connected layer connects every neuron in the previous layer to every neuron in the next layer, and the output is passed through a softmax function to obtain class probabilities.

During training, the network is optimized by minimizing a loss function, such as cross-entropy loss, using backpropagation.[24] Backpropagation computes the gradient of the loss with respect to each parameter in the network, and updates the parameters in the opposite direction of the gradient to minimize the loss.

Overall, CNNs are able to learn hierarchical representations of the input data through the convolutional layers, which can capture increasingly abstract features as the network gets deeper. This makes them highly effective for tasks such as image classification, object detection, and image segmentation.

AI Art

Many new technologies such as DALL-E by OpenAI or Imagen by Google have demonstrated AI's ability to generate art given a text input. Figures 1.4 and 1.5 give examples of AI generated art. It is clear from these examples that the ability of AI to generate art is comparable to the ability of humans to do so.

While the implementation of these generative AI models differs depending on the tool being used, many of the current state of the art AI art generators rely on diffusion mod-



Figure 1.4: DALL-E Generated Image Given Input ”1960’s art of cow getting abducted by UFO in midwest”[79]



Figure 1.5: Image Generated on Prompt ”swimming pool filled with a galaxy on a moonlit night”[80]

els.[25] Diffusion models are a type of generative model that can learn to generate high-quality samples from complex distributions, such as images or text. They use a process called diffusion to generate samples by iteratively adding noise to a seed sample.

Training a diffusion model for image generation typically involves optimizing the model parameters to maximize the log-likelihood of the training data.[36] The training process can be divided into two main steps: (1) pre-training the model using an autoencoder or a similar architecture, and (2) fine-tuning the pre-trained model using the diffusion process.

In the pre-training step, the model is typically trained to learn a low-dimensional representation of the input images using an autoencoder or a similar architecture.[36] This involves training an encoder network to map the high-dimensional input images to a low-dimensional latent space, and a decoder network to map the latent vectors back to the high-dimensional image space. The pre-training step helps the model to learn a more efficient representation of the input images and can speed up the fine-tuning process.

In the fine-tuning step, the model is trained to generate high-quality images by simulating a diffusion process that transforms a noise vector into an image.[36] The fine-tuning process involves optimizing the model parameters to maximize the log-likelihood of the training data, which measures the similarity between the generated images and the real training images. This can be represented by a stochastic differential equation (SDE) of the form:

$$dx(t) = b(x(t), t) \cdot dt + \sqrt{2 \cdot D(x(t), t)} \cdot dW(t)$$

where $x(t)$ is the state variable at time t, $b(x(t), t)$ is the drift term that specifies the mean change in $x(t)$ at time t, $D(x(t), t)$ is the diffusion term that specifies the variance of the change in $x(t)$ at time t, and $dW(t)$ is a Wiener process that represents the increment of Brownian motion over a small time interval dt.

To optimize the model parameters, the log-likelihood of the training data is typically maximized using gradient-based optimization techniques, such as stochastic gradient descent (SGD) or Adam.[36] The gradients of the log-likelihood with respect to the model parameters are computed using backpropagation through the diffusion process and the energy function. The log-likelihood can be written as:

$$\log p(x_1, x_2, ..., x_n) = \log p(x_1) + \log p(x_2|x_1) + ... + \log p(x_n|x_{n-1}, ..., x_1)$$

where $x_1, x_2, ..., x_n$ are the training images, and $p(x_i|x_{i-1}, ..., x_1)$ is the probability density function of x_i given the previous images $x_{i-1}, ..., x_1$.

Overall, the training of a diffusion model for image generation involves pre-training the model to learn a low-dimensional representation of the input images using an autoencoder,

and fine-tuning the model by simulating a diffusion process that gradually transforms a noise vector into an image and optimizing the model parameters to maximize the log-likelihood of the training data. The fine-tuning process involves adding Gaussian noise to the image and updating the model parameters at each diffusion step, and optimizing the log-likelihood using gradient-based optimization techniques.

ChatGPT

ChatGPT is a natural language processing (NLP) model developed by OpenAI using deep learning techniques. It is based on a Transformer architecture and was trained on a large corpus of text data using unsupervised learning.

The Transformer is a type of deep neural network architecture that is widely used in natural language processing tasks, including language translation, text classification, and question answering. It was introduced in a 2017 paper by Vaswani et al. and has since become one of the most popular architectures in NLP. The transformer architecture is based on a self-attention mechanism, which allows the model to weigh the importance of different parts of the input when generating its output.[74] The self-attention mechanism allows the model to learn the relationships between words in a sentence and to understand the context in which they are used. This allows ChatGPT to generate text that is grammatically correct and semantically coherent. In traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs), the input is processed sequentially or by local regions,

respectively. However, the self-attention mechanism allows the model to attend to all the input positions simultaneously.

The Transformer model consists of an encoder and a decoder.[74] The encoder takes in a sequence of input tokens and processes them to produce a fixed-dimensional output, while the decoder takes in the output of the encoder and generates a sequence of output tokens. Both the encoder and the decoder are composed of multiple layers, each of which consists of a self-attention mechanism followed by a feedforward neural network.

The Transformer architecture is built off of previously researched Sequence-to-Sequence (seq2seq) models. These models include both an encoder and a decoder and take one sequence of text and convert them to another sequence.[70] For instance by taking sequences of words from one language and converting them to sequences of words from a different language. Long Short-Term Memory (LSTM) models had long been used for this type of procedure. LSTM models are a type of recurrent neural network (RNN) used in deep learning that are designed to handle the problem of vanishing gradients in traditional RNNs.

In traditional RNNs, the output of each neuron is fed back into the input of the same neuron in the next time step.[63] This allows the network to use its previous state to inform its current state and make predictions about future states. However, in deep RNNs, this can lead to the problem of vanishing gradients, where the gradients of the loss function with respect to the parameters of the network become very small, making it difficult for the

network to learn long-term dependencies.

LSTM models solve this problem by introducing a memory cell, which allows the network to selectively forget or remember information from previous time steps.[68] The memory cell is controlled by three gates: the input gate, the forget gate, and the output gate. The input gate controls how much new information is allowed into the memory cell, the forget gate controls how much old information is forgotten from the memory cell, and the output gate controls how much information is outputted from the memory cell to the next time step. As it relates to seq2seq models, the LSTM allows us to remember key input text that it denotes as important. This is important for language processing tasks because the sequence and context of key words are very important to grasp the meaning of a sentence.

For a seq2seq model, LSTMs would often be used for both the encoder and decoder along with an attention mechanism.[70] The attention mechanism discerns which keywords in the input text are important and is able to feed this information to the decoder to improve decoding.

The Transformer model uses this seq2seq approach, but does not use LSTM models and the RNNs that they are built upon.[74] The Transfomer model architecture is given in Figure 1.6.

The encoder is the set of layers on the left hand side while the decoder is the set of layers on the right.[74] With RNNs, we are able to keep track of how sequences are fed into the

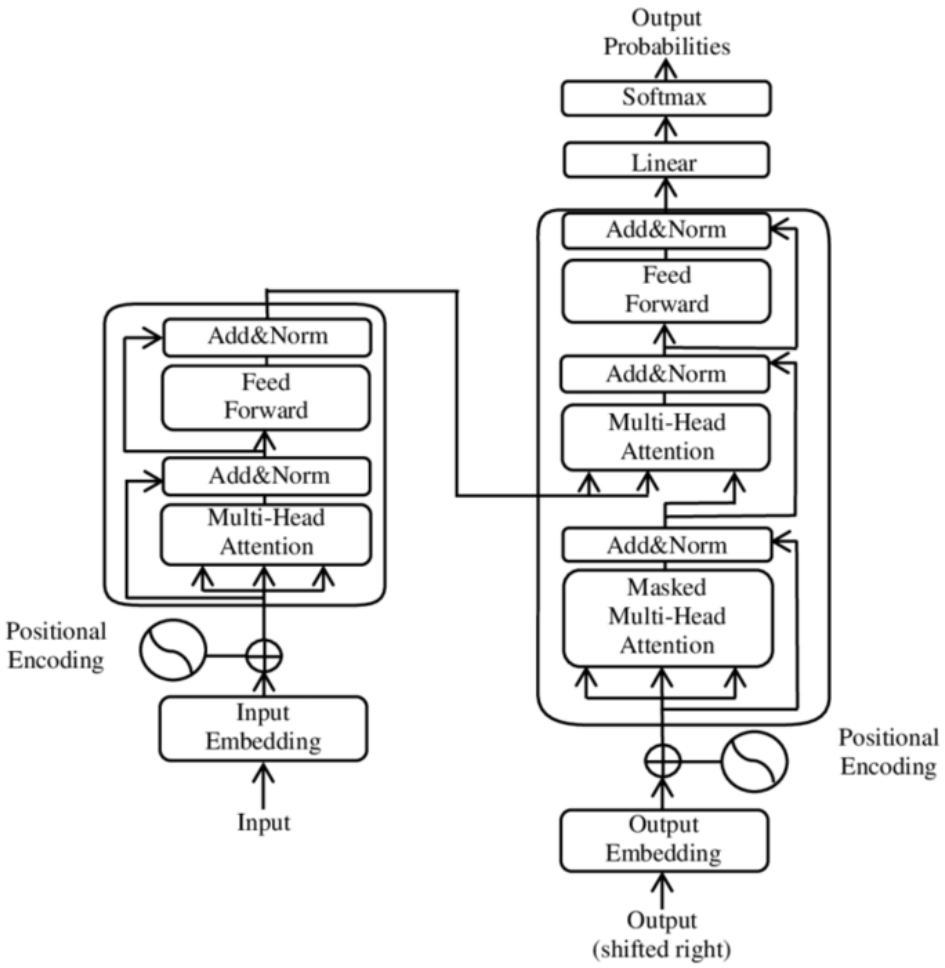


Figure 1.6: Transformer Model Architecture[82]

model and the order of the inputs. However, in this case, we do not have this ability. So, the Positional Encoding steps involve giving every word in our input sequence a relative position that is embedded into the vector representation of the word.

The Multi-Head attention layers make use of multiple Scaled Dot-Product Attention layers.[74] They allow the model to attend to different parts of the input sequence and selectively focus on the most important information.The scaled dot product attention layer takes the following inputs: a set of queries (Q), a set of keys (K), a set of values (V), and a scaling factor.[56] The queries and key-value pairs are typically derived from the input sequence, and the scaling factor is used to adjust the magnitude of the dot product to prevent it from becoming too large or too small. The queries are a set of vectors that represent the current position in the output sequence that the attention layer is generating. The key and value vectors are learned embeddings of each token in the input sequence. They are used to compute a set of attention weights that indicate how much each position in the input sequence contributes to the current output position.

The attention layer first calculates a similarity score between each query and each key, using the dot product of their respective embeddings.[40] To compute attention scores, we take the dot product of the Query vector Q with each Key vector K. The resulting scores are then divided by the square root of the dimensionality of the embeddings, which is the scaling factor.

$$\text{score}(Q, K_i) = \frac{Q \cdot K_i}{\sqrt{d_k}}$$

The scores are then passed through a softmax function to obtain a set of weights that indicate how much attention should be paid to each value in the key-value pairs.[40]

The final output of the attention layer is a weighted sum of the value vectors, where the weights are determined by the softmax scores.[40] This weighted sum represents a context vector that captures the most relevant information from the input sequence for the current query. Thus, the attention mechanisms can be described as follows:

$$\text{Attention}(Q, K, V) = \textit{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Scaled dot product attention layers are particularly effective for modeling long-range dependencies in sequences, as they allow the model to selectively attend to different parts of the input sequence based on their relevance to the current task.[74] They have been shown to be highly effective in a wide range of natural language processing tasks, including machine translation and language modeling.

The training data consists of pairs of input sequences and corresponding target sequences.[56] The input sequences are usually padded to a fixed length, and the target sequences may be

truncated or extended as needed. To train the model, the input sequence is passed through a stack of Transformer encoder layers to obtain a sequence of encoded vectors, one for each token in the input sequence. The decoder takes the encoded input sequence and generates a sequence of output tokens, one at a time. At each time step, the decoder takes as input the encoded vectors of the input sequence and the previously generated tokens. At each time step, the model generates a probability distribution over the possible next tokens. The loss is then calculated using the cross-entropy loss between the predicted probability distribution and the true next token. The loss is backpropagated through the decoder and encoder layers to update the parameters of the model using gradient descent. These steps are repeated for each training example in the dataset.

Once the Transformer model has been trained, it can be used to generate output sequences for new input sequences by passing the input sequence through the encoder and using the decoder to generate the output sequence.[74] During inference, the model generates output tokens one at a time, using the previously generated tokens as input, until an end-of-sequence token is generated or a maximum length is reached.

Although this is how a Transformer model is trained in general, ChatGPT has some additional modifications which make it the state-of-the-art tool that it is. ChatGPT was trained using a variant of the Transformer architecture called GPT (Generative Pre-trained Transformer).[19] The training of ChatGPT involved first collecting a large corpus of text

data, consisting of a diverse range of text sources such as books, news articles, and web pages. The text data was tokenized into sequences of subword units using the Byte Pair Encoding (BPE) algorithm. This allows the model to handle words that are not in its vocabulary, as they can be broken down into subword units. The pre-training step involved training the GPT model on a large dataset of text data using a technique known as "unsupervised learning".[19] The model is trained to predict the next word in a sequence, given the previous words in the sequence. This is done using a variant of the language modeling objective, where the model is trained to maximize the likelihood of the next word given the previous words in the sequence. After pre-training, the GPT model was fine-tuned on a smaller dataset of text data for a specific task, such as question answering or text generation. The fine-tuning step involves updating the pre-trained model using supervised learning on a task-specific dataset, where the model is trained to minimize a task-specific loss function. Overall, the pre-training and fine-tuning steps allow ChatGPT to learn a rich representation of natural language that can be used for a wide range of natural language processing tasks.

When a user inputs a prompt or a question, ChatGPT processes the input using its trained model and generates a response based on the learned patterns and relationships between words. The model's responses are not pre-scripted, but generated on-the-fly, using a probabilistic approach to generate the most likely output based on the input.

To improve its performance, the model has been trained on massive amounts of text data, including books, articles, and web pages, which have helped it to learn the nuances of language and develop a deep understanding of how words and phrases are used in context.

We will return back to these case studies in later chapters to serve as a basis for evaluating our progress on the road to Strong AI. We will examine whether it is the case that these examples lay the foundation for future development which could potentially result in Strong AI.

Without consciousness the mind-body problem would be much less interesting. With consciousness, it seems hopeless.

Thomas Nagel

2

Philosophy of Mind: A Brief Introduction

Fundamental to the study of philosophy of mind, is the mind-body problem. The mind-body problem seeks to answer the question of how the mind and the body are related. That is, how do mental properties affect physical properties and vice versa. Mental properties are related to the thinking and rationality. They are the basis for phenomena

such as consciousness or intentionality. Physical properties related to the body's physical makeup and its movement through space and time.[60] Physical properties are said to be equally observable in space by all, whether that be the movement of an arm or the motion of atoms at a microscopic level. Mental properties, on the other hand, exist at an individual level. For instance, how the color red appears to one person may be different than another person. Or, the level of pain someone feels, although partially observable through one's behavior, is only directly observable by the individual who is experiencing it.

There are a couple of main components of the mind-body problem. The first is an ontological question related to the difference between mental and physical states.[60] Although there appears to be an intuitive distinction between the two, are they really distinct entities? Are mental and physical states the same thing? Are mental states simply a type of physical state? The second component of the mind-body problem is related to a casual question. What exactly is the interplay between physical and mental states? Does one influence the other? Do mental states cause physical states or vice versa? There have been multiple theories posited to solve the mind-body problem; several of the main schools of thought on these matters will be outlined below.

Two distinctions in philosophy of mind include dualism and monism.[37] Dualism is the belief that the mind exists as distinct from the body.[60] There are different types of dualists as well. Substance dualists believe that the mind and body are entirely different substances.

Whereas property dualists posit that the mind arises from distinct properties that cannot be explained entirely by the physical nature of the brain, but they do not go so far as to classify the mind as a different substance from the body.

Monists believe that, at an ontological level, the mind and the body are not distinct entities. As with dualists, there are different types of monists. The most widespread form of monism appears to be physicalism (also known as materialism), the belief that the only thing to exist are physical and things that may appear to transcend physical nature are nonetheless physical phenomena.[69] Idealists, on the other hand, believe that all that exists is the mind, and anything that appears physical is merely an external world that is created by the mind. Others such as neutral monists or dual-aspect monists exist but since these are minority views, they will not be elucidated further in this thesis.

Physicalism

In this section, we will outline a few different forms of the most well-known type of monism which is material monism, otherwise known as physicalism. Other types of monism include idealism and neutral monism.[45] The schools of thought that fall under physicalism include behaviorism, identity-theory, and functionalism.[69] It can be argued that functionlism does not necessarily fit neatly under physicalism, however, for the sake of this work, we will treat it as such. However, there are certainly more than just these three, and this project does not include the full scope of beliefs that are encapsulated under both monism and

physicalism.

Behaviorism

Behaviorism is the view that mental states are reducible to physical states and behavior. Although even amongst behaviorists there appear to be differences in beliefs and attitudes, there appear to be a few core tenets. Firstly, behaviorists would argue that psychology is merely the science of behavior.[26] Humans do not have mental states or an inner mind different from the way that they behave. Secondly, an individual's behavior does not depend on internal mental processes or events. Rather it is the outside environment and physical relationships that result in behavior, not mental events. And thirdly, any reference to mental phenomena or concepts should be rephrased in terms of behavior.

To concretize this school of thought, it is helpful to give an example to clarify. One example of a phenomenon that is considered a mental property or state would be pain.[58] To a behaviorist pain is nothing more than the behavior that is associated with pain. This could be a grimace when accidentally touching a hot surface or another response that is normally associated with pain. In other words, pain is not an internal process but rather is simply a set of behaviors.

Identity-Theory

Let us begin by examining an objection to behaviorism, the Super-Spartan case espoused by Hilary Putnam in his essay, "Brains and Behaviour". In this essay, he lays out the thought

experiment of super-spartans.

> "Specifically, let us take our 'super-spartans,' and let us suppose that after millions of years they begin to have children who are born fully accultured. They are born speaking the adult language, knowing the multiplication table, having opinions on political issues, and inter alia sharing the dominant spartan beliefs about the importance of not evincing pain (except by way of a verbal report, and even that in a tone of voice that suggests indifference). Then there would not be any 'unconditioned pain responses' in this community (although there might be unconditioned desires to make certain responses—desires which were, however, always suppressed by an effort of will). Yet there is a clear absurdity to the position that one cannot ascribe to these people a capacity for feeling pain."[57]

Putnam's thought experiment outlines a potential major issue with behaviorism. If mental states are reducible to a cluster of behaviors, then how does one account for someone feeling a certain "mental state" without exhibiting the behaviors that are associated with that state which certainly appears to be a possibility.

One of the main issues with behaviorism was its argument that mental processes, such as emotions, thoughts, and perceptions, be reduced to mere patterns of behavior. Behaviorism fails to account for the complexity and flexibility of human behavior. The macro responses to external stimuli may vary too much from individual to individual to be mapped onto individual emotions or states.

Identity theory seeks to serve as a response to the limitations of behaviorism, which ignored the internal mental processes that underlie behavior. The most well-known version

of identity theory is the mind-brain identity theory. Similarly, to behaviorism, mind-brain identity theory rejects the notion of a separate realm of mental processes.[67] However, while behaviorism rejects the notion of mental states altogether, identity theorists posit that mental processes are identical to specific physical states and neurophysiological processes in the brain. That is, a mental state is a brain process.

Going back to our example of feeling pain. A behaviorist may identify pain as the behavior of grimacing upon touching a hot surface. However, to an identity theorist, the mental state of pain is associated with the specific neuron firing that occurs as a result of touching the hot surface.

Functionalism

To a mind-brain identity theorist who argues that a mental state is equivalent to brain processes, how do we account for organisms that may not have human brains? This was the question that was posed by David Lewis in his essay, "Mad Pain and Martian Pain".

> "Also there might be a Martian who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its physical realization. His hydraulic mind contains nothing like our neurons. Rather, there are varying amounts of fluid in many-inflatable cavities, and the inflation of any one of these cavities opens some valves and closes others. His mental plumbing pervades most of his body—in fact, all but the heat exchanger inside his head. When you pinch his skin you cause no firing of C-fibers—he has none—but, rather you cause the inflation of many smallish cavities in his feet. When these cavities are inflated, he is in pain. And the effects of his pain are fitting: his thought and activity are disrupted, he groans and writes, he is strongly motivated to stop

> you from pinching him and to see to it that you never do again. In short, he feels pain but lacks the bodily states that either are pain or else accompany it in us."[43]

Since the Martian writhes and groans when in pain, he would be in pain by a behaviorist account. However, from an identity-theory approach, the mental state of pain would need to be equivalent to a brain process. However, this Martian does not have a brain, in the anatomical sense. So, are we to say that this Martian, simply because they do not have a brain and its associated processes in the way that humans do, is incapable of feeling pain? The intuitive answer is no, and this Martian is indeed capable of feeling pain. But identity-theory does not account for this. Functionalism, however, is able to accommodate such a view.

Functionalism does not equate mental states to physical processes of the brain but rather to the functional roles that they play.[42] This allows for functionalists to generalize beyond just the brain. While an identity theorist would be held to argue that mental states can only occur in humans because they are equivalent to mental states, a functionalist argues that mental states can be realized by many different types of physical systems, not just the brain. This is known as the principle of multiple realizability.

Dualism

Plato was one of the earliest philosophers to discuss dualism. In his *Phaedo*, he discusses how true substances are not material entities but rather they exist as immaterial forms.[60]

Descartes, in his *Meditations* was one of the first to propound a concept of dualism in the West with his substance dualism. In this conception, he believed that there are two types of substances which include matter and the mind.[35] What makes something matter is that it exists in space.[60] What makes something a mind is that it thinks. One of the main issues with Descartes' conception was the problem of how the two substances, which are entirely different in nature, can interact. There have been many different conceptions of dualism over the years, however, in the context of philosophy of mind, it is the notion that mental and physical entities are different in nature.

Dualism and these various forms of physicalism make up many of the main schools of thoughts as it relates to the mind-body problem. However, there is also the "hard problem of consciousness".[14] This relates to how humans are able to have phenomenal experience. That is, how can we explain the internal subjective experiences that individuals have when they are exposed to a certain stimulus? While we will not expound the various beliefs and schools of thoughts related to this problem as it is beyond the scope of this project, there are many other views in philosophy of mind that seek to provide a basis for these phenomenal experiences beyond just dualism and physicalism.

It appears that the main issue with simulating human mental states in machines is the problem of subjective experience. Where does it come from? How do we replicate it? While this concept is not explicitly stated in the texts of many Islamic scholars, we will dive into

the origin of intellect and the uniquely human abilities of thought and rationality in the following chapter.

Case Studies

From a functional perspective it appears that all of these case studies could indeed serve a part in constituting a human mind. The CNNs are able to recognize objects, generative models are able to mimic human creativity to create art, and the large scale language models can output text in a manner similar to a human. In this way, they all serve functional roles of human activity, and as such, would be akin to the human mental states of the same processes in the body according to a functional perspective.

Some might make the argument that these technologies, more specifically the generative AI art models and the large scale language models, while close to simulating the human abilities that they are meant to mock, do not quite replicate a humans abilities to the skill and level that humans are able to do them. However, we can see that a mere decade ago object recognition tasks, while impressively accurate could not mimic a human's ability to discern objects. Now, due to a massive increase in data and training, are able to, in some instances, outperform humans. There is little reason why, after training with enough data, the same cannot be said about the other two technologies. As such, it seems that from a functional perspective, Strong AI is indeed possible.

Furthermore, if we were able to create a mechanistic form that incorporates these al-

gorithms. For instance, a robotic body with a similar anatomy to humans, it appears as though Strong AI could indeed be feasible from a behaviorist perspective as well. Simulating the behaviors of certain states such as pain does not seem entirely impossible in the future given where we already are in terms of technological development. And it could only be a matter of a few decades before we create machines that behave in the same way that we humans do. From a behaviorist perspective, if you incorporated the algorithms in this case study into a larger robotic form, Strong AI indeed appears feasible as well.

Mind-brain identity theory and dualism are a bit trickier to deal with. Since, to an identity theorist, mental states are brain states, then it appears as though any computational method would not be sufficient in reproducing human mental states. As it stands, these algorithms rely on math which function quite differently from the way that the human brain works. So, it appears as though Strong AI through computation would not necessarily be achievable from an identity theorist point of view. However, it is unclear whether, given a complete understanding of the way the brain works, a mathematical simulation of the brain within a computer would be enough to constitute mental states. Furthermore, it appears that for Strong AI to be achievable from a dualist perspective, the mind, which exists as an entity of a different nature from the body, would need to somehow become instantiated into a machine. It is again unclear as to how this would happen, but it does not necessarily rule out Strong AI.

The writing of a man is the symbol of his intellect and the evidence of his merit.

Imam Ali

3

The Islamic Perspectives

Prior to understanding Islam's view on the realizability of Strong AI, it is necessary to understand the nature of the soul in Islam. Western psychology tends to place great emphasis on the brain and the mind as the locus of rationality and cognition; however, the Islamic understanding takes quite a different approach to the matter. Under many Islamic

perspectives, the rational faculties and the intellect are an outgrowth of the soul.

In early Islamic thought, there seemed to be two main branches of thinking.[47] The first is the kalam, or Islamic speculative theology, which comprised of theologians (*mutakallim*) who developed rational bases for answering questions related to defending the religion and the tenets of Islamic faith. The second is the falsafa, which was an outgrowth of the Hellenistic tradition and built upon mainly Platonic and Aristotelian ideas. Kalam theologians and Arabic philosophers (*falasifa*) were two distinct groups of scholars in the Islamic intellectual tradition who approached questions about the nature of God, the universe, and human existence from different perspectives.

Kalam theologians were concerned primarily with defending and elaborating Islamic doctrine using reason and logical argumentation. They emphasized the importance of revelation as a source of knowledge and believed that the teachings of the Qu'ran and the Hadith must be interpreted in a manner that was consistent with reason. Kalam theologians were influenced by Greek philosophy, especially the works of Aristotle and Plato, but they rejected some of the more radical ideas put forward by Greek philosophers, such as the eternity of the world.

Arabic philosophers, on the other hand, were influenced by Greek philosophy to a much greater extent and saw reason and logic as the primary means of understanding the world. They were interested in a wide range of topics, including metaphysics, ethics, and politics,

and they sought to develop a comprehensive philosophical system that could explain the nature of reality. Arabic philosophers were known for their emphasis on the importance of the intellect, and they believed that human beings could attain knowledge of the divine through the use of reason.

The Views of the Mutakallimun

The Mutakallimun had an atomistic metaphysical system.[66] They believed that the world is made up of atoms. Not atoms in the scientific sense we think of today (which can be divided into protons, neutrons, and electrons), but indivisible particles which cannot be further broken down. These atoms may have properties/accidents that either inhere into individual atoms or the composition of many atoms can result in properties as well. All created beings were considered compositions of atoms and accidents that were held together by the power of God.

The kalam theologians were not a monolithic group, and there were several different schools of thought within the kalam tradition, each with its own unique approach to theology and methodology. The Ash'ari School was one of the most prominent schools of kalam theology, named after its founder Abu al-Hasan al-Ash'ari (d. 935 CE).[52] The Ash'ari school emphasized the importance of divine unity and the doctrine of predestination. It was concerned not only with religion but also with epistemology and the philosophy of nature and is one of the main Sunni schools of Islamic thought. The Maturidi school was

founded by Abu Mansur al-Maturidi (d. 944 CE) and shared many beliefs with Ash'ari but put a greater emphasis on reason than Ash'ari. The Mu'tazili school was founded in the 8th century and was known for its emphasis on the importance of the oneness of God, human responsibility, and the creation of the Qur'an.[12] The Mu'tazili school also believed in the concept of human free will, and it rejected the idea of predestination. There were other schools such as the Athari and Jahmiya, but it is beyond the scope of this thesis to explore them all. As it relates to kalam theology, we will mostly be delving into the Ash'ari school of thought for this paper.

In Ash'ari theology, what makes a being animate is that the property of life inheres in its atoms. For Ash'ari scholars, there was never a need to elucidate a concept of a soul or spirit.[66] The attributes that comprise a human are all properties which inhere in the atoms that compose the human body. These include properties such as knowledge, thinking, and perception. However, this lack of a conception of the soul was at odds with frequent mentions of spirit (*ruh*) or self (*nafs*) of human beings in the Qur'an and Hadith. As such, Ash'ari kalam theologians would seek to formulate conceptions of the soul which satisfied this need, but they often admitted to these conceptions not being adequate enough to satisfy the rigorous analysis usually associated with the other works of kalam theology. Some had a conception similar to Aristotle, where Aristotle believed that the soul was the form of the body) and believed that the soul could be equated to the body. However, others be-

lieved that it was a kind of subtle body (*jism latif*) which inhabits the body.[34]

Many of the early kalam theologians were occasionalists. They believed that God produced the world at a specific point in time, and He recreates everything within it from one moment to the next.[41] They believed that it was God who caused all things. For instance, to a kalam theologian, it would be God who created the atoms which make up my hair and the accident of blackness which inheres in these atoms. And if my hair were to change color, perhaps by dying it, it would be God who choosing not to recreate it with blackness and creating it with the different color instead.

Human Action & Causality

Many Mu'tazilite theologians had an exception to this general rule of occasionalism. In order to reconcile ideas about divine justice, they believed that humans had agency and were able to create their own actions. If humans did not have the power to produce their own actions, then how could they be held responsible for their actions and accountable by God? As such, according to Mu'tazilite theologians, both humans and God had agency since humans were able to cause certain accidents.[59] They believed that accidents could be directly produced by humans or produced by generation. The former involves actions such as choosing to raise one's arm. The latter occurs when a human creates the cause that resulted in the accident rather than creating the accident itself. This could involve something like breaking a window by throwing the stone. The breaking of the window was not di-

rectly caused by the individual throwing the stone, but it is certainly a result of it. Humans, since they created the cause of the of the accident that was produced by generation, would still be responsible for the accident.

Ash'ari disagreed with the Mu'tazilite view as he believed that it inaccurately downplays the power of God. He believes that it is God who is the sole arbiter when it comes to producing atoms and accidents.[59] The Ash'ari point of view is that God creates powers which can be acquired by humans to perform their actions. This concept of acquisition is denoted by the term *kasb* in Arabic. Although the concept of acquisition would likely require more elucidation then what can be offered here, we will briefly try to summarize the Ash'ari concept. For instance, let us go back to our example of breaking a window by throwing the stone. The person who throws the stone does not break the window. Rather, the breaking of the window is a consequence of the action. It is God who creates the power for the window to break. This power of the person to break the window is the power of the specific action of throwing the stone in a certain manner at a certain time. It is God who creates this power and it is an accident that is acquired by you temporarily. One could argue that the stone flying through the air as a result of being thrown is also a consequence of the action. One could continue arguing that the movement of the limbs is a consequence of the action of thinking about making voluntary movement. The line between an individual's action and the consequence of it becomes seemingly unclear. In the Ash'ari conception, what

is within the individual's power is extremely limited. And it is these consequences which are all in the realm of God's power which are acquired by the individual temporarily.

The Views of the Falasifa

Much of the work of the Arabic philosophers was drawn from translations of Aristotle's *De Anima* and *Parva Naturalia* and Greek commentaries on the works of Aristotle. Some philosophers also took inspiration from the foundational work of Plato. However, since most of the works of Plato were not translated into Arabic, Arabic philosophers drew from neo-Platonists such as Plotinus, Poryphry, and Proclus.[18] In all, the falsafa tradition was deeply influenced by Neoplatonism and Aristotelianism.

The Greek Tradition

Greek thinkers were instrumental in informing the conceptions of the soul propounded by Arabic philosophers. In his dialogue, *Phaedo*, Plato outlined his beliefs as it relates to dualism and his theory of Forms[23]. His theory of Forms is meant to explain the relationship between the physical world and abstract concepts. According to Plato, the physical world is merely a shadow or a reflection of the world of Forms.

Plato believed that the world of Forms is eternal, unchanging, and perfect, while the physical world is imperfect, transient, and constantly changing. The Forms are abstract, universal concepts that exist outside of time and space, and they are the true reality that can be apprehended only by the mind, not the senses[65]. Thus, human bodies are merely

imperfect instances of Forms. It is these Forms which allow us to understand the world around us as they provide the foundation for intelligibility. It is an understanding of the Forms which is what the intellect uses as a basis for understanding.[60]

It is clear that Plato is a proponent of a dualist conception of the soul. In the *Phaedo*, Plato not only argues for the immortality of the soul, but he believes that the intellect itself is an immaterial entity.[23] This is because in order to grasp the Forms, which are immaterial, then the intellect must be immaterial as well, so as to be disposed to apprehending the Forms. The soul is in a constant effort to leave the body so it can subside in the real of the Forms and perfectly understand them which may require various reincarnations before this can occur.

Plato's conception is known to have some issues which he never addressed. For instance, what is the association between a soul and the body? How does a particular soul become associated with a certain body and how does this relationship change prior to birth and after death of the human body? Aristotle did not share Plato's belief in the Forms as existing entirely separately from their instances in the physical realm.[60] Rather, he believes that forms or universals are not abstract entities that exist independently of particulars, but rather they are inherent in the particulars themselves. In other words, forms are not separate from the physical world, but are inherent in it.[30]

Aristotle believed that every particular thing or substance has a form, which is its essence

or defining characteristic that gives it its identity.[53] For instance, the form of a human being is what makes a particular individual a human being, and distinguishes it from other things that are not human beings. As it relates to the soul, this conception gave Aristotle the ability to say that the soul is simply the form of the body. That is, the soul is simply one's essence as a human being.

Notably, Aristotle's definition of the soul in *De Anima* permeated into the thought of many of the falasifa. In this work, Aristotle considers the soul to be the animating principle of the human form, more specifically, "the actuality of a body that has life".[31] Furthermore, Aristotle conceived of the soul as tripartite, with three different faculties. These included the nutritive/vegetative which is present in all living things and governs growth and reproduction. The sensitive/animal faculty belongs to both animals and humans and is associated with voluntary motion and perception. And lastly, the intellective/rational faculty is present only in humans and gives them the ability to carry out higher level thinking.

Aristotle's treatment of the soul as a property of the body rather than an explicitly immaterial substance as stated by Plato has resulted in many thinking that Aristotle's theory of the soul is built on materialist grounds. However, this is an active and ongoing debate.[60]

However, it is clear that Aristotle did not localize the intellect to any physical entity even though he believes that it is a part of the soul.[31] He argued that a material entity would not be able to understand all of the forms, and it must be an immaterial substance which is able

to grasp these intelligibles. He posits that physical entities are limited in the range of things that they are able to understand. For instance, the eye can perceive light but is not able to perceive sound. Similarly, if the intellect was purely a physical entity, it would only be able to perceive a limited range of things that are physical in nature. However, because we are able to think about any material thing and are not limited in this sense, the intellect must be immaterial.

While there is still debate on Aristotle's metaphysics and whether he believed that the soul was an immaterial substance, he certainly allows for the possibility that the human intellect is separate from the body but also potentially separate from the rest of the soul. Regardless, Arabic philosophers built upon the ideas that were conceived in the Greek tradition to devise their own accounts which they sought to align with the beliefs of Islam.

Al-Kindi

Al-Kindi (also known as Abu Yusuf Ya'qub ibn Ishaq al-Kindi) was a prominent Islamic philosopher, scholar, and scientist who lived during the 9th century (c. 801–873 CE).[52] He is considered to be one of the most important figures in the development of Islamic philosophy and was known as the "Philosopher of the Arabs." Al-Kindi wrote extensively on a wide range of subjects, including metaphysics, theology, ethics, and politics. He also made significant contributions to the fields of mathematics, astronomy, and medicine. One of his key contributions to philosophy was his argument for the immateriality of the soul.

In his work, *That There are Incorporeal Substances*, Al-Kindi applied ideas from Aristotle's *Categories* to show that the the human soul is an immaterial substance.[1] He first shows that the soul is indeed a substance and then goes on to prove that it is immaterial. From *Categories*, he argues that the essence of some thing will share both the name as well as the definition of that thing. By Aristotle's conception, the soul is simply the essence of the human being, and the human being is a substance. Therefore, the soul must also be a substance. He then goes on to claim that the soul is also a species because the soul is an intellectual form, and intellectual forms are said to be species as per Aristotle's conception. These species are immaterial, and as such, the soul must also be immaterial.

In this line of reasoning, it appears that al-Kindi mixes up the idea of species in *Categories* with the concept of forms in *De Anima* and *Metaphysics*.[3] Al-Kindi's conception of the soul also brings to mind various questions. For instance, how can many human souls be individuated when they are all a single species of human? While al-Kindi as an early philosopher of Islam contributed greatly to Islamic thought and had written on the metaphysics of the soul, there were numerous other philosphers who had their own opinions on the nature of the soul.

Al-Razi

Abu Bakr Muhammad ibn Zakariya al-Razi (also known as Rhazes or Rasis) was a Persian physician, philosopher, and alchemist who lived during the Islamic Golden Age in

the 9th and 10th centuries (c. 865–925 CE).[20] He is considered one of the most important figures in the history of medicine. In addition to his medical work, al-Razi was also a philosopher who wrote on various subjects, including ethics, metaphysics, and epistemology. He was critical of the Aristotelian system and emphasized the importance of empirical observation in the pursuit of knowledge.

Al-Razi rejected Aristotle's account of the soul and prefers an approach drawn from Plato's *Timaeus*.[4] He agrees with a tripartite conception of the soul into the desiderative, the spirited, and the rational as outlined in the *Republic* which is relatively aligned with Aristotle's division as well.

Al-Razi suggests that the soul may be immortal, but he does not provide a clear or definitive explanation of how this is possible.[4] He discusses the idea of reincarnation and suggests that the soul may transmigrate into a new body after death.

Plato believed that there was a divine, living force that permeated and animated the entire universe, which he called the "world soul" or "cosmic soul".[75] According to Plato, the world soul was responsible for maintaining order and harmony in the universe. It was the source of all life and motion, and it controlled the movement of the celestial bodies.[17] The world soul was also the source of human souls, which were believed to be fragments of the world soul that had become detached and individualized. Plato believed that the world soul was immortal and divine, and that it existed outside of time and space. He believed that it

was the ultimate cause of all things, and that it provided a connection between the material world and the realm of the Forms or Ideas. This idea of a world soul was central to his philosophy and can be found in his work the *Timaeus*.[75] Al-Razi agrees with Plato on this idea of the world soul. Since al Razi was a proponent of Plato's metaphysical scheme as it relates to the soul, it appears that he would also take a substance dualist approach to the relation between the soul and the human body, similar to Plato.[4]

Although al Razi was an important figure in Islam and wrote on the theory of the soul, we find much more written on the topic in the works of Ibn Sina.

Ibn Sina

Ibn Sina, also known as Avicenna in the Western world, was a Persian polymath who lived during the Islamic Golden Age.[2] Ibn Sina was a prolific writer and made significant contributions to various fields, including philosophy, medicine, astronomy, mathematics, and theology. His works were influential not only in the Islamic world but also in Europe, where his writings were translated into Latin and studied extensively during the Middle Ages.

Although there are key differences between the beliefs of Ibn Sina and Aristotle, much of Ibn Sina's core doctrines have a grounding in Aristotelian thinking.[28] First and foremost, Ibn Sina agrees with Aristotle's vision of the tripartite soul and the belief that the soul is what gives the body life and allows it to maintain its existence.[28,39]

Ibn Sina agrees with the idea that there are three different faculties of the soul which correspond to different types of life.[5] These include the vegetative which is associated with self-nutrition, growth, and reproduction, the animal, which corresponds to voluntary motion and perception, and finally, the human, which allows for the ability to think and act rationally.

Although, in his *De Anima*, Aristotle never gives an account of whether the soul and human intellectual capacities are material or immaterial, but Ibn Sina is a proponent of substance dualism.[5] That is, he believed that the human essence is indeed not a material substance, but rather exists outside the realm of physical substance.

The combination of Aristotle's psychological framework with a belief in substance dualism opens his conception to a number of difficulties, but Ibn Sina tries to grapple with them in his work, *Shifa' Fi al-nafs*.[39] One of the main consequences of his views is the difficulty with how an immaterial substance is connected to a material body, especially when essences are supposed to be indivisible. More specifically, how can a single unified essence instantiate itself into many particular instances of human individuals?

The answer that Ibn Sina provides is that the human essence can be mapped into physical forms in a manner similar to how objects are mapped to their unique coordinates in time and space.[39] For instance, if we have two circles of the same radius existing simultaneously, then the way to distinguish the two would be by attributing each one with unique

coordinates in space. Similarly, for the individuation of the human soul to occur, there must be distinct human bodies which receive the individuated souls. In this way, the human essence is instantiated into an individual soul when a human body is formed. Thus, there is a mutual dependence between the soul and the body. While the body requires the soul to become living, the soul requires an individual body to come to existence and instantiate itself as an individual entity that is distinct from the one, indivisible human essence. More specifically, it is the Active intellect which emanates an individual human soul, and according to Ibn Sina, this only occurs when there is a suitable human form being developed inside a mother's womb.[39] As such, the existence of a human body is a necessary but not sufficient condition for the existence of an individual soul that is associated with that body.

Although Ibn Sina's philosophical stance does not represent a complete departure from Hellenistic thought, he introduces new concepts of ideas that were not present before his time. One of these is the concept of self-awareness which does not appear to have been discusses in the Peripatetic tradition.[39] Although there was some allusion to the idea of self-awareness in prior works such as *Kitab mu'adhala al-nafs* or the works of Mu'ammar ibn Abbad al-Sulami, Ibn Sina appears to be the first to integrate the idea of self-awareness into a broader philosophical schema.

Ibn Sina posits the existence of a soul as necessary for the living nature of human be-

ings.[28] The extent of the human's faculties which involve nutrition, growth, reproduction, perception, and voluntary movement are too broad and flexible to not be governed by some sort of entity that controlled these actions. To Ibn Sina, the laws of the physical world were not enough to describe just how an individual is able to carry out the full range of abilities that they have with the level of freedom they are able to.

Although Ibn Sina posits the existence of a soul, he considers the question of whether this human soul is an immaterial substance or a material form which is the case for plant and animal souls by his account.[39] He offers the "flying man" thought experiment as a means of showing that the human soul is indeed an immaterial substance.

> "So we say: one of us must imagine (*yatawahhama*) himself so that he is created all at once and perfect but his sight is veiled from seeing external [things], that he is created floating in the air or in a void so that the resistance of the air does not hit him – a hit would have to sense – and that his limbs are separated from each other so that they do not meet or touch each other. [He must] then consider whether he affirms the existence of his self (*wujuda dhatihi*). He will not hesitate in affirming that his self exists (*li dhatihi mawjudatan*), but he will not thereby affirm any of his limbs, any of his intestines, the heart or the brain, or any external thing. Rather, he will affirm his self (*dhatahu*) without affirming for it length, breadth or depth. If it were possible for him in that state to imagine (*yatakhayyala*) a hand or some other limb, he would not imagine it as part of his self (*dhatihi*) or a condition in his self (*shartan fi dhatihi*). You know that what is affirmed is different from what is not affirmed and what is confirmed is different from what is not confirmed. Hence the self (*dhat*) whose existence he has affirmed is specific to him in that it is he himself (*huwa bi 'aynihi*), different from his body and limbs which he has not affirmed. Thus, he who takes heed (*al-mutanabbih*) has the means to take

> heed of (*yatanabbaha*) the existence of the soul (*wujudi al-nafs*) as something different form the body – indeed, as different from any body – and to know and be aware of it (*annahu 'arifun bihi mustash'irun lahu*)."[39]

In sum, Ibn Sina considers a man suspended in the air, stripped of his sense perception. Furthermore, they do not have any limbs or bodily form. They are completely deprived of any form of stimulus or input. Of note, Ibn Sina considers this man to be created all at once without any contact with the environment. As a result, they have no experiences, memory, or imagination for this would require a basis in the physical realm and sensory data. Ibn Sina would contend that this "flying man", although they have no perception, imagination, or experience, would still be able to affirm their own existence, that they are aware of their being. In this way, the soul, which affirms its own existence, must exist outside the realm of physical substance, demonstrating Ibn Sina's dualist perspective.

Al-Ghazali

No discussion of Islamic history would be complete without an outline of the view of Al-Ghazali, a leading Muslim theologian, philosopher, and jurist who lived in the 11th and 12th centuries CE.[9] He is considered one of the most influential scholars in Islamic intellectual history, and his writings have had a lasting impact on Islamic theology, law, and spirituality. Imam Ghazali's writings have had a profound impact on Islamic thought and continue to be studied and debated to this day.[27] His legacy can be seen in the many Islamic scholars, philosophers, and mystics who succeeded him.

Ghazali's most famous work, *Ihya Ulum al-Din" (Revival of Religious Sciences)*, is a comprehensive treatise on Islamic spirituality and ethics.[76] The work covers a wide range of topics, including worship, ethics, and purification of the soul. It has been widely read and studied by Muslims around the world and is considered a classic of Islamic literature.

In addition to his spiritual and ethical writings, Ghazali also wrote extensively on Islamic theology and philosophy. He is known for his critique of the philosophical views of the ancient Greek philosopher Aristotle, which had been adopted by many Muslim philosophers of his time.[27] Ghazali argued that the methods of philosophy were insufficient for understanding the nature of God and that the only way to gain true knowledge of God was through mystical experience and intuition.

Al-Ghazali denounces many Arabic philosophers and the practice of falsafa in his work *The Incoherence of the Philosophers (Tahafut al-Falasifa)*.[5] In this work, he claims that the Arabic philosophers subscribe to Greek philosophical conceptions even when they are at odds with Islam. He seeks to reconcile the views of the kalam theologians with the falasifa, and in doing so, he introduces some of Ibn Sina's philosophy into Muslim theology. Al-Ghazali's approach was accepted by many later Muslim theologians and had an influence on Latin medieval thinking.

In his works, Al-Ghazali affirmed that many of the theories of Arabic philosophers regarding the soul did not contradict religion. Those theories of the soul that were at odds

with Islam according to Al-Ghazali were any beliefs that the intellect is the main guide to gaining knowledge about the nature of reality. He believed that while the intellect is important, Islam also emphasizes other facets of the soul such as the heart. Focusing on these and the revelation are necessary to have proper and accurate knowledge as opposed to looking at everything from a purely rational perspective.

The Intellect

It appears that most Arabic philosophers believe that intellection is an endeavor that is carried out by the soul.[5] And it is the human rational faculty which enables an individual to engage in intellection. Book III of *De Anima* provides the foundation for theories of intellect propounded by many Arabic philosophers. There are four main stages in the progression of the intellect. The first is the agent intellect which is immortal and eternal. It is an immaterial substance and it is not a faculty of individual souls. It is the cause of human understanding either by making certain objects intelligible or by making possible the potential intellect. The potential (material) intellect is an innate capacity within the human soul which allows for the reception of intelligibles. The habitual (speculative) intellect is the human potential intellect after it has acquired various intelligibles and it is responsible for allowing individuals to think on command. Lastly, the acquired intellect is achieved once the habitual intellect has been perfected. This occurs when an individual has acquired all possible intelligibles.

There were several Arabic philosophers who had exception to this general framework of the intellect.[5] For instance, al-Kindi calls the habitual intellect the acquired intellect, and he calls the acquired intellect the second intellect and it is intellect when it is using knowledge it has learned previously. Ibn Rushd believed that potential intellect is like the agent intellect in that it is one for all humans and a separate substance, and he adds a fifth intellect called the "passive intellect" (imagination/cogitative faculty) which he associates with a bodily faculty. Ibn Sina further breaks down habitual intellect into two stages. This includes the habitual intellect which is related to the acquisition of primary intelligibles (principle of non-contradiction, principle of identity, principle of excluded middle), and the actual intellect which is responsible for the acquisition of secondary intelligibles that can be reasoned from first principles. Moreover, Ibn Sina's concpetion of the acquired intellect is similar to that of Al Kindi.

The actualization of the potential intellect into reality is said not to occur without first some external cause.[6] It is the Active Intelligence (*al-'aql al-fa'al*) also known as the Holy Spirit (*al-ruh al-qudus*) which serves this purpose. The Active Intelligence is, ultimately, a manifestation of God. This Active Intelligence is continuously acts in order to transform the potential human intellect into the actual human intellect, by conferring knowledge of universal forms and eternal truths, and without it, the human intellect could never be properly actualized.

A Contemporary Synthesis

In *Prolegomena to the Metaphysics of Islam*, Syed Muhammad Naquib Al-Attas outlines a more thorough view of the metaphysics of the soul by consolidating the works of Al-Ghazali, Ibn Sina, and verses of the Qur'an. For the remainder of this section, I will be outlining the conception that is provided by Al-Attas in his book. As such, this perspective is limited in its scope, and certainly does not cover all of the views that are present in Islam. However, it does offer a consolidated view of several prolific Islamic thinkers. In this view, the heart (*qalb*), self (*nafs*, also sometimes translated directly to soul), spirit (*ruh*), and intellect (*aql*) are all facets of the soul.[6] These facets have both physical, material bases as well as spiritual, immaterial aspects of them. It is the immaterial aspects of these faculties which refer to what we consider to be the basis for the essence of humans.

The physical manifestation of the heart refers to the same one as the West, that is, the organ that pumps blood throughout parts of the body.[6] The heart is said to originate the spirit and carry it throughout the body. Without the spirit, the human's body can no longer function and results in death. The intellect is located to parts of the brain and allows for what we normally consider as intellect in the West. As such, it allows us to abstract concepts and create mental representations of physical phenomena, evaluate truths about the world, and carry out reasoning. And the self denotes the individual being and is akin to modern conceptions of psyche or ego. It is often said to be located in the chest or the

stomach. These four facets comprise a unified soul, which is what give humans their reality.

The soul is said to have different states (*ahwal*), and each of the facets arise when the soul is working towards different ends.[6] Thus, when the human being is engaging in reasoning, the soul is called the intellect. When the body is being governed it is called the self. The heart is the state of the soul when it is emoting or receiving illumination relevant to spiritual connections with God. The *ruh* is said to be the source of life. Thus, each of these four faculties correspond to specific facets of a single, unified soul. The soul is constantly manifesting itself through these different facets and provides the basis for what it means to be human.

What is interesting about the soul in Islam is that it is not limited to humans.[6] Rather, there are different powers of the soul (*quwa*) depending on the form of the body that the soul inhabits. However, in relation to different bodies, the composition of the soul may not necessarily be the same. In plants, the soul has powers of nutrition (*al-ghadhiyyah*), growth (*al-namiyyah*), and reproduction (*al-muwallidah*). These powers, while they exist in plants, also exist in animals and humans.

In animals, there also exist the powers of action (*al-muharrikah*) and perception (*al-mudrikah*). These powers can be further broken down.[6] The power of action functions in two ways. The first being the arouser of action (*al-bai'ithah 'ala 'l-fi'l* where it is able to direct and create the intention of movement towards the things it desires and away from

the things that it avoids. The second is as the active (*fa'ilah*) function which actually begins and controls the movement of the individual.

Perception can be broken down in to both the external senses as well as the internal senses.[6] These external senses are the common modes of sense perception: touch, smell, taste, sight, and hearing. The internal senses on the other hand include the common sense, the representative faculty, estimative faculty, retentive faculty, and the imaginative faculty. The common sense (*al-hiss al-mushtarak*) obtains data from external sense perception. It is an innate ability of human beings to distinguish between forms and make perception possible. The representative faculty (*al-khayaliyyah*) takes the images and forms obtained from the common sense and stores this information to maintain images, meanings, and representations which are dispensed to the estimative faculty (*al-wahmiyyah*). The estimative faculty is involved in judging between good and bad and forming opinions. The retentive faculty (*al-hafizah*) is involved in preserving meanings and judgements from the estimative faculty similar to how the representative faculty retains sense data from the common sense. Lastly, is the imaginative faculty (*al-mutakhayyilah*) which is able to perceive forms, distinguish them, and also manipulate them in order for the soul to understand their meanings and relate them to other forms or images. It is through the imaginative faculty and its development that one is able to understand and work with abstract ideas which go beyond sense perception. This imaginative faculty is considered to be unique to humans and is

what allows us to perform the rational thinking, deliberation, and understanding that we do. These faculties of the soul are not distinct, rather, it is the soul manifesting itself in different ways depending on the organ in which the function is being localized or the time in which the function is being carried out. They can be thought of as different modes of the same, single soul.

The rational faculty of the soul can also be broken down into two different powers.[6] This includes the active (*'amilah*) and the cognitive (*'alimah*). The active faculty of the intellect refers to the ability of the mind to produce new ideas and concepts. The active faculty is often associated with the power of the soul to create and innovate, and it is seen as a key element in human agency. The cognitive faculty of the intellect, on the other hand, refers to the ability of the mind to understand and comprehend existing knowledge. It is the part of the intellect that is responsible for reasoning, judgment, and critical thinking. The cognitive faculty is often associated with the power of the soul to understand and reflect on the world around us, and it is seen as a key element in human intelligence. When the intellect is working with different facets of the soul, it will carry out different functions. For instance, if the intellect is working with the perceptive power of the animal faculty of the soul, it will give rise to human skills and arts. In conjunction with the imaginative faculty, the intellect allows humans to produce logical thought and deduce conclusions from a given set of premises. Both the active and cognitive powers are seen as complementary and

necessary for the full development of human intelligence and agency.

4

The Verdict

The purpose of this research project was to understand if Strong AI is realizable from an Islamic perspective. However, the premise of the research problem is itself problematic. There is indeed no *one* Islamic perspective. Rather, as we have seen from the elucidation of only a very small number of Islamic scholars since the inception of the religion,

there is certainly great variety and diversity in thought as it relates to ideas on the intellect and metaphysics of the soul. Very little work has been done as it relates to reconciling the current and future states of computation and technology with the Islamic understanding of the human mind, intellect, and soul. And while this is a first effort attempt at doing so, it certainly does not encompass the full range of beliefs on the matter. That is, this exploration is an ongoing project that requires much more research and investigation which outside of the scope of an undergraduate thesis. With that said, in this project, we have outlined the beliefs of several prominent Islamic thinkers to give a first pass at explaining what a few of the Islamic perspectives are on the matter of Strong AI and whether it is achievable.

As we have outlined earlier Strong AI is a much more rigorous thesis compared to Weak AI which posits that we can indeed reproduce human mental states and thought with computation and machines. This means that all of the abilities that humans possess can be captured and replicated in a different substrate via computation. To get to the root of the issue, we must see what if human mental states are so unique that they cannot replicated in this manner.

Let us first begin by summarizing the views of the falasifa. Out of all of the Arabic philosophers we have looked at, they all have a conception of the soul. Al-Kindi, al-Razi, and Ibn Sina all have dualistic conceptions of the soul. Ibn Sina, with his "flying man" is one of the

first to seek to give an account of self-awareness and offers a thought experiment in his effort to support his claims. While the "flying man" thought experiment does indeed have its objections, it conveys Ibn Sina's belief that the soul can indeed exist even when its physical foundation is stripped. They believe the soul to be an immaterial substance, and the intellect is due, at least in part, from the makeup of the soul. From the falsafa thinkers that we have examined, human mental states and intellection do indeed require a soul to be carried out. As such, for Strong AI to be realizable, it must be the case that the soul can be instantiated in something like a machine.

Of the philosophers that we have examined, it is only Ibn Sina who really outlines a procedure in which the soul becomes instantiated into an individual human form. It is the Active intellect which emanates an individual human soul upon the formation of a suitable body. The body cannot become living without the soul, but the soul cannot become individuated without first having a suitable human form. Ibn Sina posits that this suitable human form arises when a fetus develops inside the mother's womb. From this position, it is clear that Ibn Sina would reject the idea that machines can possess human mental states and a soul for they do not possess the proper form to be receptive of a soul. And without a soul, it appears that the human intellection required to uphold the Strong AI thesis is untenable.

Earlier, we outlined the general framework of the intellect that is propounded by many

Arabic philosophers with the transitions from the Active intellect → Potential intellect → Habitual intellect → Acquired intellect. However, with the current research we have carried out, it is not entirely clear from the views of Al-Kindi and Al-Razi whether this progression of intellection can occur in machines or not. To Ibn Sina, it certainly cannot. Although they posit the soul to be an immaterial substance, that does not rule out the fact that a machine could indeed receive this immaterial entity. More research must be done into the views of these and other philosophers to gauge whether they believed that the human body is the only form which is receptive of the soul and the intellect.

In his, *Incoherence of the Philosophers (Tahafut al-Falasifa)*, Al-Ghazali writes on the topic of resurrection. More specifically, he objects to those philosophers who refuse the idea of resurrection.[46] Al-Ghazali believes that even after the body decomposes and the soul is separated from the body, God creates a new body for the soul.

> Yes, in addition, [this] indicates the revival and resurrection thereafter-namely, the resurrection of the body. This is possible by returning [the soul] to the body, whatever body this might be, whether [composed] of the matter of the first body [or from that] of another or from matter whose creation commences anew.[46]

Al-Ghazali goes on to say that the new body that is created does not necessarily need to be made of the same exact form as the initial body.[46] As such, from Al-Ghazali's perspective, it is not immediately clear that Strong AI is impossible. If a soul can be received by a form which is not the same as a human body, it may well be the case, and it is within God's

power to make something like a machine also possess a soul.

On another note, when looking at the perspective of kalam theologians, it certainly seems that Strong AI is possible. Whether it is probable is another matter. From the kalam theologian perspective, many of whom were atomists, it is certainly allowed for the property of life and speech to inhere in non-human entities such as birds or stones.[66] As such, there is no reason to rule out that the same properties of life can inhere into something like a machine.

One of the important consequences regarding whether Strong AI is realizable is how would we treat these machines that have human mental states in terms of ethical and legal rights. However, one important thing to note is that just because the property of life is inhered into a certain entity, it does not necessarily mean that the entity would be held morally responsible or be entitled certain rights to them. Al-Ghazali held that God is free to bring suffering to any animal he wills even if they have not done anything wrong.[83] And he chooses to hold humans accountable but not birds even though both may have the property of life. God is not obligated in any way for this to be the case, and it is not necessarily *because* humans have reason and intellect. Rather it is because God simply chooses to do so.

It follows that even if Strong AI is possible by the property of life inhering into a machine, it is not necessarily the case that they are entitled to the same rights and responsibilities as humans. There is nothing obligatory with respect to God.[83] As such, there is no

reason that God would treat Strong AI on the same station as humans. It could certainly be the case, but it is not required.

Let us take a moment to return to the case studies provided earlier. This included object recognition algorithms, generative art algorithms, and large language models. One may contend that these algorithms already possess the abilities that many Islamic scholars would argue require an intellect or a soul. For instance, object recognition algorithms are able to distinguish between the forms of different objects and so are able to discern between intelligibles in a certain way. Generative art algorithms are able to possess new creative works which in a sense, replicate the imaginative faculty that humans are said to uniquely posses by many. So, how do we reconcile the fact that there already exist artificial intelligence technologies that already are able to do actions that some scholars would argue require intellective powers that are uniquely human.

One thing to note about all of these algorithms is that, at some point in the training pipeline, supervised learning algorithms are being put to use. That is, labelled datasets are used to train algorithms in order to correctly classify data and predict outcomes correctly. It follows that these labelled datasets represent the true understanding of the data, and these labels were made by humans using their powers of intellection. And in a sense, the labelled dataset is simply an abstraction of the human intellect which is what is used to train these models. Thus, they all required the intellective powers of humans and would

not have been possible without it. Many Arabic scholars would likely argue that these initial powers of intellection that are carried out by humans are what demonstrate true understanding and they are what require a soul and reception of knowledge from the intellect.[5] In this way, the algorithms presented in these case studies do not demonstrate true understanding and would not be able to function had they not been given data that was parsed using the human intellect while humans do not require this initial labelling but are able to discern patterns on their own.

We can see that there are many different views within Islam as it relates to human intelligence, rationality, and metaphysics. This results in different answers to the Strong AI thesis. While some thinkers, such as Ibn Sina would likely argue that Strong AI is not possible, there are others such as Al-Ghazali and the kalam theologians, whose conceptions could allow for some form of Strong AI, in the sense that machines can have human mental states and abilities. However, whether these Strong AI systems are afforded the same rights, responsibilities, and powers as humans is unclear. This project is a first attempt at integrating the beliefs of Islam into our modern debate of artificial intelligence technologies. We sought to take a broad overview of Islamic thought, and survey different beliefs as it relates to Strong AI. However, this study only scratches the surface of the range of beliefs held in Islam and their relation to Strong AI. In future studies, we seek to delve deeper into the views and conceptions of different thinkers to get a better understanding of machine think-

ing and consciousness. And we hope that this endeavor opens up a new area of inquiry examining the intersection of religion, computation, and artificial intelligence.

References

[1] Adamson, P. (2006). *Al-Kindi (Great Medieval Thinkers)*. Oxford University Press.

[2] Adamson, P. (2016). *Philosophy in the Islamic World*. Oxford University Press.

[3] Adamson, P. (2020). al-kindi. *The Stanford Encyclopedia of Philosophy*.

[4] Adamson, P. (2021). Abu bakr al-razi. *The Stanford Encyclopedia of Philosophy*.

[5] Adamson, P. & Taylor, R. C. (2005). *The Cambridge Companion to Arabic Philosophy*. Cambridge University Press.

[6] al Attas, S. M. N. (2013). *Prolegomena to the Metaphysics of Islam*. Kuala Lumpur, Malaysia: CreateSpace Independent Publishing Platform.

[7] Anderson, J. & Rainie, L. (2018). Artificial intellignece and the future of humans. *Pew Research Center*.

[8] Barak, B. (2022). *Introduction to Theoretical Computer Science*. Harvard University. https://files.boazbarak.org/introtcs/lnotes_book.pdf.

[9] Bowering, G., Crone, P., & Mirza, M. (2013). *The Princeton Encyclopedia of Islamic Political Thought*. Princeton University Press.

[10] Bringsjord, S. & Govindarajulu, N. S. (2022). Artificial intelligence. *The Stanford Encyclopedia of Philosophy*.

[11] Brown, H. (2023). What exactly is chatgpt? *CBS*.

[12] Campanini, M. (2012). The mu'tazila in islamic history and thought. *Religion Compass*.

[13] Campbell, C. C. (2023). Catch the next big tech ai wave: Startups flock to openai's chatgpt and dall-e. *Forbes*.

[14] Chalmers, D. (1995a). Facing up to the problem of consciousness. *Journal of Consciousness Studies*.

[15] Chalmers, D. J. (1995b). Minds, machines, and mathematics: A review of shadows of the mind by roger penrose. *PSYCHE*.

[16] Cole, D. (2020). The chinese room argument. *The Stanford Encyclopedia of Philosophy*.

[17] Cornford, F. M. (1997). *Timaeus and Critias (Oxford World's Classics)*. Plato's Cosmology: The Timaeus of Plato.

[18] D'Ancona, C. (2022). Greek sources in arabic and islamic philosophy. *The Stanford Encyclopedia of Philosophy*.

[19] Deshpande, A. & Wettig, A. (2023). How does chatgpt work? an overview of large language models. *Princeton Institute for Computational Sciences and Engineering (PICSciE)*.

[20] Fakhry, M. (2004). *A History of Islamic Philosophy*. Columbia University Press.

[21] French, R. (2000). The turing test: the first 50 years. *Trends in Cognitiive Sciences*.

[22] Galeon, D. (2017). Ray kurzweil: Ai will not displace humans, it's going to enhance us. *Futurism*.

[23] Gallop, D. (2009). *Phaedo*. Oxford University Press.

[24] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[25] Gordon, R. (2022). 3 questions: How ai image generators work. *MIT CSAIL*.

[26] Graham, G. (2023). Behaviorism. *The Stanford Encyclopedia of Philosophy*.

[27] Griffel, F. (2020). al-ghazali. *The Stanford Encyclopedia of Philosophy*.

[28] Gutas, D. (2016). Ibn sina [avicenna]. *The Stanford Encyclopedia of Philosophy*.

[29] Hackett, C. & McClendon, D. (2017). Christians remain world's largest religious group, but they are declining in europe. *Pew Research Center*.

[30] Hamlyn, D. W. (1984). *Metaphysics*. Cambridge University Press.

[31] Hamlyn, D. W. (1993). *De Anima*. Oxford University Press.

[32] Hardesty, L. (2017). Explained: Nueral networks. *MIT News*.

[33] Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*.

[34] Hasan, U. & Osama, A. (2016). *Islam & Science*. Muslim World Science Initiative.

[35] Hatfield, G. (2018). René descartes. *The Stanford Encyclopedia of Philosophy*.

[36] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *arXiv*.

[37] Honderich, T. (2005). *The Oxford Companion to Philosophy*. Oxford University Press.

[38] Johnson, S. (2022). A.i. is mastering language. should we trust what it says. *The New York Times*.

[39] Kaukua, J. (2015). *Self-Awareness in Islamic Philosophy*. Cambridge University Press.

[40] Lee, J. S. (2021). Transformers: A primer. *Columbia University*.

[41] Lee, S. (2020). Occasionalism. *The Stanford Encyclopedia of Philosophy*.

[42] Levin, J. (2021). Functionalism. *The Stanford Encyclopedia of Philosophy*.

[43] Lewis, D. (1983). Mad pain and martian pain. *Philosophical Papers*.

[44] Lucas, J. R. (1961). Minds, machines, and godel. *Philosophy*.

[45] Mandik, P. (2010). *Key Terms in Philosophy of Mind*. Continuum.

[46] Marmura, M. E. (2000). *Al-Ghazali's The Incoherence of the Philosophers*. Brigham Young University Press.

[47] McGinnis, J. (2022). Arabic and islamic natural philosophy and natural science. *The Stanford Encyclopedia of Philosophy*.

[48] Megill, J. (2019). The lucas-penrose argument about godel's theorem. *Internet Encyclopedia of Philosophy*.

[49] Mishra, A. (2017). Deep-learning networks rival human vision. *Scientific American*.

[50] Mozur, P. (2017). Google's alphago defeats chinese go master in win for a.i. *The New York Times*.

[51] Nagel, E., Newman, J., & Hofstadter, D. R. (2008). *Godel's Proof*. NYU Press.

[52] Nasr, S. H. (2006). *Islamic Philosophy from its Origin to the Present*. State University of New York Press.

[53] Nussbaum, M. C. & Rorty, A. O. (1992). *Essays on Aristotle's De Anima*. Oxford University Press.

[54] Oppy, G. & Dowe, D. (2021). The turing test. *The Stanford Encyclopedia of Philosophy*.

[55] Penrose, R. (1996). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

[56] Phuong, M. & Hutter, M. (2022). Formal algorithms for transformers. *arXiv*.

[57] Putnam, H. (2010). Brains and behavior. *Philosophical Papers*.

[58] Rachlin, H. (2010). Pain and behavior. *Behavioral and Brain Sciences*.

[59] Richardson, K. (2020). Causation in arabic and islamic thought. *The Stanford Encyclopedia of Philosophy*.

[60] Robinson, H. (2020). Dualism. *The Stanford Encyclopedia of Philosophy*.

[61] Roose, K. (2022). An a.i.-generated picture won an art prize. artists aren't happy. *The New York Times*.

[62] Russell, S. & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Pearson.

[63] Schmidt, R. M. (2019). Recurrent neural networks (rnns): A gentle introduction and overview. *arXiv*.

[64] Searle, J. (1984). *Minds, Brains and Science*. Harvard University Press.

[65] Sedley, D. (2003). *The Cambridge Companion to Greek and Roman Philosophy*. Cambridge University Press.

[66] Shihaden, A. (2012). Classical ash'ari anthropology: Body, life and spirit. *The Muslim World*.

[67] Smart, J. J. (2022). The mind/brain identity theory. *The Stanford Encyclopedia of Philosophy*.

[68] Staudemeyer, R. C. & Morris, E. R. (2019). Understanding lstm – a tutorial into long short-term memory recurrent neural networks. *arXiv*.

[69] Stoljar, D. (2022). Physicalism. *The Stanford Encyclopedia of Philosophy*.

[70] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with nueral networks. *arXiv*.

[71] Turing, A. (1950). Computing machinery and intelligence. *Mind*.

[72] Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*.

[73] Tzezana, R. (2017). Singularity: Explain it to me like i'm 5-years-old. *Futurism*.

[74] Vaswani, A. et al. (2017). Attention is all you need. *arXiv*.

[75] Waterfield, R. (2008). *Timaeus and Critias (Oxford World's Classics)*. Oxford University Press.

[76] Watt, W. M. (2023). al-ghazali. *Encyclopedia Britannica*.

[77] Wikimedia (2010). Multilayerneuralnetwork.

[78] Wikimedia (2015). Convolutional neural network.

[79] Wikimedia (2022a). 1960's art of cow getting abducted by ufo in midwest.

[80] Wikimedia (2022b). Midjourney artifiical intelligence art.

[81] Wikimedia: Mitchell, Machine Learning (2012). Perceptron.

[82] Wikimedia: Yuening Jia (2019). Transformer model architecture.

[83] Yaqub, A. M. (2013). *Al-Ghazali's Moderation in Belief*. The University of Chicago Press.