# An Intellectual History of the Integration of Gödel's Incompleteness Theorems into the Artificial Intelligence Debate and its Theological Implications

Kailee Madden

Senior Thesis
History Honors Program

## Table of Contents

**Page**

**An Intellectual History of the Integration of Gödel's Incompleteness Theorems into the Artificial Intelligence Debate and its Theological Implications**

*"We must know. We shall know!" – David Hilbert*

Mathematical knowledge has a contradictory place in history. Because of its focus on formal proof, it is both one of the least problematic and most problematic areas of human knowledge. Formal mathematical proofs have to start somewhere, that is, with basic assumptions known as axioms. These assumptions are generally conclusions from other proofs. Eventually, however, one reaches a point where there are no longer any previous conclusions to use. Yet still, there must be some given truth, some assumption, in order to have any sort of system from which a proof can be derived.[1] This requires mathematical intuition. But sometimes, even the best of intuitions falter. Such was the case for David Hilbert, a highly influential German mathematician, a founder of proof theory and mathematical logic, and a great proponent of formalism – a program for grounding mathematics on a solid and complete logical foundation.

In 1920, Hilbert proposed that all mathematics follow from a finite system of axioms, or truths, and he called for a formalization of all mathematics in axiomatic form along with a proof that this axiomatic form was consistent using only methods denoted by Hilbert as finitary. Finitary proofs are those that can be written on a sufficiently large piece of paper, that is, the argument can be translated into a finite set of symbolic propositions starting from a finite set of axioms.[2] This is known as Hilbert's Program, and it permeated the field of mathematics, with many mathematicians from the period working on problems that he

---

[1] Goldstein, Rebecca. *Incompleteness: The Proof and Paradox of Kurt Gödel*. New York: W.W. Norton, 2006. Pgs. 121-123.

[2] Zach, Richard. "Hilbert's Program." Stanford Encyclopedia of Philosophy. January 06, 2015. Accessed November 28, 2018.

proposed. On October 7, 1930, in an address before the Society of German Scientists and Physicians, Hilbert said, "the true reason why [no one] has succeeded in finding an unsolvable problem, is in my opinion, that there is no unsolvable problem. In contrast to the foolish Ignoramibus, our credo avers: We must know. We shall know!" [3] He had no idea what mathematician Kurt Gödel already knew and had presented the previous day at a roundtable discussion: Hilbert's Program was impossible as stated.

Born in Brünn, Austria-Hungary, now Brno, Czech Republic, Kurt Gödel was nicknamed "Mr. Why" for his insatiable curiosity.[4] Gödel went on to become a logician, mathematician, and philosopher. He even dabbled in physics. At 25 years old, in 1931, Gödel published his two incompleteness theorems, the second of which completely disrupted Hilbert's Program. The first incompleteness theorem argued for the existence of unprovable statements within consistent formal systems and the second incompleteness theorem expanded upon the first by arguing that one of these unprovable statements was the statement of the system's own consistency. From the second incompleteness theorem comes the implication that consistency proofs could not be carried out with solely finitistic reasoning as Hilbert desired. These theorems drastically changed the field of formal logic.[5]

Now, in the field of formal logic, Gödel's incompleteness theorems are ubiquitous. After all, they were a critical turning point in mathematical logic as a way to represent paradoxes through numbers and prove their existence in various constructed mathematical systems. However, in the past seventy years this familiarity has expanded well beyond formal logic and mathematics. In fact, people have begun to apply these incompleteness

---

[3] Dawson, John W. *Logical Dilemmas: The Life and Work of Kurt Gödel*. Wellesley, MA: K Peters, 1997. Pgs. 68-71.
[4] Ibid. Pg. 1.
[5] Franzén, Torkel. Gödel's Theorem: An Incomplete Guide to Its Use and Abuse. Wellesley, MA: K Peters, 2005. Pg. 39.

theorems to a variety of industries and questions, such as psychology, physics, theology, and artificial intelligence. Many applications of the theorems are misuses and derive false conclusions, such the claim that the U.S. Constitution must be incomplete. In this example, the Constitution is a formal system, so the Founding Fathers must have had to choose between consistency and completeness in the Constitution. Apparently, they chose consistency, thereby necessitating the integration of the Judicial Branch to close the gap of incompleteness.[6] This is a clear example of a misapplication. The Constitution does not meet the requirements of a formal system, and these are mathematical theorems, they require a certain amount of elementary arithmetic to be possible within the formal system, that is, a sufficiently strong arithmetic.[7] However, one area in which the theorems have been aptly applied is in the debate regarding the feasibility of strong artificial intelligence (AI),[8] that is, the questions of whether a computational machine can model human intelligence in all ways, including capturing the elusive idea of consciousness.

One major figure in these debates about the feasibility of AI, who leverages Gödel's incompleteness theorems, is Sir Roger Penrose. An English mathematical physicist, mathematician, and philosopher of science, Penrose is renowned for his contributions in mathematical physics, especially his work in general relativity and cosmology. He has also

---

[6] Franzén, Torkel. Pgs. 77-78.

[7] A formal theory of arithmetic T is considered sufficiently strong if it case-by-case captures all decidable numerical properties. A decidable property of numbers is one for which there exists a mechanical algorithm for deciding whether given a natural number *n*, if *n* has that property or not. Smith, Peter. *An Introduction to Gödel's Theorem*s. Cambridge: Cambridge University Press, 2014. Pg. 30.

[8] One note for the rest of the paper is that the artificial intelligence referred to throughout this paper is more correctly classified as strong AI, that is, a machine which has intellectual capability functionally equal to or greater than a human in a way that it can truly model the mind and encapsulate the elusive idea that we denote consciousness. However, I will reference it as AI or strong AI, but it means the same – a machine that models the human mind. Another important definition is that of computationalism, the idea that intelligent behavior is causally explained by computations performed by the brain, hence meaning that the mind could be modeled by a machine. For simplicity, I will generally just refer to the feasibility of strong AI, as there are slight nuanced differences between the two.

won many awards, including one that he shared with Stephen Hawking for their singularity theorems, a set of results in general relativity that explain when gravitation produces a location in space-time where the gravitational field of a celestial body becomes infinite such that it is independent of the coordinate system. While it is the minority opinion, Penrose holds the stance that strong AI is not possible because Gödel's incompleteness theorems mean that there must be something non-algorithmic in the mind, and therefore a machine cannot represent the human mind.

In this senior thesis, I investigate the way in which Gödel's incompleteness theorems were integrated into the debate surrounding strong artificial intelligence, as well as how this debate pertains to theological questions. In order to accomplish these tasks, I first explain the theorems and the main AI arguments that have leveraged the theorems to make their points. Once this context has been established, I move onto the historical background in which the theorems were developed and the ways in which, over time, the theorems permeated the field of mathematical logic as well as eventually the fields of computing and AI, specifically focusing on physicist Roger Penrose's argument against the feasibility of strong AI. Finally, I investigate the way in which Gödel's theological and philosophical beliefs may have influenced his work, and how the emergence of strong AI could have theological implications. Again, this will be done from a historical, rather than philosophical, perspective. My aim is not to add my voice and opinion on the question of whether AI is possible by deriving an answer from Gödel's theorems, but rather, to illuminate the way in which these ideas emerged, developed in the mathematical community, and spread to situations beyond mathematical logic, as well as tease out a thread of relationship between the theorems and theology.

Before embarking on this journey, it seems apt to establish the fact that such an intellectual history is worth writing. Even though many mathematicians were slow to comprehend the full significance of Gödel's theorems, once this significance was grasped, the theorems greatly influenced and changed the field of mathematical logic. Beyond just mathematical logic though, Gödel's theorems have remained astoundingly relevant and, in fact, they are overly applied and greatly misused because of their popularity. For example, the two incompleteness theorems have been used to supposedly both prove and disprove the existence of God. For instance, Torkel Franzén who argues:

> But it is most likely safe to say that no mathematical theorem has aroused as much interest among nonmathematicians as Gödel's incompleteness theorem, which appeared in 1931. The popular impact that this theorem has had in the last few decades can be seen on the Internet, where there are thousands of discussion groups dedicated to every topic under the sun. In any such group, it seems, somebody will sooner or later invoke Gödel's incompleteness theorem.[9]

Although the vast majority of these supposed applications of Gödel's theorems are incorrect, they demonstrate the way that people have been fascinated by the theorems. In addition, the incompleteness theorems do directly apply to the debates regarding the feasibility of strong AI.[10]

The way in which these theorems apply to AI debates is also highly relevant. Over the past ten years, artificial intelligence has become one of the top buzzwords and most talked

[9] Franzén, Torkel. Pg. 1.
[10] Ibid. Pg. 86.

about technological topics. Strong AI has also been heavily featured in films, such as *Ex Machina* and *Her*. As computational power has grown, new algorithms and methods have been developed, such as artificial neural networks – a framework for many machine learning algorithms to work together and process complex data inputs – , and computer scientists and engineers have reached a point where they do not understand the output of some of the tasks they delegate to machines. Thus, it is apt to reflect on the limitations of computational power in order to reevaluate the trust that is put in these machines.[11] It is also worth reflecting on the history of the strong AI debates, as that can illuminate the way in which we have reached our modern relationship with the strong AI idea.

Beyond the established value of studying Gödel's incompleteness theorems and their relationship throughout history to feasibility of AI debates, it is important to differentiate my own work from that which has already been done. Rather than adding my personal opinion derived from an interpretation of Gödel's theorems to the plethora of personal opinions already in existence, I aim to illuminate the way in which the theorems and the major AI arguments, specifically Penrose's, developed and became integrated into the larger conversation of AI and mathematical logic. By looking at these arguments from a historical perspective, I hope to provide focused insight into their growth and change, which then can be used to better understand the arguments themselves and their applicability in our current day. While there are histories already of the life of Gödel and his work, such as *Logical Dilemmas* by John Dawson, as well as plenty of histories regarding the developments of AI and AI debates, my work of specifically focusing on the integration of Gödel's theorems into

[11] Jongeneel, C.j.b, and H. Koppelaar. "Gödel Pro and Contra AI: Dismissal of the Case." *Engineering Applications of Artificial Intelligence*12, no. 5, 1999. Pg. 655.

the AI debate from a historical perspective is a more unique approach, particularly when linking it back to Gödel's theological leanings and theology more generally.

**Gödel's Incompleteness Theorems**

Often, Gödel's two incompleteness theorems are treated as one entity. This is due to the fact that they are regarding principles of consistency and completeness in the same formal system. The theorems, as they stand today, are as follows:

> First Incompleteness Theorem: Any consistent formal system F within which a certain amount of arithmetic can be carried out is incomplete. That is, there are statements of the language of F that cannot be proved or disproved in F.
>
> Second Incompleteness Theorem: For any consistent system F within which a certain amount of elementary arithmetic can be carried out, the consistency of F cannot be proved in F itself.

Let us break this down. A formal system is a system of axioms equipped with rules regarding how the symbols of the system can be used, which allow one to generate new theorems. In less jargon, this means that within a formal system there are a collection of assumptions or truths, a collection of symbols that follow particular rules, and new ideas that

can be created using the assumptions and rules within the system. The rules within this formal system are such that a basic amount of arithmetic is possible. Without this caveat it would be possible to apply these theorems to a multitude of fields in which they do not belong, however, they are mathematical theorems and, as such, require a certain amount of arithmetic to be possible within the formal system.[12] Another important note is that a formal system attempts to eradicate the influence of intuition, since intuition can be misleading. It attempts to break a system down into the most basic assumptions, so that it is divested of appeals to intuition, as much as possible. Without some axioms, or assumptions, it would be impossible to actually start proving anything, yet, it is important to reduce these assumptions to the most basic truths possible in order to avoid false assumptions. From these assumptions, by following the rules of the system, new theorems or statements of truth can be proven or disproven.[13]

Next, we can look at consistency and completeness. To be consistent means that there exists no statement such that the statement itself and its negation are both derivable in the system. So, a consistent formal system cannot have proven contradictions. For example, the statement that "The ball is red" and the statement's negation that "The ball is not red" cannot both be provably true.

On the other hand, to be complete means that, for every statement of the language of the system, either the statement or its negation can be proven in the system. This means if you can create a statement by following the set of rules for the system, then you must be able to prove that statement to either be true or false in order for the system to be complete. So, if

[12] Franzén, Torkel. Pgs. 83-87.
[13] Goldstein, Rebecca. Pgs. 126-129.

the statement that "The ball is red" can be created within the system, it must either be provably true or provably false.

Returning to the two incompleteness theorems, we can generate new descriptions that are more devoid of jargon.

First Incompleteness Theorem: If a system is consistent, then the system is incomplete, meaning there are statements that can be created from the assumptions and rules of the system, but cannot be proven to be true or false.

Second Incompleteness Theorem: If a system is consistent, then the statement about the system's own consistency cannot be proven to be true within the system itself.

An example of the first theorem would be the well-known Liar's Paradox.[14] An example statement would be: "This statement is false." If the statement is false, as it claims, then it is a true statement, but that creates a contradiction. On the other hand, if it is true, then that means it is false. Clearly, such a paradoxical statement cannot be proven to be true or false. Gödel discovered a way to prove the existence of these paradoxes in consistent mathematical systems. He represented statements by numbers, a method known as Gödel

[14] There are many variations of the Liar's Paradox, and similar paradoxes that have puzzled academics in philosophy and logic throughout the years. One of the first is the Epimenides paradox, from around 600 BC, which states, "All Cretans are liars." The problem with the Liar's paradox, and other similar statements, is that they seem to show that common beliefs about truth-value lead to contradictions. There have been a number of proposed resolutions of these paradoxes, such as Saul Kripke who argued that if a statement is "grounded", that is, its truth relates to some evaluable fact about the world, then it has a truth value. If the statement is "ungrounded", then it has no truth-value. Statements such as the Liar's Paradox are "ungrounded" and therefore there is no paradox since they have no truth-value. This is not the sole solution, and is also not accepted by everyone, but is an example of how some have attempted to resolve paradoxes.

coding or Gödel numbering.[15] Specifically, he used this method for the statement "this sentence is not provable", also known as the Gödel sentence "G". Gödel differentiated between provability and truth, demonstrating in his theorem that in a theory "T", "G" is true, but not provable in "T".[16]

To understand the second theorem we can consider the statement about the system's own consistency. Essentially, Gödel demonstrated that the statement "This system is consistent" cannot be proven to be true or false within the system. Thus, the second incompleteness theorem is an example of the first.

Even though Gödel first presented his theorem about the incompleteness of consistent formal systems, he quickly followed it up with the second theorem about the inability of a consistent formal system to prove its own consistency. It was this second incompleteness theorem that was most influential at the time. The second incompleteness theorem was most relevant because it meant that Hilbert's Program, the standard for mathematical logic, was incorrect as it stood at the time and had to be modified. Later, however, the first incompleteness theorem became the more popularly relevant theorem. The implications of the first incompleteness theorem are that regardless of the method, there is never a way to encapsulate definitions of numbers and operations in a set of axioms, giving a consistent

---

[15] Gödel coding represents symbols as prime numbers to the n+1. Take, for instance, the word "balloon". Each unique symbol "b", "a", etc. would be given a number to represent it: "a" = 1, "b" = 2, "l" = 3, "n"= 4, "o" = 5. Then stepping through the statement "balloon", each symbol of the statement would be represented by a prime to the power of that symbol's unique number plus one. So, "balloon" becomes $2^{2+1}$, $3^{1+1}$, $5^{3+1}$, $7^{3+1}$, $11^{5+1}$, $13^{5+1}$, $17^{4+1}$. These numbers are all multiplied together $2^3 * 3^2 * 5^4 * 7^4 * 11^6 * 13^6 * 17^5$. Together they become a very large number that can be broken apart again into this exact combination according to the Fundamental Theorem of Arithmetic, also known as the Unique Prime Factorization Theorem, which states that every integer greater than 1 is either a prime or can be represented by a product of primes such that the representation is unique. Due to this theorem, it is possible to use Gödel coding to generate a large number representative of the statement, and to break that large number back apart again into the symbols that represent the statement.

[16] Gödel never formalized this differentiation between truth and provability, though he did use and understand it. Instead, a mathematician named Alfred Tarski published a result in 1933, regarding truth and provability, known as Tarski's Indefinability Theorem.

theory "T". There will be truths of basic arithmetic that cannot be proved within "T". However, we, as humans, know some of these arithmetical truths even though they seem to be beyond the capability of logic and definitions. The question of how we can understand these truths that cannot be formally captured seems to lead to the mind-machine problem, that is, the debate about whether or not human minds can be captured by a machine. Again, it is important to draw the distinction between truth and provability here – Gödel demonstrated that there are truths that are not provable within the bounds of mathematical logic. For the debates about the feasibility of strong AI, both incompleteness theorems are highly relevant, however, the first is what initially implies that Gödel's theorems are worth integrating into the debate.

A quick clarification before further explaining Gödel's methodology for his proofs: using paradoxes, such as the Liar's Paradox, to understand the incompleteness theorems is for ease of understanding but the theorems are not exact representations of these paradoxes. Instead, they are similar ideas. Gödel's proof actually coopts the structure of self-referential paradoxes, such as the Liar's Paradox, but reshapes that structure for its own ends.[17] As previously mentioned, a formal system requires a certain amount of arithmetic to be possible, so the incompleteness theorems do not actually directly apply to paradoxes, and the Gödel sentence G refers to a statement's provability, not its truth.[18] The Gödel statement G is assumed to be true in the formal system, so it is the statement's provability that is in question, not its truth. Rather than exact representations then, paradoxes are heuristic grips for those wrestling with Gödel's theorems. They are a way of understanding what it is that the theorems mean, but this meaning is solely within a formal mathematical system. This

[17] Goldstein, Rebecca. Pg. 165
[18] Franzén, Torkel. Pgs. 84-86.

idea will be returned to in the next chapter as a way of understanding how and why the incompleteness theorems have been misapplied to other fields.

In his proof of the first incompleteness theorem, Gödel took a few major steps. First, Gödel laid out his formal system. Second, he devised a method to assign a unique number to statements in the formal system, called Gödel numbering. The method leverages the power of prime numbers to encode and decode statements, thereby enabling Gödel to use arithmetical statements that are also making metamathematical statements. Third, Gödel created a proposition that is true because it says that it is unprovable within the formal system, and the system is consistent meaning that if the proposition is unprovable then it is true. Finally, Gödel devised a way to prove that this proposition is unprovable within the formal system. To do so, Gödel derives a particular case of something called the diagonalization lemma.

> The Diagonalization Lemma: If T is a strongly primitive recursive adequate, primitive recursive axiomatized, theory and φ(z), is an open well-formed formula with one free variable, then there is some sentence γ such that $T \vdash \gamma \equiv \varphi(\gamma)$, that is, from T we can prove that γ is equivalent to φ(γ).

Essentially, the diagonalization lemma states that if T meets some particularly outlined requirements and a function phi meets its requirements, then there exists a statement gamma such that gamma is equivalent to phi of gamma. And gamma is provable in T, meaning that phi of gamma is provable in T. Applying this lemma to the first incompleteness theorem, because of the nature of Gödel numbering, gives the following:

1. For any propositional function of one variable F(x)
2. There will exist a number *n* such that the Gödel number of F(*n*) = *n*
3. So Gdl(F(*n*)) = *n*

This particular case of the diagonalization lemma is then used to prove that a Gödel statement G, "this statement is not provable", is true if and only if G is not provable.[19] As long as the formal system is consistent, then this statement is not provable, thus the system is incomplete and the first incompleteness theorem has been established.

For the second incompleteness theorem, the proof relies on the first incompleteness theorem. Let C be the statement of the formal system's consistency, T be the formal system, and G the Gödel statement constructed for the formal system according to the first theorem, meaning that it is a true but unprovable statement. G is assumed to exist and can be used in the second incompleteness theorem proof because it was already proven to exist. The first incompleteness theorem also proved that G is true but not provable within the formal system T. Now put these together to get the following statement:

$$T \vdash C \Rightarrow G$$

There are two claims within this statement. One, C is provable from T. And two, C implies G. This means if it was possible to prove C, then since C implies G, we would also be able to prove G. But from the first incompleteness theorem we know that G is not provable, so therefore C must not be provable, or else we would have a contradiction. So a consistent

[19] Goldstein, Rebecca. Pgs. 168-183.

formal system's consistency cannot be proven within that formal system, thus the second incompleteness theorem is established. [20, 21]

It is interesting to note that Gödel's theorems were actually initially much weaker and specific; however, with the help of other mathematicians they have been broadened in scope and strengthened. Specifically, Gödel referred to omega consistency in his 1931 publishing, that is, a specific type of consistency that is stronger than general consistency. Later in 1936, J. Barkley Rosser, an American logician, demonstrated that the first incompleteness theorem held true for general consistency, thereby strengthening the theorem.[22, 23] It may seem counterintuitive, but a stronger assumption means a weaker theorem and a weaker assumption means a stronger theorem, since the less you have to assume, the broader and stronger in scope the theorem is. Thus, by proving the theorem held true for general consistency, Rosser strengthened Gödel's theorem. In addition, when Gödel wrote his theorems, he argued only for their applicability in a system called P.[24] Other mathematicians have generalized this though to apply to many other axiomatic systems, and today Gödel's

---

[20] Ibid., Pgs. 183-184.

[21] These are not proofs of the theorems, but rather, just explanations of the general ideas used in the proofs. In addition, the footnotes are just brief additions to explain some of the important concepts further but do not in anyway constitute a comprehensive representation of mathematical logic necessary to understand Gödel's proofs. For the full details of the proofs, a basic build up of logic, and other relevant theorems, consult Peter Smith's *An Introduction to Gödel's Theorem*.

[22] Franzen, Torkel. Pg. 3.

[23] A repercussion Rosser's theorem is that there exist an uncountably infinite number of Gödel sentences, whereas before, using omega consistency, there were a countably infinite number. A set is countably infinite if its elements can be put in one-to-one correspondence with the set of natural numbers, and therefore you can count to any element in a finite amount of time even though counting all elements would take forever.

[24] This formalized theory P is considered a variant of a type-theoretical system PM. On the other hand, the first detailed proof of the second incompleteness theorem, since Gödel had only outlined the proof in his 1931 paper, was published in 1939 and was only applicable for the system PA, Peano-Arithmetic. It is also interesting to note that in 1936, mathematician Gerhard Gentzen proved the consistency of Peano-Arithmetic by nonfinitary methods in a theory that was not Peano-Arithmetic. So, even though he proved the consistency of the formal system, it was not contradictory to Gödel's second incompleteness theorem.

incompleteness theorems are constructed such that they apply to formal systems within which a basic amount of elementary arithmetic can be carried out and some basic rules of arithmetic can be proved.[25] So Gödel's theorems were generalized, and then they were applied to a variety of fields.

**AI Applications and Arguments**

Philosopher of science Ernest Nagel and mathematician James R. Newman were some of the first to make Gödel's incompleteness results accessible to a wider audience by representing the main results of the theorems, as well as the mathematical work and philosophies leading up to its discovery. Nagel's work primarily concerned the philosophy of mathematics, and he was also a logical positivist, thus his book regarding Gödel's theorems helped to propel the myth that Gödel's work supported logical positivism. Logical positivism was the belief that only statements that are verifiable by empirical observation are meaningful, whereas Platonism asserts that it is the ideas and theories, not their representations through observation, that are meaningful. Gödel, despite attending group meetings in Vienna primarily run by logical positivists, was a strong Platonist. On the other hand, Newman was a mathematician, a historian of mathematics, and had previously been a practicing attorney. He is most famous for his invention of the word "googol" to represent a very large but finite number, $10^{100}$.

[25] Franzén, Torkel. Pg. 3.

In 1958, when Nagel and Newman published their book *Gödel's Proof*, Gödel was not nearly as widely known as he is today, at least not in quite the variety of academic fields. Instead, his incompleteness theorems were primarily relevant within the field of mathematical logic. This was a large part of the motivation behind *Gödel's Proof*. Recognizing the broad philosophical impact and cultural significance of the conclusions established by Gödel's incompleteness theorems, Nagel and Newman sought to "make the substance of Gödel's finding and the general character of his proof accessible to the nonspecialist."[26] Even within the philosophical and mathematical logic communities though, the incompleteness theorems were often misunderstood, particularly in the immediate years following their publication. For example, Austrian-British philosopher Ludwig Wittgenstein, whose work was primarily in logic, philosophy of mathematics, and philosophy of mind, wrote a letter in the 1960s with remarks indicating that he thought Gödel had found an inconsistency in arithmetic, meaning 2+2 =4.0001 instead of 2+2=4, rather than incompleteness within consistent formal systems.[27] Since the theorems were generally only discussed in relation to mathematical logic and the foundations of mathematics, Nagel and Newman dismissed the theoretical possibility of AI with just a few sentences. However, this was during a time when mathematicians aimed to distance themselves from their applied counterparts in the field of computing, thus when this distance began to shorten, the relevance of Gödel in relation to AI became more evident.[28] Even in the 1950s, the distance was already shortening.

Before Nagel and Newman published their book, Alan Turing had made comments that Gödel's proof was irrelevant to the question of AI. Later, in the 1960s, Gödel seemed to

---

[26] Nagel, Ernest, and James R. Newman. *Gödel's Proof*. N.Y.: New York University Press, 1958. Pg. 4.
[27] Dawson, John W. Pg. 77.
[28] Hodges, Andrew. "In Retrospect: Gödels Proof." *Nature* 454, no. 7206. 2008. Pg. 829.

dispute this claim.[29] The debate has yet to be resolved. Currently, most computer scientists tend to believe that Gödel's incompleteness theorems do not hinder the possibility of AI. Yet, though small, there is still a staunch opposition that claims otherwise. For instance, philosopher J.R. Lucas believes that Gödel's first incompleteness theorem proves that minds cannot be explained as modern machines.

Though the implications of Gödel's incompleteness theorems to the field of computability and artificial intelligence had been brought up before, it was not until British philosopher John R. Lucas that the debate truly began with vigor. Lucas has written on a variety of philosophical topics, such as free will, philosophy of mind, business ethics, and philosophy of religion, but is primarily known for his paper "Minds, Machines, Gödel", published in 1961, in which he asserts that a machine could not represent a human mathematician.

According to Lucas, "Gödel's theorem must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system."[30] More simply put, a machine is essentially a physical form for a formal system. It operates according to specifically defined rules and performs a set of operations. Since it is a concrete example of a formal system, Gödel's incompleteness theorems apply. Specifically, the first incompleteness theorem is relevant. This means that any consistent machine or computer will have a formula that is true but that the machine is unable to prove. An outside machine could be programmed to prove this formula or statement, however, in this new machine there would be a new formula that could not be proved. And so, there becomes an endless cycle. Lucas then differentiates this from minds, writing, "for every machine there is

[29] Ibid.
[30] Lucas, J. "Minds, Machines and Gödel." Philosophy 36, 1961. Pg. 113.

a truth which it cannot produce as being true, but which a mind can."[31] Essentially, Lucas argues that human minds are not like formal systems, because they have consciousness, which allows a mind to make a meta-argument about itself and circumvent the issue of incompleteness. Thus, there will always be something that a machine cannot prove that a mind can. In addition, even though taking all the machines in the world and linking them together would likely surpass the intellectual capacity of a human mind, Lucas clarifies that this group of machines cannot model the mind. It is about machine and mind not being the same: the mind is not simply a computational machine. Lucas published this paper in 1961, asserting that most mathematical logicians agreed, but were reluctant to put forth their views without a fully formulated argument against mechanism, that is, the idea that machines can simulate human minds. Lucas constructed his paper to provide this complete argument.

Like most philosophical papers, there were a number of objections raised against Lucas's assertion that Gödel's first incompleteness theorem proved that machines could not model minds. One objection, by Professor Judson Webb from Boston University, was that he could construct a machine that can calculate the Gödelian formula, as Gödel provided a mechanism for such a calculation. If the machine can calculate the Gödelian formula for a different machine, then Webb shows that it can also do so for itself. In this way, then the machine could continue to calculate the necessary formula, and could keep up with a human mind that is constructing the Gödelian formulas for these machines.[32]

There are also academics that believe that neither Lucas, nor objectors, such as Webb, are correct. Professors Jongeneel and Koppelaar from Delft University of Technology in the Netherlands, write that neither of these arguments is sufficient, as they are actually about the

[31] Ibid. Pg. 115.
[32] Jongeneel, C.j.b, and H. Koppelaar. Pg. 657.

limitations, or lack thereof, of human imagination. For if and only if the human imagination is unbounded, will there always be a Gödel formula that a human mind can construct that cannot be proved by a machine.[33] Since the limitations of the human imagination cannot be proved to be unbounded or bounded, Koppelaar and Jongeneel believe that "Gödel's theorem cannot be used conclusively to argue either the superiority of the human mind over the machine, or the equal capacities of both."[34] While Lucas had his own response to the objections that were raised, they revolved around the philosophical ideas, rather than any physical aspects of the brain, and so computer scientists have tended to dismiss this particular argument. This is why physicist Roger Penrose is so pivotal. Though the way in which he relates Gödel's first incompleteness theorem to the mind-machine problem is similar to Lucas, Penrose has provided physical explanations, that is, quantum gravity, as to how the mind is non-algorithmic and has constructed an argument beyond Lucas's in more detail and depth.

The argument that Penrose constructs against the feasibility of strong AI essentially builds off of Lucas's argument. Penrose asserts that Gödel sentences – statements that cannot be proven within a formal system, such as a statement about a formal system's own consistency – are provable by human mathematicians. Furthermore, because human mathematicians can prove these statements, Penrose argues that mathematicians cannot be represented by formal systems and therefore must have a non-computable algorithm within their minds. This is then generalized for all humans.[35] He writes,

---

[33] Ibid. Pg. 658.
[34] Ibid.
[35] Penrose, Roger. *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford University Press. 1989. Pg. 480.

> The inescapable conclusion seems to be: Mathematicians are not using a knowably sound calculation procedure in order to ascertain mathematical truth. We deduce that mathematical understanding – the means whereby mathematicians arrive at their conclusions with respect to mathematical truth – cannot be reduced to blind calculation![36]

In essence, the argument mirrors Lucas's quite well. Hence why it is often referred to as the Penrose-Lucas argument. However, one distinction is that Penrose takes it a step further and attempts to explain what this non-computable algorithm within human minds is. He created a theory called Objective Reduction to explain this phenomenon in the brain, and attempts to reconcile general relativity with quantum mechanics. As of yet, no one has disproven this theory, and his ideas are all considered to be eventually testable although extreme. Even though some believe Penrose's explanation, it is the minority opinion. Penrose's, and to a lesser extent Lucas's, arguments will be explored in more detail in Chapter 2.

In this way, Gödel's theorems disrupted the field of formal logic and Hilbert's Program, and were later integrated into AI debates through the arguments of Lucas and Penrose. Now that the basics of the theorems and relevant arguments have been established, it is time to return to before Gödel had published his profound theorems, investigate the theorems' incorporation into the AI debate from a historical and information transmission-based perspective, and analyze the relationship between theology and the theorems in the context of Gödel's life and the AI debates.

---

[36] Penrose, Roger. *What is Intelligence?: Mathematical intelligence.* Editor Jean Khalfa. Cambridge University Press, Cambridge, United Kingdom, 1994. Chapter 5, Pgs. 107-136.

## Chapter 1: Kurt Gödel

*"Mathematics, viewed rightly, possesses not only truth, but supreme beauty."*
*– Bertrand Russell*

Mathematician Kurt Gödel greatly influenced the field of mathematical logic.[37] Although initially misunderstood, once his incompleteness theorems were fully grasped they had repercussions reaching far beyond the field. It is for these theorems, as well as his completeness theorem for first order predicate calculus, that Gödel is most well known among laymen. Some of his other ideas, however, are equally intriguing though they do not hold the same gravity for the field of mathematics. For example, a paper that Gödel gave to Einstein in 1949 supplied a mathematically sound solution to Einstein's field equations of general relativity that allowed for the existence of closed time-like curves, which meant that time travel would be possible. [38] Before investigating the way in which Gödel's theorems eventually permeated the field and provided the basis for some philosophical arguments against strong AI, it seems apt to provide the context of Gödel's life, look at the spread of his theorems, and then examine the definition of computability and relevance of Alan Turing.

[37] He also completed extensive work in set theory, furthered intuitionistic logic and arithmetic, and incessantly reevaluated philosophical ideas such as realism, Platonism, and rationalism.

[38] Gödel gave this solution to Einstein as a gift for his birthday; however, it actually disturbed Einstein somewhat that such a mathematically sound solution could be provided to his equations. In Gödel's paper, he uses a cosmological term that has a negative value rather than being equal to zero. His solution is a static and not expanding universe, meaning that this is not a realistic model of the universe since it did not contain Hubble's expansion of the universe. Gödel as much in his paper, but this has not doused the fascination that many have found with such a concept of the universe. It also demonstrates the importance of choosing the correct constants, and how mathematical equations can vary.

## Gödel and the Incompleteness Theorems

Gödel was born in Brünn, Austria-Hungary to an ethnically German family. His father, Rudolf Gödel was Catholic, while his mother, Marianne Gödel, was Protestant, but Kurt and his older brother, Rudolf, were raised Protestant.[39] Following World War I, Kurt Gödel, along with the rest of his family, automatically became citizens of Czechoslovakia, because of the way that the Habsburg Empire had been broken up. However, Gödel was never comfortable with this and so later became an Austrian citizen. In the build up to World War II, when Germany annexed Austria, Gödel automatically became a German citizen. Then, after immigrating to the United States and once World War II had concluded, Gödel became a U.S. citizen. He lived in the U.S. for the rest of his life.

While still a young child, Kurt suffered from rheumatic fever and after recovering became convinced that he had sustained permanent damage, despite doctors confirming otherwise. Even though he had little to no reason to worry, Gödel was always concerned over his health because of this event. This kept him away from many athletic activities as a child, so he spent most of his time with books and studying. He was exceptional in all subjects, but took particular interest in mathematics and physics.[40]

He attended a selective school where he received top marks in all subjects, except for once in mathematics. World War I had little effect on Gödel and his family's lives, but the family did lose a fair bit of money due to their investments in the German War Loans. Gödel graduated from his school, the Realgymnasium on June 19th, 1924, and that following autumn he moved to Vienna to attend the University of Vienna. Gödel only attended the

[39] Dawson, John W. Pgs. 1-3
[40] Dawson, John W. Pgs. 1-3

University of Vienna even though it was quite common at the time to spend a term or year at several different universities. He received his doctorate and then stayed on to become an unpaid lecturer in 1933, and eventually a paid lecturer.[41]

Beginning in 1933, Gödel made a series of visits to the United States, primarily visiting the Institute for Advanced Study at Princeton, where he would lecture and teach classes for a semester at a time. This continued for a number of semesters. Gödel would visit the United States to lecture, then return to Vienna. Throughout this time, even though there were still logicians who did not fully understand Gödel's results, he was beginning to gain fame within the mathematical community. Some more prominent mathematicians advocated for him, since he would not do so for himself. It was primarily because of the pivotal results of Gödel's incompleteness theorems that he was invited to lecture at Princeton, as well as later at the University of Notre Dame.

Gödel was a visiting professor at the University of Notre Dame in the spring of 1939, having been invited by Professor Karl Menger. His coming was announced in the school newspaper, as it was for all new professors, but also in a short story about his prominence in the mathematical logic community due to his incompleteness theorems.[42,43,44] In 1937, Menger had left the University of Vienna, accepting a professorship at the University of Notre Dame. This was during the time that many academics were searching for ways to escape the growing reach of Nazism. While at the University of Vienna though, Menger had run a mathematical colloquium that Gödel had participated in since 1929. So, when Notre

---

[41] Ibid. Pgs. 14-23.
[42] "Many New Teachers Join Faculty." *The Notre Dame Scholastic*, 24 September 1937, University of Notre Dame Archives.
[43] "Many Faculty Changes," *The Notre Dame Alumnus*, October 1937, University of Notre Dame Archives.
[44] Gödel, Kurt - Math, UDIS 105/29. University of Notre Dame Archives. Notre Dame, IN.

Dame was looking to recruit prominent European scholars in order to strengthen its reputation as an academic institution, Menger suggested Gödel, and Gödel, with some negotiation in the typical Gödel way, accepted.[45]

During his time at Notre Dame, Gödel lectured three hours a week on the power of the continuum and his consistency proofs. He had previously lectured on this material at Princeton, but he rewrote his notes for his lectures. Gödel also assisted Menger in joint seminar on introductory logic. There have been a number of times when Gödel's advanced lecture notes were almost published, but nothing came of it. In addition, none of Gödel's correspondence with his wife Adele, his mother, or his brother survived from this period.[46]

Although Gödel did express belief in an afterlife and theism during his life, he was not a Catholic, and actually expressed disdain towards some Catholic doctrines in his letters to his mother. This is understandable though, particularly as in Austria a Catholic clerico-fascist state had developed leading up to the annexation of Austria by Nazi Germany in 1938. Nazi Germany also invaded Czechslovakia, however, this occurred while Gödel was teaching at Notre Dame. Since his mother had returned to Brno, Czechoslovakia at the end of 1937, this caused Gödel much concern and was one of his reasons for returning to Vienna.[47] He was also motivated to return in order to secure his position as a lecturer at the University of Vienna.[48]

When World War II broke out in September 1939, Gödel's friends, who had been trying for some time to convince him to move permanently to the United States, helped to arrange for his and his wife Adele Nimbursky's safe passage from Vienna to the U.S.

---

[45] Dawson, John W, Jr. "Logic at Notre Dame: Kurt Godel at Notre Dame." *Notre Dame Journal of Formal Logic*.
[46] Ibid.
[47] Ibid.
[48] Dawson, John W. *Logical Dilemmas.*

Eventually, he was able to obtain permission from both governments and took the Trans-Siberian Railway to Japan then sailed to San Francisco. They arrived in the spring of 1940 and returned to the Institute for Advanced Study.

In 1946, Gödel became a permanent member of the Institute for Advanced Study at Princeton, then in 1953 he became a full professor, and in 1976 a professor emeritus. He resided in the area for the rest of his life. Gödel received a variety of accolades, including the first Albert Einstein Award and an honorary Doctorate in Science from Harvard University. He also became a U.S. citizen in the winter of 1947. Gödel died on January 14th, 1978, due to malnutrition and inanition caused by personality disturbance. With his wife in the hospital and unable to prepare his food for him, Gödel did not eat and his paranoia caused him to starve himself to death. Even though Gödel's death may have been dark, his legacy was great. Many of Gödel's theorems were the start of entirely new branches of mathematics, and he has often been compared to Einstein[49] in terms of intellect and contributions to his respective field.[50]

Not only was Gödel a renowned logician and mathematician, but he also contributed to the fields of philosophy and physics. During his daily walks with Einstein at Princeton University, they would often discuss various physics problems. In 1946, Gödel published a model that solved Einstein's field equations – this was supposed to be a birthday present for Einstein – in which time is cyclical and by making a round trip on a rocket with a wide enough curve, it would theoretically be possible to travel into any region of the past, present, or future, then back again. Einstein was rather disturbed by this model of time, but admitted

[49] Einstein and Gödel were actually very good friends! Originally, Gödel had wanted to study physics, while Einstein had wanted to pursue theoretical mathematics. So, in that way they were images of what the other could have been.

[50] Goldstein, Rebecca. Pgs. 250-252.

that the mathematical deductions were sound. Gödel on the other hand, greatly enjoyed this idea of cyclical time. One suggestion is that perhaps Gödel thought of his life as incomplete and might have liked to imagine that it was possible for him to return to his days as a youth in Vienna.[51]

In the years before 1931, when Kurt Gödel had yet to establish himself as a renowned mathematician, he was living in Vienna, a city full of intellectual and creative activity that had not been diminished by the fall of the Habsburg Empire in 1916. Throughout the late nineteenth century in the Habsburg Empire, the voting population expanded beyond liberal elites, thereby pushing politicians to use ideas such as nationalism to unite a less educated voting population. By basing politics on ethnic or religion grounds, politicians were able to mobilize the masses who had just received the right to vote.[52] Due to this appeal to ethnic identity politics, as well as a stock market crash, distrust of liberals grew and many liberals were evicted from the very government they had created.[53] These liberals then often turned to art and culture, thereby adding to the vibrant and lively culture that permeated Vienna. Within this artistic culture, there were also movements pushing the boundaries of thought and art. For example, Sigmund Freud, a scientist at the forefront of psychological research, found Vienna to be his place for innovation and intellectual pursuit, as did Gustav Klimt, a famous artist and leader of the Secession[54] movement.[55] Other intellectuals from abroad also flocked to Vienna, such as Trotsky, Stalin, Lenin, and Theodore Herzl.[56]

---

[51] Ibid. Pgs. 256-260.
[52] Schorske, Carl E. *Fin-de-siglo Vienna: Politics and Culture*. Cambridge: Cambridge U.P., 1985. Print.
[53] Judson, Pieter M. "Culture Wars and Wars for Culture." *The Habsburg Empire: A New History*. Cambridge, MA: Belknap of Harvard UP, 2016. Print.

[54] The Vienna Secession was an art movement beginning in 1897 that objected to the conservatism of Vienna art institutions, particularly the Vienna Künstlerhaus. Many of the members were artists who had resigned from the Association of Austrian Artists, and their new work was not only aesthetic, but also, philosophical and

Despite the collapse of the Habsburg Empire, in which this vibrant intellectual environment arose, Vienna continued to flourish during Gödel's time there. The city had a culture separate from Austria as whole, and Vienna even had the only real university in Austria. Cafés were common gathering places, where culture and ideas were exchanged, but there were also many somewhat formal discussion groups that met weekly. One of the most prominent of these meeting groups revolved around philosopher Moritz Schlick[57] – this circle was where logical positivism[58] spread. Kurt Gödel was invited to join this group, and did so from 1926-1928, while an undergraduate student at the university. Due to his association with this group, Gödel was often construed as a positivist and thus his theorems have been understood through that context. However, Gödel had become a Platonist[59] in 1925 and strictly adhered to Platonism.[60]

---

political. The primary concern of Secession artists was the possibilities of art outside of the confines of academic tradition and that was influenced by the dynamic nature of Vienna at the time.

[55] Schorske, Carl E.
[56] Morton, Frederic. *Thunder at Twilight: Vienna, 1913-1914*. London: Methuen, 2001. Print.

[57] Moritz Schlick was a philosopher and physicist, as well as the founding father of logical positivism. He led a group of philosophers and scientists who met regularly in Vienna, and initially this group was called the Ernst Mach Association. Later, they became known as the Vienna Circle. When tensions rose in Germany and Austria, leading up to World War II, many of the Vienna Circle's members left for the U.S. or U.K., however, Schlick decided to stay. Schlick was murdered in 1936 on the steps of the University of Vienna where he worked. This was a rather traumatic event for Gödel and signaled to much of the academic community in Vienna that the political environment was becoming toxic (leading up to World War II) and it was time to emigrate. While Gödel did not adhere to logical positivism, he did attend meetings of the Vienna Circle and did not voice his disagreement to the philosophy, which explains why many people initially interpreted his results in the context of logical positivism, under the assumption that Gödel must have adhered to that philosophy.

[58] As briefly mentioned in the introduction, logical positivism essentially adheres to a theory of knowledge where only statements that can be verified via empirical observation are considered meaningful. It was a Western philosophy movement that was flourishing during the 1920s and 1930s, particularly in Vienna and in Berlin.

[59] Platonism, as evident from the name itself, was the philosophy of Plato and his followers. The theory is that ideas are timeless, objective entities that provide true knowledge, whereas physical objects are imperfect representations of these unchanging ideas. In contrast to logical positivism, Platonism finds statements to be meaningful even when they are not analytical, conclusively verifiable, or observable. Platonism and logical positivism take rather opposing positions, so it was quite ironic that so many people interpreted Gödel's theorems as supporting logical positivism when, in fact, Gödel asserted that they supported Platonism. In

Gödel's adherence to the principles of Platonism is important in the way that it colored his understanding of mathematical truth, as well as with regards to how it relates to his theological leanings. According to logical positivism, all meaningful thought can be reduced to sense perceptions, implying that mathematics, as a collection of nonempirical statements, is meaningless. However, Gödel believed that concepts and truths ought to be reduced to Platonic ideas, not sense perceptions. This is a fundamental difference and incompatibility between logical positivism and Platonism, and can be seen to argue either against or for the reality of mathematical truth. Not only were there profound differences between Gödel's thoughts and beliefs and those of the positivists, but Gödel interpreted his first incompleteness theorem, that a sufficiently consistent system is incomplete, as disproving the positivist antimetaphysical position. Yet, there were many positivists who misunderstood and took the incompleteness theorems to mean that mathematics was on shaky ground and therefore sense perceptions must be trusted instead.[61]

While Gödel's adherence to Platonism likely informed some of his work, it is also possible that some of his theological beliefs had a similar effect on the problems he decided to pursue, as well as the conclusions he drew. Gödel described his personal beliefs as "baptized Lutheran (but not a member of any religious congregation). My belief is theistic, not pantheistic, following Leibniz, rather than Spinoza." [62] Beyond this vague theological statement, there are other indications that Gödel had some sort of belief in a Christian-like religion, including the fact that he worked on an ontological proof, that is, a formal argument

---

addition, Platonic realism asserts that mathematics does not create its objects, that is, classes and concepts within mathematics, but that they already timelessly exist and mathematicians are just discovering them. This was a view that Gödel also seemed to support.

[60] Goldstein, Rebecca. Pgs. 70-75.
[61] Ibid. Pgs. 111-112.
[62] Wang, Hao. *Reflections on Kurt Gödel*. Cambridge, Mass. U.a.: MIT Pr., 1987. Pg. 18.

for God's existence. This will be explored further in Chapter 3, as will the theoretical connections between Gödel's work and theology.

Before Gödel published, or even worked on his incompleteness theorems, he first worked on the completeness of first-order calculus. It is believed that Gödel first began work on his dissertation, the work that led to his completeness theorem, in 1928 or early 1929. Just a bit earlier, Gödel's mathematical interests had shifted from classical mathematical fields toward logic and foundations.[63] Informally, the completeness theorem states that any first-order result that is true in all models of a theory must be logically deducible from that theory, and vice-versa. It establishes a connection between semantic truth and syntactic provability in first-order logic, that is, if a statement is logically valid then there is a finite deduction, or proof, of the formula.

While Gödel's dissertation, the completeness of first-order calculus, was an important result, it has not gained the same level of fame outside of the mathematical community as his incompleteness theorems. Additionally, many people mistakenly take the names of the different theorems, completeness and incompleteness, to imply incompatibility between them. However, there are two different senses in which the term "complete" can be used within mathematical logic. In the case of the completeness theorem, complete can be taken to mean able to prove whatever is valid. On the other hand, in the incompleteness theorems, complete means able to prove or disprove every sentence that can be constructed within the formal system. Beyond the different meanings of completeness, the theorems also apply to different systems. The incompleteness theorems are applicable for higher-order formal systems, whereas the completeness theorem is only applicable for first-order

[63] Dawson, John W. Pgs. 53-54.

formalizations.[64] Applying the theorems together gives a fascinating result that "no first-order axiomatization of number theory can be adequate to the task of deriving as theorems exactly those statements that are true of the natural numbers."[65]

After finishing his dissertation on completeness, Gödel began work on a problem posed by Hilbert – the influential German mathematician who supported formalism – that of giving a finitary consistency proof for the axioms of analysis.[66] As briefly mentioned in the introduction, this idea of a finitary proof was pivotal for Hilbert's Program. In this way, Gödel was not attempting to destroy Hilbert's Program, but rather, he was attempting to solve one of the key problems and thought, in doing so, he would be advancing the program. However, the result he discovered, his first incompleteness theorem, was unexpected, and from that Gödel developed the second incompleteness theorem, which really threw a wrench in Hilbert's Program and insistence on finitary consistency proofs.

The first person that Gödel informed of his result was philosopher Rudolf Carnap, a member of the Vienna Circle and an advocate of logical positivism. On August 26th, 1930 they met at a café and while the primary topic of conversation was their travel to an upcoming conference, Gödel also discussed his first incompleteness theorem a bit, though Carnap did not fully comprehend the implications of such a result at that point. Three days later, they again met at a café and discussed the first incompleteness theorem.

Gödel's first announcement of his findings to a larger crowd was a few days later at the Conference on Epistemology of the Exact Sciences, running from September 5-7.[67] This conference, held in Königsberg, was a follow-up conference to one that had been held the

[64] Ibid. Pg. 67.
[65] Ibid. Pg. 68.
[66] Ibid. Pg. 61.
[67] Ibid. Pg. 68.

previous year in Prague. The Prague conference was in connection with a conference of mathematicians and physicists. Similarly, in 1930, the conference was being held just before the sixth Assembly of German Physicists and Mathematicians happened.[68] Between the two conferences was a third, the ninety-first annual meeting of the Society of German Scientists and Physicians, which was the conference where Hilbert delivered the opening address.

On the last day of the Conference on Epistemology of the Exact Sciences, during a roundtable discussion, Gödel offhandedly made a remark about incompleteness. As quoted by Dawson, Gödel said "one can even give examples of propositions that, while contentually[69] true, are unprovable in the formal system of classical mathematics."[70] There were no follow-up questions to this statement, and only one mathematician, von Neumann, asked Gödel about further details.[71] Von Neumann became an advocate of Gödel's work, and was truly fascinated by it, as evidenced by his follow up letter in late November having discovered the second incompleteness theorem on his own.[72] By this time though, Gödel had already submitted an abstract of his work, containing both incompleteness theorems, to the Vienna Academy of Sciences.[73] Evidently, Gödel's incompleteness theorems did not make the immediate splash that one may have anticipated. Instead, it took time for more mathematicians and logicians to grasp the gravity of what he had proved, and for the field to reflect those findings.

---

[68] Reichenbach M., Cohen R.S. (1978) The Königsberg Conference on the Epistemology of the Exact Sciences [1930f]. In: Reichenbach M., Cohen R.S. (eds) Hans Reichenbach Selected Writings 1909–1953. Vienna Circle Collection, vol 4a. Springer, Dordrecht.

[69] This means with regard to content rather than context. So in this case, there are propositions that have content that can be seen to be true, but they are still unprovable.

[70] Dawson, John W. Pg. 69.
[71] Ibid.
[72] Ibid. Pg. 70.
[73] Ibid.

Once Gödel became recognized as one of the most important mathematical logicians, his findings were extrapolated to apply to other fields, but these were misapplications. While terms such as "consistency" and "completeness" are clearly defined in a technical sense within mathematical logic, they have many other applications in ordinary language, meaning that it can be easy to construe significance from the incompleteness theorems in applications to which they should not be involved.[74] One example of a misapplication that was previously mentioned is the U.S. Constitution. Another even more fascinating example is the commonly cited misapplication to the Bible. Since some religious people believe that all answers can be found in the Bible, the thinking goes then, this must mean the Bible is a complete formal system. But Gödel's incompleteness theorems indicate otherwise, thus it must be incomplete or inconsistent. However, the Bible and other religious texts are not formal systems. The Bible has no axioms, no rules of inference, and no theorems. It certainly does not contain a sufficiently strong arithmetic. These are misapplications of the first incompleteness theorem that neglect the essential condition that the system needs to be capable of formalizing a certain amount of arithmetic. It must contain a sufficiently strong arithmetic.[75]

Despite these misapplications of the incompleteness theorems, one legitimate application, though to what extent is debated, is the feasibility of AI debate. This mind-machine problem is one that Kurt Gödel even commented on in correspondence with a mathematician and philosopher named Hao Wang. Later in his life, Gödel was in touch with

---

[74] As a refresher, these are the technical definitions of consistency and completeness:
To be consistent means that there exists no statement such that the statement itself and its negation are both derivable in the system. So, a consistent formal system cannot have proven contradictions.
To be complete means that for every statement of the language of the system, either the statement or its negation can be proven in the system. This means if you can create a statement by following the set of rules for the system, then you must be able to prove that statement to either be true or false in order for the system to be complete.

[75] Franzén, Torkel. Pgs. 77-78.

Wang, who chronicled Gödel's philosophical and theological ideas and wrote several books on the topic. In Wang's analysis and writing about Gödel, he also noted Gödel's beliefs on minds, machines, and the applications of the incompleteness theorems. Gödel, as quoted by Wang, believed that "Either the human mind surpasses all machines (to be more precise it can decide more number-theoretical questions than any machine) or else there exist number-theoretical questions undecidable for the human mind."[76] So, Gödel applied his own incompleteness theorems, the first one in particular, to the debate regarding whether or not machines can become adequate models of the human mind. This is the feasibility of strong AI debate.

Knowing that Gödel understood the applications of his theorems to this debate, and believed them to be relevant, gives credibility to all the subsequent arguments regarding this topic. This is not a fanciful application of the theorems that is misjudging and misapplying them, but rather, a critical application to a complex topic of minds and machines. At the same time, just because they are relevant, does not mean that the conclusion ought to be that strong AI is not possible. In fact, a majority of computer scientists who are informed of or involved in AI debates believe that strong AI is possible.[77]

Though Gödel originally announced his first incompleteness theorem at a conference, he officially published both theorems, though the second was only an outline, in a German scientific periodical in 1931. The relatively short paper held the title "On Formally Undecidable Propositions of Principia Mathematica and Related Systems." Gödel was only 25 years old when he published this paper. In general, reading papers in periodicals,

---

[76] Goldstein, Rebecca. Pg. 203.
[77] Aaronson, Scott. "Can computers become conscious?: My reply to Roger Penrose." *Shtetl-Optimized* (blog). 15 June 2016.

attending scientific conferences, and the exchanging of letters containing proofs or ideas was the way that this type of mathematical information was disseminated. Gödel's incompleteness theorems were dispersed through all of these methods.

As with many mathematical theorems, Gödel's incompleteness theorems were advanced by the work of other mathematicians.[78] Pivotal among these advancements and clarifications is J. Barkley Rosser's trick. As mentioned in the introduction, Rosser was an American logician who greatly advanced the applicability of Gödel's incompleteness theorems, and is also well known for his part in the Church-Rosser theorem, a theorem applicable in lambda calculus. Seeking to strengthen the first incompleteness theorem, Rosser introduced what became known as a Rosser sentence R. Rather than the Gödel sentence G, which stated "This statement is not provable," the Rosser sentence R claims "If this sentence is provable then its negation is provable with a proof of smaller Gödel number than this one's."[79] Applying this then to Gödel's proof gives the conclusion that the requirement for omega consistency can be weakened to consistency thereby strengthening the theorem and expanding its applicability. This was known as Rosser's trick and, because of it, there exist an uncountably infinite number of Gödel sentences.[80]

Beyond Rosser's trick, published in 1936, there are a variety of other contributions that mathematicians and logicians have made that are relevant for Gödel's incompleteness theorems. For instance, David Hilbert, the mathematician known for Hilbert's Program,

---

[78] The rest of this chapter becomes more mathematical, but I have done my best to write it in a way that is accessible to readers of all levels. If some of the mathematical ideas become confusing that is okay. The most important part is to understand how Gödel's incompleteness theorems relate to AI through Turing machines. This will resurface again in the following chapter, so if nothing else, just remember that modern computers are essentially forms of Turing machines, thus the incompleteness theorems apply.

[79] Franzén, Torkel. Pg. 43.
[80] Smith, Peter. Pgs. 98-100.

helped to formally prove the second incompleteness theorem, along with Paul Bernays, in 1939. This was quite ironic considering it was the second incompleteness theorem, that a consistent formal system cannot prove its own consistency, that disrupted Hilbert's Program and demonstrated that it was not possible to prove the formalization of all mathematics in axiomatic form through finitary proof methods. Gödel never actually gave a detailed proof for the second incompleteness theorem, but rather, just a fully convincing outline. The main challenge in proving the second incompleteness theorem is showing that various facts regarding provability that were used in the proof of the first incompleteness theorem can be formalized within the formal system using a formal predicate for provability. Essentially, there were further steps beyond proving the first incompleteness theorem that were necessary to use some of those facts in the second incompleteness theorem's proof.

Formalization of the First Theorem: $PA \vdash Con \rightarrow \neg Prov(G\ddot{o}del\ number\ of\ G)$

In this scenario, the formal system is Peano-Arithmetic[81], represented by PA, Con is the well-formed formula stating the consistency of PA, which claimed to be provable from PA, and from this claim it is implied that not Prov of the Gödel number of G is true, but this can be equated with the provability of the Gödel sentence G. So essentially, FFT is that the consistency of PA is provable from PA, which implies that G can be proved. However, we

[81] Considering the consistency of PA is an intriguing problem, since according to the two incompleteness theorems, PA's consistency cannot be proven within PA itself. Gentzen's consistency proof, published in 1936, shows that the axioms of First-Order Peano-Arithmetic do not contain any contradictions, thus the system is consistent, as long as another system, an extended version of primitive recursive arithmetic, does not contain any contradictions either. The two systems used in the proof are neither weaker nor stronger than one another. Not only is this proof important for understanding the consistency of PA, but also, it highlights a common misunderstanding of Gödel's second incompleteness theorem. This misunderstanding is the idea that the consistency of a formal system can only be proven in a stronger formal system, however, as shown in Gentzen's proof, this does not have to be the case, rather the caveat is that such proofs can occur in other formal systems that are not weaker.

already know that G cannot be proven so this causes a contradiction, meaning that if PA is consistent then the well-formed formula Con, which expresses its consistency, is not provable in PA.[82]

Another way that Gödel's incompleteness theorems were advanced was through Alfred Tarski's work on truth and provability. A Polish-American mathematician and logician, Tarski was educated in Poland then immigrated to the United States in 1939. He is most well known for his work on metamathematics, model theory, and algebraic logic, but he contributed to a variety of other mathematical fields as well.

> Tarski's Indefinability Theorem: No predicate of an adequate arithmetical language L can express the numerical property $\text{true}_L$ (the property of numbering a truth of L).

Essentially, this means that while syntactic properties of a formal system (such as provability) can be expressed inside the system itself through the process of Gödel numbering, it is impossible to express certain semantic properties (such as arithmetical truth) within the system. From this observation, it is possible to then derive incompleteness again as truth is not provability within a formal system, so under the assumption that the system is consistent and everything that is provable within the system is true, then there must be truths that the system cannot prove.[83] Tarski proved this result in 1933. However, in 1931 Gödel had already sent a letter noting this necessary distinguishing between truth and provability in his incompleteness theorems, thought he did not prove it.

---

[82] Smith, Peter. Pgs. 107-111.
[83] Ibid. Pgs. 106-107.

One final way in which the incompleteness theorems were strengthened was through their applications to a variety of formal systems. When Gödel proved his results, they were only applicable for a system called P. However, as mentioned previously, they have been expanded as to apply to other formal systems with sufficiently strong arithmetics,[84] such as Robinson Arithmetic and First-order Peano-Arithmetic.[85] The incompleteness theorems were generalized from the original theory, P, in which Gödel proved his result, to any axiomatized formal theory T that satisfies two particular conditions:

1. T is primitive recursive axiomatized
2. T is strongly primitive recursive adequate[86]

These rules essentially clarify what type of mathematical theories, such as Peano-Arithmetic, to which the incompleteness theorems apply. In the rules there is mention of an idea called primitive recursion.[87] This is a specific type of recursion. Informally, recursion can be understood as repeatedly applying a rule to its results. Most functions that are studied in

[84] This is just a way of clarifying what qualifies as a basic amount of elementary arithmetic for a formal system. It is particularly important because this requirement of a formal system, having a sufficiently strong arithmetic, is what keeps the incompleteness theorems from being misapplied to issues such as whether the U.S. Constitution is consistent. Sufficiently strong is generally defined as being able to capture all decidable properties, decidable relations, and all computable functions of a formal system. More simply, this is being able to prove lots of well-formed formulas (wffs) about properties of individual numbers, prove relations between these numbers, and capture all computable functions. (Smith pags 30 and 31)

[85] Franzén, Torkel. Pg. 3.
[86] Smith, Peter. Pg. 95.

[87] A function f: $N^k \rightarrow N$ is primitive recursive if it is built from:
Constants: $C^n_m(x_1 \ldots x_m) = m$
Projection: $P^n_m(x_1 \ldots x_m) = x_m$ , given m <= n
Successor: $S(x) = x + 1$
While using the following:
Composition: If $f_1 \ldots f_n$ and g are primitive recursive then $g(f_1(x \text{ bar}) \ldots f_m(x \text{ bar}))$ is primitive recursive
Primitive recursion: f(0, x bar) = g(x bar), f(n+1, x bar) = h(n, x bar, f(n, x bar))
From lecture by Dr. Bays

number theory, such as addition and division, are primitive recursive, and primitive recursion implies recursion.

## Computability and Turing Machines

Recursion theory, also called computability theory, originated in the 1930s, and Gödel's work was a large part of this. In 1933, Gödel along with French mathematician Jacques Herbrand, created a formal definition of general recursive functions. These functions, also known as μ-recursive functions, are similar to primitive recursive functions and their inductive definition builds upon that of primitive recursive functions. However, not every μ-recursive function is a primitive recursive function. General recursive functions, that is, μ-recursive functions, are precisely the functions that can be computed by Turing machines.[88] Within recursion theory comes the idea of Turing computability, which is when a set of natural numbers is a computable set, also known as a recursive set. Specifically, the definition is as follows:

[88] Jacques Herbrand had already died by the time that Gödel published this work, however he had contributed to it. Herbrand was considered to be one of the greatest up and coming mathematicians, but he died while mountain climbing in the French Alps at age twenty-three. There are two theorems named after him, one regarding mathematical logic and the other regarding homological algebra. In early 1931, he submitted his principal study of proof theory and general recursive functions, and then added an appendix after reading Gödel's publication on the incompleteness theorems to explain how the results did not contradict one another. Later that year, in July, he fell to his death. When Gödel created a formal definition of general recursive functions in 1933, it was expanded and formalized from Herbrand's work.

> Turing computable (recursive) set: a set of natural numbers is a Turing computable set if there is a Turing machine, that, given a number n, halts with output 1 if *n* is in the set and halts with output 0 if *n* is not in the set.[89]

While there are more formal mathematical definitions, this way of understanding recursion is useful because it relates to Turing machines. Cue the entrance of Alan Turing.

Well known for his role in code breaking during World War II, Alan Turing was an English mathematician, computer scientist, philosopher, cryptoanalyst, and even a theoretical biologist. In 1936, Turing published a paper, *On Computable Numbers, with an Application to the Entscheidungsproblem*, where he leveraged results and methods, such as Gödel numbering, from Gödel's paper to design hypothetical devices called Turing machines. Given any computer algorithm, a Turing machine capable of simulating the logic of that algorithm can be constructed. The goal of the paper was to determine the uncomputability of the decision problem, the Entscheidungsproblem.[90] Unbeknownst to Turing, Alonzo Church, an American mathematician and logician, had been working on the same problem, but had proved his results using lambda calculus, which was less accessible, particularly for non-

[89] Copeland, Jack B. "The Church-Turing Thesis", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.).

[90] In the 1920s, the question of whether there is a systematic way to find a solution to every mathematical problem was formulated. This became known as the decision problem. In order to demonstrate that there cannot be a computational procedure, or an algorithm, that solves every mathematical question, Alan Turing first had to construct a mathematical model of computability, and then show that no computational procedure could solve the decision problem. This problem is also known as the halting problem, a problem of determining, from a description of an arbitrary computer program and an input, whether the program will finish running, that is, halt, or continue to run forever. Turing notes that if there were a computable procedure to solve every mathematical question, and then there would be a computable procedure to prove that an individual program halts or to prove that it never does. However, he demonstrates that sometimes it cannot be proved one way or another., that is, that a general solution to the decision problem is impossible.

mathematicians, than Turing's proof.[91,92] Together their work became known as the Church-Turing thesis.

> Church – Turing thesis: Any function that is computable by an algorithm (a calculational procedure) is a computable function.[93]

Before this formal definition of computability, mathematicians would use the informal term "effectively calculable" to describe functions that were computable via paper and pencil methods. The Church – Turing thesis not only officially defined computability, but also provided a way to relate Gödel's incompleteness theorems to the mind via Turing machines. Informally, these machines can be defined as follows:

> Turing machine: a mathematical model of a machine that mechanically operates on an infinite tape with symbols. These symbols can be read and written by the machine one at a time, using something called the tape head. There is a finite set of elementary instructions that the machine operates according to, including: read, fetch, write, move, and repeat.[94]

---

[91] Alonzo Church published his paper about a month prior to Alan Turing, prompting Turing to rush his results to publication. During their work on these papers, neither man knew that the other was working on the same problem. They both proved the same result, but used different methods. In particular, Church defined the notion of computability using lambda calculus whereas Turing used Turing machines. Lambda calculus is a formal system that consists of constructing lambda terms according to particular rules and performing reduction operations on those terms.

[92] Copeland, Jack B.
[93] Ibid.
[94] De Mol, Liesbeth. "Turing Machines", *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.).

For instance, a Turing machine will read in the symbol, fetch the matching rule, write the designated symbol, transition to a new state, move right or left depending on whether the matching rule's last column had +/- 1, and repeat until Halt is reached. Whatever is on the tape is the program's output. In this way, a Turing machine is all-or-nothing. Its purpose is to compute one function then reach the Halt state. A simple example of what this type of instruction looks like is "in state 37, if the symbol is 0, write a 3." A more complete example of a Turing machine is as follows:

Input: binary string Output: binary string + 1

S = {S, $S_0$, $S_1$, Y} where S is the set of possible states to be in (moves from the start state to other states until reaching the end state and stopping)

A = {B, S, 0, 1} where A is the alphabet i.e. the possible symbols (B=blank, S=start, 0 and 1 are the numbers in the binary string)

Rough idea: go to the right end of the string, change 1's to 0's and go left, change first 0 to 1, stop (these instructions would be transition functions rather than words)

Input: 1001 Output: 1010

Input: 1111 Output: 10000

It is important to note that the instructions according to which the machine operates are an intrinsic part of the machine. This helps to clarify that a Turing machine is not a mathematical model of computers, but rather, it was constructed as a method of formalizing the definition of computation. There is also a particular type of Turing machines denoted as a universal Turing machine (UTM) that can simulate an arbitrary Turing machine on an

arbitrary input. In this way, a UTM is not limited to one particular set of instructions or one particular purpose, since it can simulate any arbitrary Turing machine, but rather, it is a general-purpose computer. It is generally accepted that Turing machines, specifically UTMs, can be used to model anything that a digital computer could do, as a modern computer essentially operates like a Turing machine absent the infinite tape, instead using some other form of data storage, such as the memory in a typical computer. Thus, modern digital computers operate like theoretical Turing machines using discrete parameters.[95] One key difference is that the TM/UTM has infinite memory on its tape, whereas modern machines only have extremely large memories that can approximate the capabilities of an infinite memory. Though different on a hardware level, the conceptual workings of a TM/UTM can relate to a modern machine, encompassing its capabilities, and the relationship is more about the theoretical capabilities of the two than their actual hardware.

In addition to defining computability, Alan Turing's results reinforced what Gödel had previously proven, that Hilbert's Program was not possible as originally intended. This is because Turing was trying to solve the halting problem, which in his paper meant finding a universal algorithm for determining whether or not a Turing machine would stop. Turing proved that there is no such universal algorithm, though this does not mean that for any particular Turing machine we cannot determine if the machine will stop. According to Hilbert's Program, there ought to be a method to prove or disprove the truth of any mathematical proposition within a formal system, but Turing demonstrated that this was not

[95] Penrose, Roger. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. London: Vintage, 2005. Pgs. 24-25.

the case.[96] Gödel had shown this earlier through his incompleteness theorems. So, Turing's results reinforced Gödel's.

Beyond the mathematical significance of Turing's results, his Turing machines are useful for understanding how Gödel's incompleteness theorems have been applied to the problem of AI.

Turing machines are susceptible to the incompleteness theorems because of the way that they are constructed as instantiations of formal systems, however, not all machines have to be Turing machines, as pointed out by philosopher Paul Benacerraf.

> A Turing machine is an instantiation of a formal system, as Turing machines are mathematically defined objects. However, this does not mean that all modern computers are Turing machines and therefore formal systems in the same sense. Theoretically, there might be machines that do not satisfy the specifications of a Turing machine and modern computers, with sufficient computational power, could accomplish any task that a Turing machine could. From this arises the question of how these non-Turing machines relate to the incompleteness theorems and whether Lucas's argument is applicable to them as well.[97]

While there may be theoretical machines that could not be simulated by a Turing machine, Benacerraf, in his reply to philosopher John Lucas, did not attempt to contrive an example of a non-Turing machine and examine how Gödel's incompleteness theorems might, or might not, affect its capabilities. Perhaps this is because Lucas did not differentiate between

---

[96] Penrose, Roger. *The Emperors New Mind: Concerning Computers, Minds and the Laws of Physics*. Oxford: Oxford University Press, 2016. Pg. 117.
[97] Benacerraf, Paul. "God, the Devil, and Gödel." *Monist* 51, no. 1 (1967). Pg. 14.

machines and Turing machines, but another likely reason is the difficulty that might arise in contriving such an example, and whether it is even necessary for addressing the possibility of using a machine to model the mind.

Beyond the importance of the Church-Turing thesis, Alan Turing also made relevant remarks regarding the feasibility of strong AI. For instance, Turing believed that mental processes could not surpass mechanical ones. He even wrote about this issue in his 1936 paper on the uncomputability of the decision problem. Turing wrote, as quoted by historian John Dawson, "the justification [for the definitions of how his machines were to operate] lies in the fact that the human memory is necessarily limited."[98] However, while Turing advocated for bounds on the power of human reason, Gödel argued that it was unbounded and was greatly troubled by Turing's claim. In 1969, Gödel asserted that he had discovered a way to refute Turing's idea. As quoted by Dawson, Gödel argued, "mind, in its use, is not static, but constantly developing."[99] In this way, human capacity for greater knowledge and understanding is unlimited, whereas Turing's results placed limits on the potentialities of formal systems.[100]

These remarks by Turing and Gödel, made before Lucas or Penrose had constructed their arguments against the feasibility of strong AI, indicated a way in which the incompleteness theorems would be integrated into the feasibility of strong AI debates and demonstrated Gödel's own interest in the topic, particularly as the theoretical possibility of strong AI intensified.

---

[98] Dawson, John W. Pg. 232.
[99] Ibid.
[100] Ibid.

## Chapter 2: Debates over the Feasibility of Strong AI

*"You've got to play with ideas that are sort of on the edge of what we know, otherwise you're stuck with what we know." – Roger Penrose*

One of the most famous advocates for the idea that strong AI is not possible, due to the incompleteness theorems, is Roger Penrose. He asserts that there is a non-algorithmic element to the human brain, an element that must exist because of the incompleteness theorems, and that this non-algorithmic element can be explained through quantum gravity. As a physicist, Penrose has incredibly impressive credentials, but when it comes to AI, computer scientists tend to dismiss his ideas. This chapter will explore the build up to Penrose, specifically examining philosopher John Lucas's Gödelian argument, explore further who Roger Penrose is, expound his ideas regarding AI, as well as counterarguments, such as those by theoretical computer scientist Scott Aaronson, and study the influence his books and lectures on debates about strong AI. Through this study it is possible to see how Penrose helped popularize the use of Gödel's incompleteness theorems to argue against the possibility of strong AI, arguments that continue to be referenced and examined today.

With regard to the application of the incompleteness theorems to the AI debate, Gödel is at times thought to be the first to connect the two, essentially advocating that a machine could not model the mind. He most clearly outlined this at a lecture in 1951, where he stated,

> "So, the following disjunctive conclusion is inevitable: Either mathematics is incompletable in this sense, that its evident axioms can never be comprise in a finite rule, that is to say, the human mind (even within the realm of pure mathematics)

> infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable Diophantine problems of the type specified…"[101]

Essentially, Gödel was arguing that his first incompleteness theorem implied that either the human mind is not a Turing machine or that there are certain unsolvable mathematical problems.[102] This is not fully denying the idea that there is a theoretical possibility of modeling the mind, just that then our comprehension of mathematics is different than what we thought. However, in a 1998 paper, John Lucas commented on Gödel's statements, asserting that Gödel "was implicitly denying that any Turing machine could emulate the powers of the human mind." Lucas believed that Gödel thought his second option, that there are certain unsolvable mathematical problems, to be clearly false, thus meaning that the human mind could not be a Turing machine.[103]

Although Gödel and Turing made their own comments regarding the feasibility of strong AI and the application of the incompleteness theorems to this discussion, the first detailed presentation of this argument against the possibility of machines to model human minds was by philosopher John Lucas.

---

[101] Gödel, Kurt. *Collected Works*. Edited by Solomon Feferman. Oxford: Oxford University Press, 1995. Pg. 310.

[102] There are a number of unsolved mathematical problems, but proving that a particular problem can never be proven or disproven is an extraordinarily difficult task. Consider the Millennium Prize Problems. These are seven problems proposed by the Clay Mathematics Institute on May 24th, 2000. One of them has been solved, but the rest still have not been proven or disproven. They are complicated problems, such as P versus NP, the question whether or not, for all problems for which an algorithm can verify a given solution in polynomial time, an algorithm can also find the solution in polynomial time. Proving this problem, or disproving it, is quite difficult, but even more so is proving that such a problem could never be proven or disproven. Saying that there are unsolvable mathematical problems require proving that a particular problem can never, despite all advances, be solved. Simply finding problems that have not yet been solved is not sufficient to make the claim that there are unsolvable mathematical problems. Thus, claiming that there are unsolvable mathematical problems is a difficult claim to make.

[103] Megill, Jason. "The Lucas-Penrose Argument about Gödel's Theorem." *Internet Encyclopedia of Philosphy*.

## Lucas and Benacerraf

According to Lucas, "Gödel's theorem must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system."[104] More simply put, a machine is essentially a physical form for a formal system. It operates according to specifically defined rules and performs a set of operations. Since it is a concrete example of a formal system, Gödel's incompleteness theorems apply. Specifically, the first incompleteness theorem is relevant. This means that any consistent machine or computer will have a formula that is true but that the machine is unable to prove. An outside machine could be programmed to prove this formula or statement, however, in this new machine there would be a new formula that could not be proved. And so, there becomes an endless cycle.

Lucas then differentiates this from minds, writing, "for every machine there is a truth which it cannot produce as being true, but which a mind can."[105] Essentially, Lucas argues that human minds are not like formal systems, because they have consciousness, which allows a mind to make a meta-argument about itself and circumvent the issue of incompleteness. Thus there will always be something that a machine cannot prove that a mind can. In addition, even though taking all the machines in the world and linking them together would likely surpass the intellectual capacity of a human mind, Lucas clarifies that this group of machines cannot model the mind. It is about machine and mind not being the same, as in the mind is not simply a computational machine.

---

[104] Lucas, J.R. "Minds, Machines and Gödel." Philosophy 36, 1961. Pg. 113.
[105] Ibid. Pg. 115.

Lucas published this paper, "Minds, Machines, and Gödel," in 1961, and asserted that most mathematical logicians agree that machines cannot model minds, but were reluctant to put forth their views without a fully formulated argument against mechanism, that is, the idea that machines can simulate human minds. Hence why Lucas constructed his paper – to provide this complete argument. This was Lucas's first construction of his argument. In response to Professor Paul Benacerraf's article "God, The Devil, and Gödel," Lucas wrote another paper, "Satan Stultified: A Rejoinder to Paul Benacerraf," clarifying his argument and addressing particular concerns.

Benacerraf is a French-born American philosopher who taught at the Princeton University until his retirement in 2007. He specialized in philosophy of mathematics and published a reply to J.R. Lucas's argument against the feasibility of a machine that could be the equivalent of the human mind. His article, published in 1967, summarizes Lucas's idea as, "man couldn't be a machine – for (he argues) man can do something which has been shown by Gödel to be beyond the reach of machines."[106] Benacerraf then attempts to deconstruct and reconstruct Lucas's argument. He does this because Lucas's original argument is dialectic, that is, the proof is through a dialogue between a mechanist who has created a machine and a human. The argument is also with regard to the theoretical capabilities of a machine and a human given an infinite amount of time and resources.[107] Benacerraf works to formalize this dialectic argument into a direct proof, but in doing so criticizes some of the weaker points of Lucas's argument, and eventually reaches a much weaker conclusion.

[106] Benacerraf, Paul. Pg. 10.
[107] Ibid. Pg. 17.

The primary issue that Benacerraf takes with Lucas's argument is that while it is possible to understand some formal systems, and therefore construct their respective Gödel sentences, this does not entail that we can construct and see the truth of our own Gödel sentences. If we cannot obtain the insight to construct our own Gödel sentences, and see their truth, then we might not be different from machines. In addition, there is the possibility that, as Lucas asserts, we are not equivalent to Turing machines, but this might not be because we are more than a Turing machine, but rather, because we are less.[108] Benacerraf's eventual conclusion, after formalizing Lucas's dialectic proof into a direct proof, is that we cannot know if we are a Turing machine, and if we are, we cannot know which one we are. He writes, "If I am a Turing machine not only can I not ascertain which one, but neither can I ascertain of any instantiation of the machine that I happen to be that it is an instantiation of that machine."[109] Since he cannot know what Turing machine he is, if he is a Turing machine, then he cannot know himself. Benacerraf invokes Socrates at this point, "If I am a Turing machine, then I am barred by my very nature from obeying Socrates' profound philosophic injunction: KNOW THYSELF."[110] Thus, not only does Benacerraf reach a weaker conclusion, but he delves into some of the implications of such a conclusion.

In Lucas's response, he asserts that Benecerraf misses the point of much of his argument and also clarifies many of his points. The primary point is that any machine that a mechanist asserts is equivalent to the human mind will not be. Lucas writes,

> "An idealised person, or mind, may not be able to do more than all logically possible machines can, between them, do: but for each logically possible machine there is

[108] Ibid. Pg. 26.
[109] Benacerraf, Paul. Pg. 29.
[110] Ibid. Pg. 30.

> something which he can do and it cannot; and therefore he cannot be the same as any logically possible machine."[111]

Thus, since for any machine, it can be shown to not be the equivalent of the mind, then there cannot exist a machine that is equivalent to the mind. Letting this claim rest, there are plenty of issues with Lucas's argument, as Benacerraf and others have asserted. For instance, Lucas's claim that "machines cannot conjure up anything, but only act according to their input and the way they are wired up"[112] seems to only apply to top-down machines, that is, machines that follow very specific instructions, rather than bottom-up machines that use their algorithms to learn.[113]

One key issue that Lucas addresses is the idea of the consistency of humans. He had previously written on this in his first article, but Benacerraf found those arguments unconvincing. So, Lucas returned with a new idea. He writes, "we decide to be consistent."[114] Whenever propositions or assumptions that we have accepted turn out to be inconsistent, then we revise them to make them consistent. And so this pattern continues in a rational manner, thereby affirming our consistency because we choose to rationally

---

[111] Lucas, J. R. "Satan Stultified." *Monist* 52, no. 1. 1968. Pg. 146.
[112] Lucas, J. R. "Satan Stultified." Pg. 147.

[113] Penrose, on the other hand, clarifies in his book "Shadows of the Mind" that both top-down systems and bottom-up systems are included in the term "computation," thereby meaning that both forms of computational procedure can be put into a general purpose computer and included in his arguments are computational and algorithmic machines. Specifically, he defines a top-down system as one that has been organized such that it follows "a well-defined and clearly understood fixed computation procedure." On the other hand, bottom-up systems are ones "where such clearly defined rules of operation and knowledge store are not specified in advanced, but instead there is a procedure laid down for the way that the system is to 'learn' and improve its performance according to its 'experience'." A highly recognized example of the latter is an artificial neural network. It is pivotal to Penrose's argument that "computational" encompasses both these types of systems, and thus cannot easily be dismissed by new algorithms and methods in computer science. (Penrose shadow of mind pg 18-19)

[114] Lucas, J. R. "Mechanism: A Rejoinder." *Philosophy* 45, no. 172. 1970. Pgs. 149-51.

"discipline our thinking and not to allow ourselves to affirm anything whatever, but to draw some distinction between truth and falsehood."[115] In this way, it is a choice that humanity has made to course-correct in order to be consistent. While Lucas believed this to be a successful argument for the consistency of humans, it did not convince much of the scientific community. Many people currently believe that the inconsistency of humans is what invalidates Gödelian arguments.[116] Or, alternatively, they believe that the consistency of humans cannot be established one way or another, meaning that there would be incompleteness within humans, thus the truth of Lucas's argument cannot be established.[117]

While Lucas was the first to present a detailed argument against strong AI using the incompleteness theorems, Penrose has furthered the argument and brought new aspects to light.

## Penrose and Further Gödelian Arguments

Born in 1931, in Colchester, U.K., Penrose visited the United States with his family in 1939, and then moved to Ontario, Canada, as his father saw the indications of the looming war. Growing up in London, Ontario, Roger Penrose's interest in mathematics was stimulated by his family, consisting of his mother, his father, and his older brother. In 1945, the family returned to England, and Penrose attended University College School in London. After specializing in mathematics, physics, and chemistry at the University College School,

---

[115] Ibid.
[116] LaForte, G., Hayes, P. J., Ford, K. M. "Why Gödel's theorem cannot refute computationalism." *Artificial Intelligence* 104. 1998. Pgs. 265-286.
[117] Megill, Jason. The Lucas-Penrose Argument about Gödel's Theorem. *Internet Encyclopedia of Philosophy*.

Penrose enrolled at University College London to earn a B.Sc. in Mathematics. Next, he attended Cambridge focusing his mathematics research in algebra and geometry.

During his time at Cambridge Penrose developed a fascination with physics, prompting him to publish papers on cosmology in the subsequent years, while simultaneously holding various positions as a mathematics or applied mathematics professor at universities such as Princeton and Oxford. There were three particular lectures that Penrose attended while at Cambridge that influenced him to broaden his academic work, that is, one by Hermann Bondi on general relativity and cosmology, one by Paul Dirac, a founder of quantum mechanics, and one by a Professor Steen who talked on mathematical logic explaining Gödel's theorem and Turing machines.[118]

Before attending Steen's talk, Penrose had, in his own words, misunderstood Gödel's incompleteness theorems, believing the repercussions to be that there were things in mathematics that just could not be proved. During the talk by Stein though, he learned what it really meant. Penrose explains as follows. Suppose you have a method of proving things with numbers in mathematics, such as Fermat's Last Theorem[119] or Goldbach's Conjecture[120]. In mathematics you've got a system of methods of proof and you can have a computer check whether your proof is correct. Gödel's theorem says that if you trust this algorithmic

---

[118] Joe Rogan Experience Podcast. # 1216, "Sir Roger Penrose". Aired December 18, 2018.

[119] Fermat's Last Theorem, also called Fermat's conjecture, is a theorem from number theory stating that no three positive integers (a, b, c,) can satisfy the equation $a^n + b^n = c^n$ for any integer n that is greater than 2. If n=1 or n=2 then there are infinitely many solutions. Penrose uses this as an example of a method of proving things with numbers in mathematics, such as if you were asked if n=3 if there existed any three positive integers that could satisfy the equation $a^n + b^n = c^n$. By Fermat's Last Theorem it would be possible to prove this false.

[120] Goldbach's conjecture is an, as of yet, unproven conjecture that states that every even integer greater than 2 can be expressed as the sum of two prime numbers. It has been proved that this holds for all integers less than $4 * 10^{18}$, but still remains unproven. This is another example of a method of proving things with numbers in mathematics, that is, using the Goldbach conjecture one could say that, for example, 8 can be expressed as the sum of two prime numbers – 3 and 5.

procedure (you believe that if it says it is true then it is true) then you can see that there is something true but that cannot be proven by this method. Gödel constructs a very specific statement and shows that if you trust this algorithm then you can see by the way that the statement is constructed that it is true, but you can also see by the way that it is constructed that the algorithm cannot prove it. This seemingly indicates that what is going on in our heads is not following these algorithmic rules.[121]

This distinction that Penrose draws, between a mathematician's ability to see the truth of a Gödel sentence despite the system's inability to prove the truth of that same sentence, prompted him to search for an explanation as to why this is the case, and focused his exploration on the idea of consciousness. In a recent interview with Joe Rogan, Penrose explained his thought process once he understood the full depth of Gödel's theorems.

> "It is something else it is something that requires our conscious appreciation of what we are thinking about. Thinking is a conscious thing and understanding is a conscious activity. So I formed the view, that conscious activities, whatever they are, not just that kind of thing but you know playing music or falling in love, whatever else they might be, are not computations, there's something else going on."[122]

As a physicist and mathematician, Penrose was not content to accept that conscious activities were simply different, but rather, he wanted to find a scientific explanation for why they

[121] Joe Rogan Experience Podcast.
[122] Ibid.

were different. He eventually settled on the idea that something in the quantum world must be responsible, but he kept the view to himself for some time.[123]

It was not until Penrose heard a radio talk by Marvin Minsky, a cognitive scientist primarily concerned with artificial intelligence research, and Edward Fredkin, an early pioneer of digital physics, explaining their theories on what they believed computers could do. After hearing the talk, Penrose was inspired to write a book about his ideas regarding Gödel and the non-algorithmic nature of consciousness. The title, "The Emperor's New Mind", was designed to be a reference to the story about the emperor's new clothes. In the story, everyone seems to believe that the emperor is wearing special new clothes, but one young child notices that the emperor does not, in fact, have on any clothes at all. Essentially, this was to be a metaphor for how everyone thinks what is happening in human minds is computational, but standing back Penrose argues that there's something else going on.[124]

The Oxford University Press, the largest university press in the world and the second oldest, published Penrose's first book, *the Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics*, on November 9th, 1989. A foreword by Martin Gardner praises Penrose's ability to write a book that is accessible by non-professionals and informed laymen. Penrose also hoped to stimulate kids' interest in science, although the reality is that he primarily received letters from retired people.[125] In order to accomplish such a task, Penrose begins by first introducing the idea of a mind being modeled by a machine and the relevant philosophical propositions that have historically dominated the debate. These

[123] Ibid.
[124] Ibid.
[125] Ibid.

philosophical thought problems include: the Turing test and John Searle's Chinese Room Argument.[126]

The Turing test was a test established by Alan Turing, in 1950, as a way to determine whether a machine had the ability to exhibit equivalent intelligent behavior to that of a human being. According to the standard interpretation, there would be an interrogator who is posing questions to both a machine and a human. The interrogator must determine, based on the two participants' answers, which is human and which is machine. In this way, this particular interpretation of the test is not about the ability to give the correct answers, but rather, about the machine's ability to give human-like answers. With regards to the Turing test, Penrose accepts that imitation does not need to be the exact same as the real thing, so he writes, "*if* the computer were indeed able to answer all questions put to it in a manner indistinguishable from the way that a human being might answer them…then, *in the absence of any contrary evidence*, my *guess* would be that the computer actually thinks, feels, etc."[127] Thus, he prepares the reader to look for the contrary evidence that he will be providing, that is, Gödel's first incompleteness theorem and also legitimizes himself by taking an intermediary stance, rather than an extreme one.

Penrose utilizes the Chinese Room Argument similarly to set up logical gaps that he will later fill using Gödel's incompleteness theorems. There are two key claims within Searle's Chinese Room Argument. First, that brains cause the mind. And second, that syntax does not suffice for semantics. Essentially, the proposition is that if a monolingual English speaker is locked in a room and given a set of rules in English for correlating particular Chinese symbols with other Chinese symbols, then the person in the room could answer

[126] Penrose, Roger. *The Emperor's New Mind*.
[127] Ibid. Pg. 10.

questions in Chinese without knowing what he was saying. The argument is that the person in the room will never know what they are actually saying in Chinese, but their responses to an outside person's questions in Chinese will be so realistic that the outside person will not know that the inside person does not speak Chinese. Searle formulated this Chinese Room Experiment to demonstrate his idea that a machine can simulate the mind but that machine will never truly have a mind or consciousness. Penrose uses the Chinese Room Argument to introduce the concept of strong AI and how it encompasses an idea of machines with understanding, or as Penrose will later explain, with consciousness. So, Searle's argument is primarily a jumping off point, and Penrose also connects it to dualism and how the strong AU standpoint seemingly drives one into an extreme form of dualism.[128,129]

Once the concept of the mind-machine debate has been established, Penrose proceeds to meander through a variety of disciplines including fields of mathematics, computational developments, classical physics, quantum physics, cosmology, and neuroscience. In this way, Penrose's first book is more of an exploration of what Gödel's incompleteness theorems may mean for the feasibility of strong AI, as well as how quantum mechanics in brain activity might be able to explain consciousness. His idea is that there is a non-algorithmic element in conscious thought processes. Despite this extra explanation though, Penrose's base argument

[128] Although Penrose uses the Chinese Room Argument as a way to explain the issue he is about to address, I think it might also generate skepticism before even reaching Penrose's separate argument. Searle argues that a machine can model consciousness but cannot actually achieve consciousness. However, how do we quantify our own consciousness? In his contrived experiment, there is meaning already in existence, that of what the Chinese characters mean. So, he is taking a Platonic point of view. Say then that the machine assigns its own meaning to these characters such that the translation rules are satisfied. Would this qualify as consciousness? My inclination is that most people would say no because it is not the correct meaning. Consider then the history of humanity's work in science. We are incessantly constructing meaning that we later discover to be false, but that does not negate our consciousness. So why would it negate the machine's? At the same time, Searle does depend on correct syntax without conscious semantics, so the case of false scientific theories is not necessarily apposite. Still though, solely based on the way in which Penrose introduces the argument, it does not immediately necessitate support for Penrose's later argument.

[129] Penrose, Roger. *The Emperor's New Mind*. Pgs. 17-21.

is essentially a repetition of Lucas's. According to Gödel's first incompleteness theorem, and its relation to computability, Penrose argues that since there will always be a Gödel sentence that is unprovable by the algorithmic methods, but a mathematician can see the truth of it by using the reflection principle and reflecting upon the meaning of the statement, then there must be something non-algorithmic about the mathematician's mind.[130] Another fascinating off-handed remark in the book seems to imply that since algorithmic things are such a limited part of mathematics, whereas the scope of non-algorithmic mathematics is greater, then it is perhaps more likely that the mind contains something non-algorithmic.[131] While the crux of his argument in no way relies upon, or even relates to this remark, it exhibits the meticulous way in which Penrose is able to leverage every bit of exploratory work to his advantage, slowly nudging the reader towards agreement.

When initially setting out to write his book, Penrose had hoped that by the time he reached the end of his writing he would know the scientific process of consciousness, however, he did not. Rather, he tapered off with something he designated as "unbelievable".[132] Specifically, the idea was that wave function collapse was the only possible physical basis for a non-computable process. However, Penrose was not satisfied with the randomness of this, so he proposed a new form of wave function collapse that occurred in isolation. He denoted this "objective reduction."[133]

Although Penrose did not fully formulate a theory as to how consciousness occurs in a non-algorithmic manner, he did catch the interest of others who were fascinated by the topic. One such person was Stuart Hameroff. Hameroff, an anesthesiologist and a professor,

---

[130] Penrose, Roger. *The Emperors New Mind.* Pg. 110.
[131] Ibid. Pg. 98.
[132] Joe Rogan Experience Podcast.
[133] Penrose, Roger. *The Emperors New Mind.*

wrote Penrose to expound upon his theory of objective reduction. Specifically, Hameroff believed that microtubules, tubular polymers that provide some of the shape and structure of eukaryotic cells, would be suitable hosts for the quantum behavior in the brain that Penrose was hoping to discover. Together they produced the theory known as Orchestrated Objective Reduction (Orch-OR) where they postulate that consciousness originates at the quantum level inside of neurons, specifically orchestrated by microtubules. Though Hameroff and Penrose are enthusiastic about the possibilities of the theory, Orch-OR has been met with much criticism from biologists and neuroscientists because microtubules are present in areas other than the brain, such as the liver. However, Penrose and Hameroff argue that they are arranged differently in the brain. Additionally, in Orch-OR the focus is on A-lattice microtubules, that is, the more symmetric ones, rather than B-lattice microtubules. It is this symmetry that Penrose believes is of critical importance because of an effect in quantum mechanics such that a highly symmetrical structure can allow there to be a large gap between the lowest energy level and the next level, thereby meaning that there can be information in the lowest energy level that is shielded from the higher levels.

While Orch-OR technically claims that microtubules are the part of the brain that is subject to quantum processes, Penrose has clarified that although microtubules are one of the best candidates for explaining consciousness, there may be other suitable biological structures. For instance, Penrose suspects that clatherin, a protein that plays a major role in the formation of coated vesicles required for synaptic nerve transmission in the brain, might be important in consciousness, especially as they are highly symmetrical just like A-lattice microtubules.[134] This openness of the theory to other possible biological structures through which quantum processes may cause consciousness makes it a difficult theory to disprove.

[134] Joe Rogan Experience Podcast.

This theory was presented in Penrose's second book about the feasibility of strong AI and the non-algorithmic nature of consciousness.

In his second book, *Shadows of the Mind: A Search for the Missing Science of Consciousness*, published in 1994, again by the Oxford University Press, Penrose clarifies his position further and addresses many of the objections that were raised against his first book. While in the first book Penrose spent more time delving into a history and background of mathematics, his second book is less exploratory, less tentative, and more focused on objections that had previously been raised. He admits as much in his preface, noting also that *Shadows of the Mind* can be read as a stand-alone work. His second book is a more focused and nuanced argument, as well as slightly different from the previous argument put forth by both Penrose and Lucas.

Before fully introducing his argument, Penrose presents the necessary knowledge to the reader, in a manner similar to his previous book, but more focused, including attempts to clarify what his stance on consciousness is, and how that is compatible with his Gödelian argument. Penrose outlines four primary classifications of views on consciousness:

1. Consciousness is just computation
2. Consciousness can be simulated by a machine, but that machine will not have real understanding
3. Consciousness cannot be simulated by a machine, but it does have a scientific explanation, just one that we don't understand yet
4. Consciousness does not have a scientific explanation

Penrose believes the third classification, that a machine cannot simulate consciousness but consciousness does have a scientific explanation. The first part, that a machine cannot simulate consciousness is what he derives from Gödel's first incompleteness theorem. The second part, that consciousness still has a scientific explanation, is what prompts Penrose to delve into quantum physics and neuroscience, in an attempt to ascertain how consciousness arises in such a way that a modern machine cannot model it.

Beyond the significance of consciousness, an emphasis is also placed on the fact that computability is a mathematical idea, formalized by Alan Turing, and independent of any particular computing machine. Or, more specifically, he clarifies that a computation is an action taken by a Turing machine and that computation and algorithm are essentially synonymous.[135] Thus, Penrose is arguing that, with regard to the mathematical definition of computability, there is no way to create a computing machine that can model the mind. Quantum machines are also not sufficient for modeling the mind, as they are susceptible to the same algorithmic issues as our current modern computer.

Eventually, after building up sufficient background knowledge, Penrose formulates his argument against strong AI. It can be roughly summarized as follows:

1. Gödel's incompleteness theorem means that mathematical insight cannot be represented by an algorithm.
2. Mathematical insight is dependent on consciousness, thus consciousness cannot be represented by an algorithm.
3. Then a conscious physical system (human minds) cannot be simulated by a computer.

[135] Penrose, Roger. *Shadows of the Mind.* Pg. 17.

4. There must be a scientific explanation for consciousness though; luckily, modern physics is just strange enough to fit the bill.
5. The interaction between quantum mechanics and general relativity is poorly understood, and in particular, though some things in quantum mechanics are understood, such as entanglement, measurement is not understood, and the collapse of the wave function in quantum mechanics doesn't make sense.
6. Perhaps the brain exploits some large-scale quantum coherence to achieve consciousness, with this effect operating in the cytoskeletons of neurons.[136]

While Penrose does depend on quantum processes to explain consciousness, he does not believe in trying to use quantum ideas directly in human behavior, as he believes such analogies to be far-fetched. For instance, the idea that our minds might be entangled and so could communicate across large distances makes no sense to Penrose or the larger scientific community. The mathematics used in quantum mechanics is very specific to quantum mechanics and so Penrose, as with most physicists, believes it would not scale to something as macroscopic as a human brain.[137,138] Thus, even though his ideas on consciousness may seems outlandish to some, they are still grounded in scientific and mathematic methods, and

[136] Penrose, Roger. *Shadows of the Mind.*

[137] While this is the current stance held by Penrose and others, that there is a way to differentiate between quantum and non-quantum worlds, through quantum physics and classic physics, more recently there has been evidence that this distinction may not be as clear as previously thought. For instance, in 2011 a paper was published describing an experiment in which diamonds released vibrational energy in an entangled system. Having a larger system display quantum tendencies makes dividing the quantum and non-quantum worlds more problematic. "How Big Can Entanglement Get?" PBS. September 24, 2015. Accessed March 13, 2019.

[138] Joe Rogan Experience Podcast.

Penrose distinguishes between feasible applications of quantum mechanics and those that belong solely in science fiction.

Although Penrose has written and spoken extensively on the topic of consciousness and its supposed non-algorithmic nature, he recognizes his own lack of expertise in biology, noting that he developed a theory that was testable but others may be better equipped to further test it. Beyond this, Penrose also never intended to study consciousness much and had intended his first book to be the only work he did on the problem.[139]

One primary objection is that the argument from Gödel's theorem doesn't hold and, therefore, the rest of the argument must be faulty. There are a number of routes that academics have taken to argue against the Gödelian argument, such as there being no reason to assume the AI mathematically infallible, as well as others that were addressed in relation to Lucas's version of the Gödelian argument. However, just because Penrose was inspired by Gödel's theorem to search for a scientific explanation for consciousness does not mean that said exploration is necessarily faulty even if the logic that provoked it is. In fact, Penrose has even suggested dispensing with the logical argument that arises from Gödel's incompleteness theorems and instead to utilize an argument derived from evolution and the highly improbability that an algorithm for seeing the truth of mathematical statements may have evolved from natural selection.[140]

In general, most people take issue with Penrose's usage of Gödel's incompleteness theorems, similar to the same qualms that were discussed in relation to Lucas's Gödelian argument. According to mathematician Martin Davis, Penrose makes an "erroneous claim that Gödel's theorem can be used to show that mathematical insight can not be

---

[139] Ibid.
[140] Aaronson, Scott. "Can computers become conscious?: My reply to Roger Penrose." *Shtetl-Optimized* (blog). 15 June 2016.

algorithmic."[141] One important distinction that Martin Davis makes is with regards to Gödel himself. Davis notes that Penrose insinuates in his books that, because of the implications of the incompleteness theorems, Gödel believed a machine could not model the human mind. However, Davis asserts that this is false as Gödel was extraordinarily careful to avoid controversial statements. [142] This attitude is actually why many people incorrectly believed him to be a logical positivist, because he never spoke up when attending the Vienna Circle, a meeting of primarily logical positivists. Thus, even though Davis says Gödel likely would have agreed with Penrose that mathematical intuition could not be the product of an algorithm, he would not have believed that this was a consequence of the incompleteness theorems.

Another critique of Penrose's argument comes from theoretical computer scientist Scott Aaronson who claims that among computer scientists, software developers, AI and machine learning researchers the answer to whether strong AI is feasible, that is, whether a computer could model the mind such that it truly became conscious, is obviously yes. Most people in these fields also believe that it is highly likely that AI will exceed human capabilities in around thirty years.[143] There have been other studies that indicate people in these fields believe this, although the exact timeframe tends to depend on the region of the world in which one is located.[144]

Scott Aaronson's criticism of the actual arguments that Penrose puts forth include a number of points, but one of the primary issues considers how Penrose grants humans the

---

[141] Davis, Martin. "How Subtle Is Gödel's Theorem? More on Roger Penrose." *Behavioral and Brain Sciences* 16, no. 03. 1993. Pg. 1
[142] Ibid. Pg. 3.
[143] Aaronson, Scott. "Can computers become conscious?"
[144] Ryan, Kevin J. "Elon Musk (and 350 Experts) Predict Exactly When Artificial Intelligence Will Overtake Human Intelligence." Inc.com. June 06, 2017.

capability and freedom to assume the consistency of an underlying formal system, whereas he withholds this from machines. Aaronson points out that unless the human mind can directly perceive the consistency of a formal system, whether that be peering into Platonic heavens or simply knowing, then humans must approach mathematical truth with the same fallible trial-by-error tools used for everything else. But if this is the case then the same liberty ought to be granted to AI. That is, AI should not need to be infallible if humans are not.[145]

On the other hand, some of Aaronson's qualms are simply that he wishes to remain as conservative as possible with regard to quantum mechanics, neuroscience, and biochemistry. However, this is what Penrose thinks it is necessary to avoid. Without any consideration of ideas outside of what is currently known, there is no way to move beyond what is already known. Of course this should be done within reason, that is, the ideas should be testable and not completely unbelievable.

Penrose continues to respond to criticism, such as the above, and to attempt to further his argument through research. A few instances include the following: Penrose debated with Aaronson in Minnesota in 2016, in 2013 Penrose and Hameroff published an updated version of their Orch-OR theory for the purposes of addressing criticism and highlighting new research that supported the theory, and in 2018 Penrose on Joe Rogan's morning talk show, attempting to clarify his position on consciousness, as well as expounding upon all his other work in physics. [146] Another fascinating way that Penrose's colleague, Stuart Hameroff, has begun to respond criticism is by responding directly in the comment section of blogs by

---

[145] Aaronson, Scott. *Quantum Computing since Democritus*. Cambridge: Cambridge University Press, 2015. Pg. 152

[146] Hameroff, Stuart, and Penrose, Roger. "Consciousness in the universe: A review of the 'Orch OR' theory." *Physics of Life Reviews* 11. 2014.

academics that critique the Orch-OR theory. For instance, on Scott Aaronson's and Michael Cerello's blogs Hameroff responds to their criticisms, and on Scott Aaronson's blog they proceeded to have a short conversation in the comment section clarifying their points. This is more unique, modern method of information dissemination.

In the comment section on Aaronson's blog, Hameroff clarifies number of points that Aaronson takes issue with. One particular qualm that he addresses is Aaronson's qualms with the gravitizing quantum mechanics that he claims Penrose does. Hameroff explains,

> Penrose bravely tackles quantum superposition, in which particles exist in multiple locations simultaneously. In general relativity, mass is equivalent to spacetime curvature. Roger suggested superposition of mass in two locations is equivalent to two alternate curvatures, a separation in fundamental spacetime geometry. One might imagine that if such separations were to continue, each such curvature would evolve its own universe, fulfilling the multiple worlds interpretation. However Roger proposed such separations would be unstable, and undergo reduction to a sing state at an objective threshold, held objective reduction, OR. That threshold is the uncertainty principle…OR puts consciousness into reality, avoids multiple worlds, and turns Copenhagen upside down. Rather than consciousness causing collapse, collapse causes consciousness (or is equivalent).[147]

Thus, Penrose's theory of collapse is not just significant for consciousness, but also for the way that it resolves the incapability of general relativity and quantum mechanics. In addition,

[147] Hameroff, Stuart. "Re: Can computers become conscious?: My reply to Roger Penrose." *Shtetl-Optimized* (blog).

it circumvents the need for Copenhagen or Many Worlds, two theories about how to deal with the measurement issue in quantum mechanics.

Of course, then Orch-OR theory goes further looking to microtubules inside brain neurons as the organizers and orchestrators of quantum computations that either halts by objective reduction or product a fully conscious experience with causal power. The arguments by Penrose and Hameroff are extremely technical and nuanced, hence why many people misunderstand them, and why many believe them to be false, as they are making a great number of, as of yet, unproven claims.

The application of the objective reduction reaches beyond consciousness though, as it is essentially a method of reconciling quantum mechanics and general relativity in a deterministic but non-computable theory of fundamental physics. In twentieth century theoretical physics, relativity theory, quantum theory, and chaos theory were the three major developments, however, they have primarily affected distinct domains of impact. Thus, Penrose's integration of relativity and quantum theory is highly unique and potentially influential in the larger theoretical physics domain.[148]

Another couple prominent objections to Orch-OR theory is that the brain is simply too warm, wet, and noisy for quantum behavior and that the collapse of the wave function is much shorter than that of the timescales of neuron firings.[149] However, the more recent discovery of quantum vibrations in microtubules inside brain neurons may support the Orch-OR theory, or at least require more than the claim that the brain is simply to warm and wet to dismiss the theory. Researchers have also found quantum effects to be important for

---

[148] Palmer, T.N. "Lorenz, Gödel and Penrose: New Perspectives on Determinism and Causality in Fundamental Physics." *Contemporary Physics* 55, no. 3. March 2014. Pg. 1.
[149] Sbitnev, Valeriy I. "Quantum Consciousness in Warm, Wet and Noisy Brain." *Modern Physics Letters B* 30, no. 28. July 15, 2016.

particular biological processes, such as photosynthesis. The efficiency with which light is converted into chemical energy to feed the plant is speculated to be possible through the use of the quantum effect of superposition. In this way, while Orch-OR theory remains highly controversial, scientists are coming around to the idea that theories regarding quantum effects ought not to be immediately dismissed, and that microtubules may be a good potential mechanism for explaining anesthesia.[150]

Despite the ongoing debates about consciousness and AI, and the seeming incessant invocation of Penrose's name, this is not what Penrose does primarily. Rather, his main research is in cosmology, general relativity, and black holes.[151]

One of the most influential and critical accomplishments of Penrose's career was his singularity theorem. In 1965, Penrose published the first genuine post-Einstenian result in general relativity, that is, the first modern singularity theorem. Essentially he used algebraic topology[152], specifically the hairy ball theorem[153], to prove that inside any black hole there must be a singularity. A singularity is a point in a place, such as the middle of the Oppenheimer-Snyder Dust Cloud, where the density becomes infinite so the curvature of spacetime becomes infinite, that is, the general relativity equations go to infinity thus they can be shown to fail somewhere. It is a single point in space that contains all the matter that had originally formed the black hole and all the matter that came in after the black hole's

---

[150] Volk, Steve. "Down the Quantum Rabbit Hole." Discover.
[151] Joe Rogan Experience Podcast.

[152] Algebraic topology is a branch of mathematics that studies topological spaces, that is, sets of points and neighborhoods around those points obeying particular sets of axioms, using tools from abstract algebra. One of the most commonly studied objects is a torus, an object that looks like an inner tube.

[153] The hairy ball theorem, also sometimes called the hedgehog theorem, essentially says that it is impossible to comb the hairs continuously and have all the hairs lay flat, instead some hair must be sticking up. More formally, this means that any continuous tangent vector field on the sphere must have a point where the vector is zero.

formation. Regardless of the size of the black hole, Penrose proved that the singularity remains. Others, such as Stephen Hawking, for whom Penrose was a PhD examiner, were fascinated by this work by Penrose and continued with it.

Even though consciousness and singularities are seemingly completely different, there is a connection between them. In a black hole, as the pull of gravity increases, so does the speed necessary to escape the black hole, called the escape velocity. Once a certain escape velocity, the speed of light, is exceeded, it is no longer possible to escape being swallowed into the singularity of the black hole. This point of no return is called the event horizon. What is fascinating is that non-computability is rather uncommon in physics, but black hole event horizons are an example of non-computability, that is, there is not finite algorithm for determining the position in space of a black hole event horizon at some particular time.[154] Thus, even though Penrose's work varies from far off in space to theories about human minds, there are connections, those connections being non-computability and applications of general relativity.

More recently, Penrose has been working on is conformal cyclic cosmology (CCC), also sometimes referred to as sequential aeon theory. This is the theory that the universe iterates through infinite cycles with each Big Bang singularity being the identified with the time-like infinity of the previous iteration. He has been working on this theory, searching for evidence, along with a few other physicists, such as Vahe Gurzadyan. In 2010, Penrose published a book, *Cycles of Time: An Extraordinary New View of the Universe*, in which he expounded upon this theory and generalized it for the interested layperson. Last year, at age

[154] Palmer, T. N. Pg. 4.

87, Penrose jointly published a paper about recently observed Hawking points[155] and how this may support CCC.

Another effort, for which Penrose is recognized, is his book *the Nature of Space and Time* jointly published with Stephen Hawking. It is a record of a series of lectures the two gave alternating so as to address one another's ideas. It is here that Penrose clarifies that he is not in fact a Platonist, but rather, a realist.[156] In this way, he mirrors some of Gödel's philosophical views, as Gödel was known to be a mathematical Platonist, but also a rationalist and a realist. Penrose summed up his position in the debate as follows,

> "At the beginning of this debate Stephen said that he thinks that he is a positivist, whereas I am a Platonist. I am happy with him being a positivist, but I think that the crucial point here is, rather, that I am a realist. Also, if one compares this debate with the famous debate of Bohr and Einstein, some seventy years ago, I should think that Stephen plays the role of Bohr, whereas I play Einstein's role! For Einstein argued that there should exist something like a real world, not necessarily represented by a wave function, whereas Bohr stressed that the wave function doesn't describe a 'real' microworld but only 'knowledge' that is useful for making predictions."[157]

---

[155] These are patches that have been identified within the cosmic microwave background that are significantly hotter than the surrounding area, meaning that these patches could be due to the radiation that is given off during the Hawking evaporation of black holes in the previous aeon. They were denoted Hawking points by Penrose and his colleagues. "New Evidence for Cyclic Universe Claimed by Roger Penrose and Colleagues – Physics World." Physics World. August 28, 2018.

[156] O'Connor, J. J., and E. F. Robertson. "Roger Penrose biography."

[157] Hawking, Stephen, and Roger Penrose. *The Nature of Space and Time*. Princeton: Princeton University Press, 2015. Pgs. 134-135

Penrose distinguishes himself not as a Platonist, but rather as a realist. Generally speaking, there are two primary aspects of realism: existence and independence. The latter means that, for instance, the moon's existence is unaltered, is independent of anything that someone might say about it. The former means that certain things, such as the moon, exist, as do particular facts about them, such as the moon being spherical.[158]

In this way, Penrose and Gödel may share some philosophical views, but they also both made extensive contributions to their scientific communities. Penrose has acquired a variety of awards throughout his career, such as the Albert Einstein Prize and Medal of the Albert Einstein Society, the Naylor Prize of the London Mathematical Society, the Wolf Foundation Prize for Physics, and the Eddington Medal of the Royal Astronomical Society. He was even knighted for his services to science, in 1994.[159]

Beyond his fame in the academic community, Penrose has been popularized for laypeople as the figurehead of the assertion that strong AI is not feasible. For instance, in *Business Insider*, an American financial and business news website with many international editions, published an article entitled *If you think your brain is more than a computer, you must accept this fringe idea in physics*. Within just the title it is evident that Penrose has come to represent the argument for a scientific way to understand the brain while still denying strong AI. He also has continued to speak at symposiums and on radio shows; all while doing his own physics research on other topics. Although Penrose may be popularized as representing the argument against strong AI, he is not the only one to advocate for the non-computability of consciousness. There have been other theories, such as integrated information theory, that attempt to explain consciousness in such a way that it cannot be

[158] Miller, Alexander. "Realism." Stanford Encyclopedia of Philosophy. October 02, 2014.
[159] O'Connor, J. J.

modeled by a machine, as well as entire of fields of study dedicated to this question. However, Penrose is the primary advocate of the Gödelian argument against strong AI and his Orch-OR theory has received much press because of its entirely new physics ideas.

Penrose has greatly contributed to the debate regarding the feasibility of artificial intelligence, and continues to do so today. He has popularized it, made complex mathematical theorems accessible to the public, and generally represented the belief that a machine cannot model the mind. In understanding these arguments and how they arose, it also becomes vital to recognize the influence and interaction of theological and philosophical beliefs. Not only can these beliefs influence how one develops an argument, but also, the side that one takes in the strong AI debate can have theoretical theological repercussions, which might actually be relevant within our lifetime.

## Chapter 3: Philosophical and Theological Influences and Implications

*"To explain everything is impossible: not realizing this fact produces inhibition."*
*– Kurt Gödel*

As previously discussed, Gödel's incompleteness theorems were integrated into the debates regarding the feasibility of strong AI through the arguments of John Lucas and Roger Penrose. However, this integration also meant that these mathematical theorems were pulled into the realm of philosophical discussion about the human mind, a vastly complex realm to say the least. It is worth exploring some of the philosophical and theological connections that have arisen due to this association. Specifically, this chapter explores Gödel's and Penrose's personal philosophical and theological beliefs and how these might provide insight into their application of the first incompleteness theorem to strong AI, it further expounds upon the vast debates about consciousness to which Penrose connected the incompleteness theorems, and it briefly investigates theoretical questions of how strong AI could have theological repercussions, such as addressing the seemingly facetious question whether an AI could be baptized.

According to Hao Wang, a philosopher and mathematician who wrote about and compiled resources about Gödel's philosophical and theological thoughts, Gödel's philosophical theory was a monadology Monadology is from Leibnitz's philosophy and can be understood, as simply as possible, as the doctrine of monads as the ultimate units of being. A monad is a substance without parts that makes up larger compounds. So, the idea is that the monads are the most basic substances that comprise the world. This monadology that

comprised Gödel's philosophical theory, was rationalist, idealistic, optimistic, theological, and contained a central monad, namely, God.[160,161]

Following this idea of a central monad in a theologically influenced monadology leads to two key theological arguments generated by Gödel. Gödel privately offered arguments for God's existence and an afterlife, but these arguments were separated, as he considered the issue of an afterlife apart from the problem of the existence of God.[162] For the afterlife issue, Gödel wrote four long letters to his mother in 1961 delineating his views, all of which have been translated into English by Yi-Ming Wang and are contained in Hao Wang's book *A Logical Journey*.[163] In these he writes that he is "convinced of this [the afterlife], independently of any theology."[164] This conclusion was reached by reasoning that "If the world is rationally constructed and has meaning, then there must be such a thing [as an afterlife]."[165] His argument is as follows: science demonstrates that order prevails in the world, this order indicates that the world has meaning, human beings in the world realize so little of their potentialities that these potentialities seem to be a meaningless waste, but this would contradict the idea that the world has meaning. Thus, in order to circumvent this contradiction, Gödel concluded that there must be an afterlife. In addition, Gödel suggested that his belief in an afterlife might prevail in future, and even compared it to the way in which atomic theory was originally proposed as a philosophical argument.[166] Like many of his theological musings, Gödel's ideas regarding an afterlife were primarily focused on what

---

[160] Although these all have formal philosophical definitions, they are not particularly important here, as the intuitive definitions from everyday life suffice, with the exception of monadology, which was just explained.

[161] Wang, Hao. *A Logical Journey: from Gödel to Philosophy*. Cambridge, Mass: MIT Press, 1996. Pg. 8.
[162] Ibid. Pg. 102.
[163] Ibid. Pg. 105.
[164] Ibid. Pg. 104.
[165] Ibid. Pg. 105.
[166] Ibid. Pg. 105.

he believed could be logically argued through pure reason, rather than any construction or imagination as to how such an afterlife might manifest in actuality.

This conclusion about the existence of an afterlife also connects to Gödel's theological worldview, "the concept that the world and everything in it has meaning and sense, and in particular a good and unambiguous meaning."[167] It would only be rational for everything to have meaning. The world having meaning is part of Gödel's argument for an afterlife and for his belief in a God-like being, but the two are independent arguments. If everything has meaning then, Gödel argues, "It follows directly that our presence on Earth, because it has of itself at most a very uncertain meaning, can only be the means to the end for another existence."[168] In this way Gödel asserts the existence of a being outside of humanity and he furthers this assertion in a version of St. Anselm's ontological proof of God's existence. In Gödel's proof, his conclusion is that there exists a "God-like" individual, where "God-like" means that every essential property is positive and every positive property is essential. Such a definition distances Gödel's proof from more evangelical ones, and Gödel also clarified that the proof had been undertaken as a logical investigation, rather than a theologically motivated one, and it remained unpublished as he was concerned that "a belief in God might be ascribed to him."[169]

Thus, Gödel utilized rationality, derived from science, and the idea that everything in the world has meaning, derived from the scientific principle that everything has a cause, to argue for an afterlife and for the existence of a God-like being. However, despite these indications of theological beliefs, Gödel never ascribed to a particular religion, only

---

[167] Ibid. Pg. 108.
[168] Ibid.
[169] Franzén, Torkel. Pg. 91.

describing himself as holding theistic beliefs and having been baptized Lutheran.[170] It is also worth noting that Gödel never attempted to invoke the incompleteness theorems as a means to draw theological conclusions, though many others have done so.[171]

One final consideration for Gödel's theological beliefs is his semester at Notre Dame in 1939. It seems apt to consider the possibility that such a strongly Catholic environment might have affected Gödel's own beliefs in some way. However, this seems unlikely due the short amount of time spent at Notre Dame, and the way in which his primary relationship with Notre Dame, through Professor Menger, deteriorated, although it did recover somewhat in the later years of Gödel's life. In addition, Gödel never returned to the University of Notre Dame. He only ever taught and worked at Princeton and the Institute for Advanced Study once he moved to the U.S. for good. So, in that way, Notre Dame is unique, as it was the only other American university at which he taught, even if only very briefly.

With regards to Gödel's philosophical beliefs, although his theological thoughts could also be classified as such, he believed that reflections on mathematics could provide the best method to develop and understand philosophical thought.[172] This seems quite fitting for a mathematical logician. Gödel recommended in mathematics the process of deepening knowledge of abstract concepts by using phenomenology – a "technique that should bring forth in us a new state of consciousness in which we see distinctly the basic concepts."[173] Thus he argues that by focusing on the study of consciousness and the objects of direct experience (phenomenology), it is possible to further our knowledge of abstract concepts, bringing our consciousness to a state where we can understand the most basic concepts. In

[170] Wang, Hao. *A Logical Journey*. Pg. 112.
[171] Ibid.
[172] Ibid. Pg. 287.
[173] Ibid. Pg. 157.

this context Gödel's idea of consciousness can seemingly be equated with understanding or comprehension, as he believes that deepening knowledge brings about a new state of consciousness. The use of phenomenology is specifically from a philosopher named Edmund Husserl, and Gödel derived much of his personal philosophy from the philosophical beliefs of Husserl.[174] Gödel even saw his own incompleteness theorems as evidence in support of Husserl's belief that humans utilize categorical intuitions, a concept in which Husserl extends the idea of intuition beyond its normal limitations of sense perception.[175] For Gödel, categorical intuitions were important as he believed in the feasibility of human ability to see universal connections, and Husserl's categorical intuitions represents that. This is part of Gödel's emphasis on the abstract and the universal as the center of philosophy; he also encourages many generalizations from smaller to broader concepts.[176]

Another way in which mathematics can be seen to relate to Gödel's philosophical beliefs is through his idea that assumptions must be made and intuition is necessary, as in mathematics. He writes, "the purpose of philosophy is not to prove everything from nothing but to assume as given all – including conceptual relations – that we see as clearly as shapes and colors, which come from sensations but cannot be *derived* from sensations. The positivists attempt to prove everything from nothing."[177] In this way, Gödel argued that the positivists refused to make any assumptions and instead attempted to derive basic ideas, such as shapes and colors, from sensations, something that he believed was not possible. Not only does Gödel articulate the necessity of assumptions, similar to the axioms required in order to

[174] Edmund Husserl was a German philosopher best known for establishing the school of phenomenology, that is, the philosophical study of the structures of experience and consciousness. He was born in 1859 and lived until 1938. Although he is famous as a philosopher, he also studied mathematics under Karl Weierstrass and Leo Königsberger.

[175] Ibid. Pg. 156.
[176] Ibid. Pg. 329
[177] Ibid. Pg. 173.

formulate a mathematical proof, and again emphasizes the need to utilize phenomenology, but he also sets himself up in opposition to the positivists. As previously mentioned, Gödel spent some time in the Vienna Circle, a group primarily comprised of positivists, and as he never spoke in contradiction to their beliefs, he was assumed a positivist by many. However, Gödel was assuredly not a positivist, and as evinced here, wanted others to know that. Another example of the intersection of Gödel's philosophical beliefs and his work in mathematics is his article "What is Cantor's Continuum Problem?" originally published in 1947 in *The American Mathematical Monthly*.[178] In this article, he articulated his metaphysical Platonist beliefs. Later, the article was edited by Benacerraf and Hilary Putnam and republished in 1964. However, Gödel was extremely anxious about publishing the aforementioned article, as he believed Benacerraf and Putnam to be positivists, and as such, thought that they would use their introduction to attack his ideas.[179]

Gödel also believed in the power of mind over matter. His rationalistic optimism is an optimism about the power of human reason.[180] This accords with his brief assertion in a 1951 lecture that either the human mind infinitely surpasses the powers of any finite machine or there exist unsolvable mathematical problems.[181] Said assertion, as explained in the previous chapter, was leveraged by Lucas to argue that Gödel believed his incompleteness theorems indicated the impossibility of a machine modeling the mind.[182]

One final philosophical idea of Gödel's that is worth investigating is time, which he saw as subjective.[183] Gödel believed in seeing objective reality, physical and conceptual, as

---

[178] Godel, Kurt. "What Is Cantor's Continuum Problem?" *The American Mathematical Monthly* 54, no. 9, 1947. Pgs. 515-525.
[179] Goldstein, Rebecca. Pgs. 216-218.
[180] Wang, Hao. *A Logical Journey*. Pgs. 288 and 317.
[181] Gödel, Kurt. *Collected Works.* Pg. 310.
[182] Megill, Jason
[183] Wang, Hao. *A Logical Journey.* Pg. 313.

eternal, timeless, and fixed. For the physical world specifically, he clarified that there is a natural coordinate system from the dimensions of reality, but that for the mind there is no natural coordinate system. The only natural frame of inference is time. Thus, Gödel held a belief in Kantian time. So while objective reality is timeless, Gödel also believed that our internal consciousness of time is an essential ingredient of our experience because it is the only natural frame of inference for the mind.[184] Between 1946 and 1950, Gödel wrote several articles on this concept of time. This period was also when he found new solutions for Einstein's field equations of general relativity by using a cosmological term that had a negative value rather than being equal to zero. Gödel presented the solutions to Einstein for his birthday, though the latter was actually rather disturbed by the mathematically sound solution as it allowed for the existence of closed time-like curves and therefore time travel. Gödel's purpose was to leverage the solutions to argue that our intuitive concept of time is not objective.[185] In this way, Gödel's personal philosophy and belief in Kantian time clearly influenced his mathematical work, as he found mathematically sound solutions to Einstein's gravitational equations. Thus through this particular example of Kantian time, as well as his conclusion that reflections on mathematics could provide the best method to develop and understand philosophical thought, it is evident that there was a dual philosophical and mathematical nature to some of his work.

As mentioned earlier, the vast majority of research on Gödel's philosophical and theological beliefs was conducted by academic Hao Wang and compiled into several books. Wang also corresponded with Gödel and met with him a number of times to discuss his beliefs. As a logician, philosopher, and mathematician, Wang had a number of contributions

[184] Ibid. Pgs. 287 and 322.
[185] Ibid. Pg. 319.

beyond his expansion of the collection of knowledge about Gödel. One of his most famous contributions is the Wang tile. These tiles are colored squares that tile the plane but only aperiodically. Not only was this the first noted example of an aperiodic tiling,[186] but also he showed that any Turing machine could be turned into a set of Wang tiles. Penrose is also known for similar tilings. A Penrose tiling is an example of non-periodic tiling generated by an aperiodic set of prototiles, where a prototile is one of the shapes of a tile in a tessellation, that is, a cover of a space by closed shapes. In this way, Wang and Penrose are connected through both their mathematical work on aperiodic tiling, as well as their fascination with, and connections to, Gödel.

In seeming contradiction to his later statement in *The Nature of Space and Time*, Penrose writes that he is a mathematic Platonist in *The Emperor's New Mind*.[187] However, this Platonism is distinguished from the general Platonist views that Gödel adhered to in his earlier life. In particular, Penrose uses his book to expound upon the three primary types of philosophy of mathematics: formalism, intuitionism, and Platonism. Then he notes that in this context he is a Platonist. Thus, his realist and mathematical Platonist views are

[186] An aperiodic tiling is a non-periodic tiling (for a counter-example, the sine wave is periodic, although it is not a tile, the same idea of "periodic"ness applies). There is also the additional property for an aperiodic tiling that is does not contain arbitrarily large periodic patches. Quasicrystals are a real life example of aperiodic tiling, specifically, Penrose tilings are the useful model for quasicrystals as the tilings and quasicrystals both exhibit reflection symmetry and fivefold rotational symmetry. Initially, Dan Schetman, who discovered quasicrystals in 1982, was ridiculed and told he had made a mistake because the atoms of crystals were known to form a periodic arrangement, but the model of aperiodic tilings from the mathematical community help to eventually convince scientists that quasicrystals were real. Schetman later received the Nobel Prize in 2011.

[187] "At the beginning of this debate Stephen said that he thinks that he is a positivist, whereas I am a Platonist. I am happy with him being a positivist, but I think that the crucial point here is, rather, that I am a realist. Also, if one compares this debate with the famous debate of Bohr and Einstein, some seventy years ago, I should think that Stephen plays the role of Bohr, whereas I play Einstein's role! For Einstein argued that there should exist something like a real world, not necessarily represented by a wave function, whereas Bohr stressed that the wave function doesn't describe a 'real' microworld but only 'knowledge' that is useful for making predictions." Hawking, Stephen and Roger Penrose.

compatible, as it is only in the context of mathematics that he is a Platonist, and philosophically it is accepted practice to be selectively realist or non-realist.[188]

In his writing, Penrose explains why he adheres to mathematical Platonism, and how he believes the other two philosophies cannot encompass the fullness of mathematics. For instance, formalism, the idea proposed by Hilbert, is unable to express the truth of the Gödel sentence of a particular formal system. Thus, the formalist's notion of truth is incomplete. However, there are ways to circumvent this supposed issue, such as solely referring to the idea of provability instead of the concept of truth. Or, a more pragmatic approach would be to recognize that Gödel sentences are statements that rarely, if ever, occur in serious mathematics, thus could be ignored as irrelevant.[189] On the other hand, intuitionism was started as a response to paradoxes in a way that is distinct from formalism. Intuitionism is the view that sets do not have existence, as some Platonists may claim, but rather are considered in terms of all the rules that determine their membership. This philosophy was supposed to mirror human thought, hence its name, intuitionism. According to intuitionism, existence really means constructed existence, that is, a definite construction must be presented regarding a mathematical object before said object is accepted as existing.[190]

Within mathematical Platonism there are a number of different distinctions that can be drawn, but Penrose opts to ignore such distinctions, embracing the philosophy as a whole. Different versions all would encapsulate the primary theses of mathematical Platonism though, that is, that mathematical objects exist, that mathematical objects are abstract, and that mathematical objects are independent of all intelligent agents and their language,

---

[188] Miller, Alexander. "Realism." Stanford Encyclopedia of Philosophy. October 02, 2014.
[189] Penrose, Roger. *The Emperor's New Mind*. Pg. 108.
[190] Ibid. Pgs. 113-114.

thought, and practices.[191] Additionally, Penrose takes a weaker Platonist view, rather than the stronger view that Gödel takes, meaning that Gödel believed that the truth or falseness of mathematical statements, even ones concerning enormous, nebulously constructed sets, is always an absolute Platonic matter. So, a strong Platonist, such as Gödel, would believe that even though some sets have incredibly dubious definitions, they still have a God-given quality of either truth or falseness, that is, they are not constructed ideas by humans, but rather part of the existing mathematical truth. On the other hand, a weaker Platonist might look at such sets on a sliding scale, recognizing that while some mathematics has this God-like truth quality, some is convoluted and begins to take on a matter of opinion quality.[192] To clarify though, this God-like truth quality does not mean there is a God according to a Platonist. In fact, while Gödel had his ontological proof, Penrose is a devout atheist and takes great issue whenever his theory about the impossibility of a machine modeling the mind, due to quantum activity in the brain causing consciousness, is leveraged to inappropriately support religious ends.

Returning to the incompleteness theorems and the way in which Penrose relates them to debates about the feasibility of strong AI, the crux of Penrose and Hameroff's Orch-OR theory is that, due to the first incompleteness theorem, there is something unique about the human mind, denoted consciousness, that is likely manifested through quantum activity in microtubules. It is this concept of consciousness that is highly important.

The prevalence of the word "consciousness" can be seen in magazines, self-help books, philosophical conversations, and everything between. It is a term that has permeated our culture, yet still remains rather enigmatic. Looking solely at Catholic journals,

---

[191] Linnebo, Øystein. "Platonism in the Philosophy of Mathematics." Stanford Encyclopedia of Philosophy. January 18, 2018.
[192] Penrose, Roger. *The Emperor's New Mind*. Pg. 112.

newspapers, and magazines, the plethora of articles that use the word consciousness, whether discussing how to be a better Christian, scientific discoveries and what they mean for people of faith, or any other context, is astounding. For some context, *Religious News* has 893 free articles (there are more that are not free and have been archived) from 2003 onward that include mention of consciousness, and *The Tablet* has 64 from the past two years alone. Penrose also seems to use consciousness in a way that is almost interchangeable with understanding. Among cognitive scientists, the existence of consciousness is deemed a "hard problem" whereas other "easy" problems are primarily information processing, such as driving a car, and are therefore mere computation. To clarify, consciousness is not a hard problem in the sense of NP-hard problems in mathematics,[193] but rather there is a qualitative difference between cognitive problems that can be easily solved through computations and those that have not yet been fully answered. Hameroff, who works in this cognitive science field, obviously adheres to his Orch-OR theory regarding consciousness, but also has freely speculated about the philosophical implications of his and Penrose's theory. One such conjecture is that near-death experiences might reflect a potentially short-lived quantum

[193] NP problems are those that have an efficient algorithm to verify the problem's solution while P problems are those that have an efficient algorithm to solve the problem. In this case, efficient means solvable in polynomial-time. A NP-hard problem is one that is at least as hard as the hardest problems in NP, meaning NP-complete problems. NP-complete problems are those that are at least as difficult to solve as any other NP problem. In computer science and mathematics there is lots of emphasis on the question of whether P = NP. Moreover, this is one of the seven Millennium Prize Problems that carries a million dollar prize for the first correct solution. It is clear that all P problems are NP problems, as it is easy to check that a solution is correct by solving the problem and comparing the two solutions. So if there is P problem, then it is solvable in polynomial time, which means it can be verified by solving it again and comparing the solutions, also in polynomial time. However, there is no clear way to prove or disprove if all NP problems are P problems. Gödel is actually connected to these problems as well, as in 1956 he wrote a letter to mathematician John von Neumann asking whether a certain NP-complete problem could be solved in quadratic or linear time. It wasn't until 1971 when Stephen Cook formally articulated the precise statement of the P versus NP problem, but Gödel did pose a version of it with regard to a specific problem.

afterlife.[194] On the other hand, Penrose has steered clear of postulating any philosophical implications of Orch-OR theory as a scientific explanation for consciousness.

Another fascinating way to grasp consciousness, according to cognitive psychologist Steven Pinker, is through empathy, as empathy is essentially consciousness simulating consciousness, thus part of consciousness is understanding and feeling. [195] This idea of consciousness as understanding or comprehension is similar to what Penrose indicates in the prologue of *the Emperor's New Mind* and to what Gödel seemed to indicate when discussing the deepening of knowledge, claiming that deepening knowledge of abstract concepts by using phenomenology should bring about a new state of consciousness in which one can fully understand the basic concepts. According to Anil Seth, professor of Cognitive and Computational Neuroscience, intelligence and consciousness are not the same, and that is why some people, such as Seth, think that strong AI is not possible.[196] This distinction between intelligence and consciousness is of particular interest, since other experts, such as theoretical computer scientist Scott Aaronson, do not seem to make such a distinction. Aaronson summarizes Penrose's Orch-OR theory as theorizing that objective collapse plays a role in human intelligence, when Penrose and Hameroff specify that this collapse plays a role in consciousness. While this may seem minute, labeling consciousness as the non-algorithmic element in the mind, rather than claiming that intelligence contains a non-algorithmic element, seems more similar to distinctions that neuroscientists, such as Anil Seth, may be making. Despite its ambiguous definition, consciousness is commonly referenced, and indicates a cultural phenomenon of focus on self-awareness. Modeling this

---

[194] Volk, Steve.
[195] Pinker, Steven. *How the Mind Works*. London: Penguin Books, 2015. Pgs. 18-19.
[196] Seth, Anil. "Your brain hallucinates your conscious reality," filmed in April 2017. TED video, 17:01.

self-awareness is what many, including Penrose, take issue with. Returning to Penrose's four primary classifications on consciousness,[197] it is evident that each has particular deadweight associated with it, and adhering to a particular theological or philosophical belief system could hinder one's ability to take up whichever seems most reasonable, instead boxing one in to a particular view.

Marvin Minsky, a founding father of artificial intelligence, and the man whose BBC interview prompted Penrose to write his *The Emperor's New Mind*, argued that consciousness was a word lacking of scientific rigor that ought to be replaced by a number of more specific concepts, such as reflection and decisions. In fact, Minsky believed that the study of consciousness would eventually be viewed as a waste of time. However, the scientific study of consciousness has stuck, and Penrose and Hameroff's theory has continued to be discussed, though people often criticize it. The theory itself makes it difficult to draw lines between scientific and philosophical dimensions of thinking. For instance, some people have questioned what the Orch-OR theory of consciousness could mean for the long-standing philosophical argument over free will and determinism. The indeterminacy that is intrinsic to quantum theory seems to suggest the breakdown of causal connections within the brain, thereby insinuating that perhaps the deterministic viewpoint also would break down. Penrose notes that while it may appear that he is making a case for free will, that is actually not the case.[198] He says, "It does look like these choices would be random. But free will, is

---

[197] The classifications are as follows:

1. Consciousness is just computation
2. Consciousness can be simulated by a machine, but that machine will not have real understanding
3. Consciousness cannot be simulated by a machine, but it does have a scientific explanation, just one that we don't understand yet
4. Consciousness does not have a scientific explanation

[198] Paulson, Steve. "Roger Penrose On Why Consciousness Does Not Compute - Issue 47: Consciousness." *Nautilus*. May 04, 2017.

that random?"[199] In this way, relating philosophical questions and Penrose's theory of consciousness is quite possible, thus expanding the connections and relevance of such a theory even further.

Penrose's personal beliefs also provide him with the motivation to continue studying consciousness. When asked about if he thought there was any inherent meaning in the universe, Penrose replied, "Somehow, our consciousness is the reason the universe is here."[200] This statement is rather reminiscent of Gödel's theological beliefs, but as a reverse. Gödel believed that everything in the world must have meaning since scientific principles indicate that everything has a cause. And if everything has meaning, then, since the meaning of humanity is quite uncertain, our existence must be the means to the end of another existence. In this way, Gödel argued for the existence of a God-like being.[201] However, Penrose seems to be saying that, rather than claiming our existence to have uncertain meaning, it is our uniqueness, our consciousness, that gives meaning to the universe.

Expanding, Penrose explains, "I'm not so sure our own universe is favorably disposed toward consciousness. You could imagine a universe with a lot more consciousness that's peppered all over the place. Why aren't we in one of those rather than this one where it seems to be a rather uncommon activity?"[202] But since we are in a universe where consciousness seems to be uncommon, Penrose has his motivation for continuing to study consciousness, that is, the question of what the purpose of consciousness is.

The issue of consciousness is also a topic that theoretical computer scientist Scott Aaronson believes greatly influences how people develop a personal conclusion regarding

---

[199] Ibid.
[200] Ibid.
[201] Wang, Hao. *A Logical Journey*. Pg. 108.
[202] Paulson, Steve.

the strong AI question. According to Aaronson, many people have an intuitive issue with the concept of granting consciousness to a machine or robot because of two key assumptions. First, people have the directly experienced certainty that they're conscious. Second, they hold the belief that if they were just a computation, then they could not be conscious in this way. Aaronson combines these two assumptions, leading them to their logical conclusion. He writes, "For people who think this way (as even I do, in certain moods), granting consciousness to a robot seems strangely equivalent to *denying that one is conscious oneself*."[203] It is fascinating that Aaronson also admits to, at times, believing this equivalency between accepting a machine's consciousness and denying one's own consciousness, particularly as he believes these assumptions and conclusion to be the primary reason that Penrose vehemently denies the feasibility of strong AI. Despite his interactions with Penrose, it is a stretch for Aaronson to claim that he has such special insight into Penrose's reasoning, or people's reasoning in general.

One simple way out of this difficult-to-swallow equivalency, that granting consciousness to a robot is seemingly equal to denying one's own consciousness, is by remembering that we regard other people as conscious even though we do not understand how a bundle of neurons could be conscious. Yet, even though we understand very little about neuron activity and brain function, we still accept the consciousness of other people, just as the consciousness of a machine could be accepted, if it could emulate humans in all respects, that is, pass a Turing test.

The connection between the two assumptions noted by Aaronson and philosophy and theology are evident, that understanding the personal theological and philosophical ideas at

[203] Aaronson, Scott. *Quantum Computing*. Pg. 41.

stake in the feasibility of AI debate can explain how many people derive their conclusion and decide which side to take.

For example, in Catholic newspaper *Crux Now*, David Chiang, a professor of computer science and engineering at Notre Dame, remarked, "As a Catholic I don't believe that so-called artificial intelligence will ever be intelligent. It's really an article of faith for me (rather) than a well-worked out philosophical position."[204] Even though there is a connection between Chiang's personal theological and philosophical ideas informing his stance on the feasibility of strong AI, Chiang recognizes as much. Also, it does not seem as though having theological beliefs necessitates a disbelief in the possibility of strong AI. As a counter example, co-founder of *Wired* magazine, Kevin Kelly, is an advocate for the creation of a catechism for robots because he believes that strong AI is both possible and imminent, and would like the Catholic Church to prepare for such a situation.[205]

Moving beyond the way in which personal beliefs can inform one's belief or disbelief in the feasibility of strong AI, it is fascinating to consider theoretical theological questions that could arise if there were an AI equivalent to the human mind. One question is the interaction of strong AI with religion. What would it mean, for religion, is an AI existed that had every intelligence and consciousness attributed to humans? Levi Checketts, professor of religious studies and philosophy at Holy Names University in Berkeley, wrote his doctoral dissertation on applying Catholic moral theology to new technologies. Checketts questioned, "Can AI be baptized?"[206] He argues, "the Christian must be ready to understand herself in

---

[204] Jenkins, Jack. "The (holy) Ghost in the Machine: Catholic Thinkers Tackle the Ethics of Artificial Intelligence." *Crux*. May 26, 2018.
[205] Merritt, Jonathan. "Is AI a Threat to Christianity? Are You There God? It's I, Robot." *The Atlantic.* February 11, 2017.
[206] Jenkins, Jack.

new terms if a consciousness is 'successfully' uploaded."[207] While he does not say that strong AI will definitely exist in the future, he insinuates that it is highly plausible, and that the Christian faith must adapt to a possible post-human future.

Catholic nun, Sister Ilia Delio, who teaches theology at Villanova University, explains that this question is essentially about discovering the meaning of theology in an unfinished universe. She remarks, "What is the meaning of salvation and redemption in an unfinished universe? That's what technology builds on – an unfinished universe that we help finish."[208] Thus, according to Delio, religion has to adapt and modify in order to integrate new technological advancements, such as strong AI. In a mass in 2014 Pope Francis suggested that if a Martian asked to be baptized then the Catholic Church would baptize that Martian, but then, if an intelligent robot asked seemingly the answer ought to be the same. One Presbyterian pastor from Florida has garnered recognition for his assertion that God could redeem a conscious AI. He explains, "I don't see Christ's redemption limited to human beings. If AI is autonomous, then we should encourage it to participate in Christ's redemptive purposes in the world."[209]

These questions then also intersect with questions of the soul that are often discussed in religion. How does one differentiate between consciousness and having a soul, or is there a difference? Could an AI have a soul, and is this the same as asking if an AU could be conscious? Although these thought-problems may seem far-fetched, technology for genetic cloning and in vitro fertilization are already common, but it seems safe to assume that most

[207] Checketts, Levi. 2017. "New Technologies—Old Anthropologies?" *Religions* 8, no. 4: 52. Pg. 8.
[208] Jenkins, Jack.
[209] Ibid.

Christians would agree that those humans still have a soul.[210] Thus, in some minor ways, technology has already begun to infringe upon these distinctions of consciousness and soul.

Thus, there are a number of theological and philosophical questions that arise from the concept of strong AI, but, returning to Gödel and Penrose, if one leverages the incompleteness theorems to argue that the human mind cannot be modeled by a machine, then follows Penrose's line of thought that there must be a unique scientific process because of this, and that the unique scientific process gives rise to consciousness, then seemingly these theological problems hold less significance as the AI would still be missing this consciousness. Still though, these questions will become more prevalent as technology advance, and recognizing the connections between philosophy, theology, computer science, and mathematics will become more pertinent. Whether that be philosophical beliefs informing mathematical pursuits, as evinced by Gödel, or leveraging mathematical conclusions to derive philosophical ideas, as Penrose did with the incompleteness theorems and AI, it is important to recognize these relationships between fields.

[210] Merritt, Jonathan.

## Conclusion

*"The more I think about language, the more it amazes me that people ever understand each other." – Kurt Gödel*

Despite the limitations of language, people still strive to understand one another. Part of this understanding involves connecting ideas and tracing them through history, in order to fully realize the origins and derivation of another person's stance or belief. That is what has been attempted here. This thesis is about understanding some of the relevant arguments of the time, how they have arisen, why they are worth studying, and how they intertwine with other fields, that is, the interweaving of many ideas into a cohesive picture. Gödel's theorems have permeated academia beyond mathematical logic, and helped demonstrate the interconnectedness of the mysteries and problems of this world. Having investigated the way in which Gödel's incompleteness theorems were integrated into the debate surrounding strong artificial intelligence, as well as how this debate now pertains to personal theological and philosophical beliefs, I have attempted to bring about an understanding of the rich intellectual history of Gödel's incompleteness theorems and their continued resonance in philosophical questions.

In order to achieve this understanding, I first explained the theorems and the main AI arguments that have leveraged the theorems to make their points. Once this context had been established, I moved onto the historical background in which the theorems were developed and the ways in which, over time, the theorems permeated the field of mathematical logic as well as eventually the fields of computing and AI, specifically focusing on physicist Roger Penrose's argument against the feasibility of strong AI and the relationship between Gödel's theological leanings and his stance on the incompleteness theorems' relevance to AI. Finally, I argued that taking a stance in the feasibility of AI debate is not only necessary, but has

theological implications in terms of aligning theology with modern science and answering theoretical questions about how the emergence of strong AI could affect our views of what it means to be a Christian, and investigate the way in which Gödel's theological and philosophical beliefs influenced his work. With a goal of comprehension of the theorems, as well as their connection to AI, philosophy, and even theology, I sought to trace this intellectual history beginning with the life of Gödel and ending close to the present day. In addition, recognizing this interconnected nature can help to bring about greater understanding of whichever aspect one finds most compelling.

As the field of artificial intelligence grew and developed from its theoretical roots to levels of complex computation, philosophy of AI grew as well. Neuroscientists, philosophers, mathematicians, computer scientists, and physicists, among others, populate this field of philosophy. Through this growth, there has been an integration of all these different fields, such that the arguments for and against the feasibility of strong AI have become more complex. As a general matter though, most of the AI community has accepted the idea that humans are inconsistent, meaning that the human mind cannot be represented by a formal system since there may be times when contradictions arise and remain unresolved. Due to this supposed inconsistency Gödel's results are not applicable to the debate, despite assertions by Penrose and Lucas to the contrary. In this way, computationalism cannot be proven false by Gödel's incompleteness results as it is claimed that the results are consistent with the computational theory of mind.[211,212] Hilary Putnam, famously known for the Putnam

[211] As mentioned in the introduction, there is a difference between computationalism, mechanism, and the feasibility of strong AI. Since I am addressing the feasibility of strong AI debates, from a historical perspective, I have glazed over some of the differences between these three ideas, as generally speaking arguments for and against one or the other are applicable for all the ideas. The definitions are as follows.
Computationalism: the theory that minds are information processing systems where consciousness and cognition are just forms of computation

mathematics exam that occurs each year in December, first put forth this computational theory of mind, in its modern form, in 1961.[213] Lucas even addressed Putnam in his reply to Benacerraf's paper *Satan Stultified*.[214] In addition, there is the idea that in reality machines and humans have limited resources and time, so employing Gödel's incompleteness theorems, which apply to what can theoretically be proven with infinite energy, memory, and time, is incorrect and that neither humans nor machines need to be able to prove everything in order to be intelligent or equal to one another.[215]

From a theological perspective, this understanding of the theorems, the AI debates, and the various conceptual definitions can help prevent any particular belief from pigeonholing the construction of the rest of one's beliefs. As Scott Aaronson notes, many people have an intuitive issue with the concept of granting consciousness to a machine or robot because of two key assumptions. First, people have the directly experienced certainty that they're conscious. Second, they hold the belief that if they were just a computation, then they could not be conscious in this way. Aaronson combines these two assumptions, leading them to their logical conclusion. He writes, "For people who think this way (as even I do, in certain moods), granting consciousness to a robot seems strangely equivalent to *denying that one is conscious oneself*."[216] By recognizing the interconnectedness of all these topics

---

Mechanism: phenomena are solely determined by mechanical principles, therefore they can be adequately explained by certain mechanical principles alone
Strong AI: an artificial intelligence construct that has mental capabilities and functions that mimic the human brain, including intelligence and consciousness (this then applies for feasibility of strong AI debates, that is, the possibility of what has just been defined)

[212] LaForte, G., Hayes, P. J., Ford, K. M. "Why Gödel's theorem cannot refute computationalism." *Artificial Intelligence* 104. 1998. Pgs. 265-286.
[213] Putnam, Hilary. "Brains and Behavior." 1961.
[214] Lucas, J. R. "Satan Stultified."
[215] Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Upper Saddle River: Pearson, 2016. Pg. 950.
[216] Aaronson, Scott. *Quantum Computing*. Pg. 41.

though, and by attempting to achieve a relative conceptual understanding of each, such an immediate felt issue can be circumvented and reconsidered from more theoretical standpoints.

The strong AI debates are particularly compelling because they are increasing in relevance as the theoretical possibility of a machine that could model the human mind inches closer to reality, and thus singularity, the hypothesis that the invention of artificial superintelligence will abruptly trigger runaway technological growth and profoundly change human civilization.[217] In fact, North American experts believe that in the next seventy years AI will outperform humans in all capacities, whereas experts from Asia believe this will occur in only thirty years.[218] This does not necessarily mean that these machines will be conscious as that depends more on whether one equates consciousness with intelligence or

[217] A brief aside, in my own personal exploration of these arguments I have found myself enticed by Penrose with regard to his ideas that there may be something non-algorithmic within the human mind and that there may be concepts such as quantum gravity that are currently beyond our comprehension. However, I disagree with the primary argument that Gödel's first incompleteness theorem necessitates this non-algorithmic design, and that this means it is impossible to model the mind, that is, create strong AI. Essentially, I am saying since I am neither a PhD in quantum physics nor neuroscience, I do not consider myself qualified to dismiss his arguments in that regard, particularly as they are written in a highly convincing format and his credentials lend even more credibility. But, coming from a mathematics background and having studied the arguments of experts on the subject, I do disagree with his application of Gödel's theorem. As Scott Aaronson asserted, Gödel's first incompleteness theorem seems almost thrown in after Penrose had already come to the conclusion that strong AI is not possible. Returning to Lucas's application of Gödel's theorem (as this is the entirety of Lucas's argument), I primarily take issue with the assumption of the consistency of humans (there are other issues I have but this is my primary concern). Lucas addresses one counterargument by Benacerraf regarding this problem by essentially saying that humans choose consistency. What if they do not though? In making this assumption of consistency or that humans choose consistency, it seems to me that we enter into the realm of determinism, that is, the lack of free will. Of course if things were deterministic then wouldn't it be possible to model them on a Turing machine? It certainly seems so. Then let us assume things are not deterministic. One of the primary assumptions that anti-free will arguments often make is that non-determinism implies randomness, but since randomness is just a distribution of probabilities, there is still no free will. Aaronson's response is that this is a false assumption. Computers run on algorithms but those algorithms do not know in advance the inputs they will be given, else they could hard code the outputs. In this way, free will is the decision of inputs. From this idea then we seem to return to the idea that humans run on algorithms. In essence, we have made a circular path returning to the idea that it is possible to model the mind. Consider then if we are consistent but despite this, we say we can see, using our collective ability to generate mathematical insight, the truth of our own Gödel sentence, whereas the computer cannot. But we do not know that the computer cannot since it does not run on truth but on probability. So we need to give it the ability to make meta-arguments, else it won't be a good model of us in the first place. But with the ability to make meta-arguments it will be able to prove its own Gödel statement, hence the incompleteness theorems would not apply.

[218] Ryan, Kevin J.

not. In addition, strong AI is, as Steven Pinker might say, a problem now rather than a mystery. Meaning it has moved from something at which "we can only stare in wonder and bewilderment, not knowing what an explanation would even look like" to a problem where "we may not know its solution, but we have insight, increasing knowledge, and an inkling of what we are looking for."[219] This growth of understanding is added to by all the different fields and approaches to understanding the mind, as well as how humans might mathematically relate to machines.

From a more personal perspective, I was drawn into this topic by George Zarkadakis's book *In Our Own Image*, which is representative of a larger cache of philosophy and history of AI books. In addition, the interconnectedness of many different fields, such as history, mathematics, philosophy, computer science, and theology, was addressed in *In Our Own Image*, and again emphasizes that often paths to understanding wind through a variety of interconnected fields.

Although such an intellectual history of the incompleteness theorems and their connection to so many other fields is worth writing, in this process comes the construction of a particular narrative-defined method of research and writing. Particularly with a topic so broad as strong AI, the way in which it is approached helps to define how readers formulate their own beliefs. On this front, I have done my utmost to remain impartial to the philosophical leanings of the major players, Kurt Gödel and Roger Penrose, while still critically analyzing the way in which their work has been influenced. Inevitably, I have failed to some extent. In addition, due to the breadth of the topic, a significant amount of critical

[219] Pinker, Steven. Pg. ix.

information has been ignored, thereby constructing a narrative in which the primary voices are only those of Gödel and Penrose.

There is also the question of why this was done from a historical perspective rather than computer science or mathematics or philosophy or theology. In essence, it is more feasible to focus on the broader picture, that is, the historical interactions between multitudes of fields and integrate them together, than to finitely focus on one field and have to ignore the relationships with other fields. How one develops an understanding of one field can inform one's understanding of another, thus it is important to cultivate this awareness of connection. For instance, philosophical views can push one towards particular scientific viewpoints and scientific viewpoints can in turn be extrapolated to have philosophical significance. An example of the former is the way in which Penrose explains why he cannot agree with Searle that a machine could simulate consciousness but that machine will not have real understanding. Penrose asserts that everything must have an explanation, therefore consciousness cannot be simulated by a machine but it does have a scientific explanation, thus he embarks on his quest to discover the scientific explanation. Gödel on the other hand, provides an example of the opposite type of implication. He extrapolated from the scientific principle that everything has a cause, to conclude that everything must have meaning, and therefore since there is no evident meaning to the existence of humanity, humans must be the means to the end for another existence, that is, a God-like figure. Intellectual history is about this understanding and connection. It is about comprehension and questioning, rather than refuting any specific viewpoint.

In summation, this intellectual history wove together the way in which Gödel's incompleteness theorems have connected artificial intelligence, physics, neuroscience, philosophy, and theology throughout the twentieth and twenty-first centuries.

Bibliography

**Primary Sources**

Benacerraf, Paul. "God, the Devil, and Gödel." *Monist* 51, no. 1. 1967. Pgs. 9-32. doi:10.5840/monist196751112.

Checketts, Levi. 2017. "New Technologies—Old Anthropologies?" *Religions* 8, no. 4: 52.

Gödel, Kurt. *Collected Works*. Edited by Solomon Feferman. Oxford: Oxford University Press, 1995. Pg. 310.

Gödel, Kurt - Math, UDIS 105/29. University of Notre Dame Archives. Notre Dame, IN.

Gödel, Kurt. "On Formally Undecidable Propositions Of Principia Mathematica And Related Systems." *Philosophical Books* 4, no. 1. 1963. Pgs. 17-18. doi:10.1111/j.1468-0149.1963.tb00774.x.

Godel, Kurt. "What Is Cantor's Continuum Problem?" *The American Mathematical Monthly* 54, no. 9. 1947. Pgs. 515-25. doi:10.2307/2304666.

Jenkins, Jack. "The (holy) Ghost in the Machine: Catholic Thinkers Tackle the Ethics of Artificial Intelligence." *Crux*. May 26, 2018. https://cruxnow.com/church/2018/05/26/the-holy-ghost-in-the-machine-catholic-thinkers-tackle-the-ethics-of-artificial-intelligence/.

Lucas, J. R. "Mechanism: A Rejoinder." *Philosophy* 45, no. 172. 1970.

Lucas, J.R. "Minds, Machines and Godel." 1961. Web.

Lucas, J. R. "Satan Stultified." *Monist* 52, no. 1. 1968. Pgs. 145-58. doi:10.5840/monist196852111.

"Many Faculty Changes," *The Notre Dame Alumnus*, October 1937, University of Notre Dame Archives, http://archives.nd.edu/Alumnus/VOL_0016/VOL_0016_ISSUE_0001.pdf.

"Many New Teachers Join Faculty." *The Notre Dame Scholastic*, 24 September 1937, University of Notre Dame Archives, http://archives.nd.edu/Scholastic/VOL_0071/VOL_0071_ISSUE_0001.pdf.

Merritt, Jonathan. "Is AI a Threat to Christianity? Are You There God? It's I, Robot." *The Atlantic.* February 11, 2017. https://www.theatlantic.com/technology/archive/2017/02/artificial-intelligence-christianity/515463/.

Penrose, Roger. "Gödel, Relativity, and Mind." *Journal of Physics: Conference Series* 82. 2007. doi:10.1088/1742-6596/82/1/012002.

Penrose, Roger. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. London: Vintage, 2005.

Penrose, Roger. *The Emperors New Mind: Concerning Computers, Minds and the Laws of Physics*. Oxford: Oxford University Press, 2016.

Penrose, Roger. *What is Intelligence?: Mathematical intelligence.* Editor Jean Khalfa. Cambridge University Press.1994. Chapter 5.

Wang, Hao. *A Logical Journey from Gödel to Philosophy*. Cambridge, Mass: MIT Press, 1996.

**Secondary Sources**

Aaronson, Scott. “Can computers become conscious?: My reply to Roger Penrose.” *Shtetl-Optimized* (blog). 15 June 2016. https://www.scottaaronson.com/blog/?p=2756

Aaronson, Scott. “Could a Quantum Computer Have a Subjective Experience?” *Shtetl-Optimized* (blog). 25 August 2014. https://www.scottaaronson.com/blog/?p=1951.

Aaronson, Scott. *Quantum Computing since Democritus*. Cambridge: Cambridge University Press, 2015.

Copeland, Jack B. "The Church-Turing Thesis", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.).

Davis, Martin. "How Subtle Is Gödels Theorem? More on Roger Penrose." *Behavioral and Brain Sciences*16, no. 03. 1993. Pg. 611. doi:10.1017/s0140525x00031915.

Dawson, John W, Jr. “Logic at Notre Dame: Kurt Godel at Notre Dame.” *Notre Dame Journal of Formal Logic*. https://math.nd.edu/assets/13975/logicatndweb.pdf

Dawson, John W. *Logical Dilemmas: The Life and Work of Kurt Gödel*. Wellesley, MA: K Peters, 1997.

De Mol, Liesbeth. "Turing Machines", *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.).

Dodd, T. "Gödel, Penrose and the Possibility of AI." *Artificial Intelligence Review*5, no. 3. 1991. Pgs. 187-99. doi:10.1007/bf00143761.

Franzén, Torkel. *Gödels Theorem: An Incomplete Guide to Its Use and Abuse*. Wellesley, MA: K Peters, 2005.

Goldstein, Rebecca. *Incompleteness: The Proof and Paradox of Kurt Gödel*. New York: W.W. Norton, 2006.

Hadley, Robert F. "Consistency, Turing Computability and Gödel's First Incompleteness Theorem." *Minds and Machines*18, no. 1. 2007. Pgs. 1-15. doi:10.1007/s11023-007-9082-2.

Hameroff, Stuart. "Re: Can computers become conscious?: My reply to Roger Penrose." *Shtetl-Optimized* (blog).

Hameroff, Stuart, and Penrose, Roger. "Consciousness in the universe: A review of the 'Orch OR' theory." *Physics of Life Reviews* 11. 2014.

Hawking, Stephen, and Roger Penrose. *The Nature of Space and Time*. Princeton: Princeton University Press, 2015.

Hodges, Andrew. "In Retrospect: Gödels Proof." *Nature* 454, no. 7206. 2008. doi:10.1038/454829a.

Joe Rogan Experience Podcast. # 1216, "Sir Roger Penrose". Aired December 18, 2018.

Jongeneel, C.j.b, and H. Koppelaar. "Gödel Pro and Contra AI: Dismissal of the Case." *Engineering Applications of Artificial Intelligence*12, no. 5. 1999. Pgs. 655-59. doi:10.1016/s0952-1976(99)00024-x.

LaForte, G., Hayes, P. J., Ford, K. M. "Why Gödel's theorem cannot refute computationalism." *Artificial Intelligence* 104. 1998. Pgs. 265-286.

Letzter, Rafi. "If You Think Your Brain Is More than a Computer, You Must Accept This Fringe Idea in Physics." June 10, 2016. https://www.businessinsider.com/penrose-says-your-brain-isnt-a-computer-2016-6.

Linnebo, Oystein. "Platonism in the Philosophy of Mathematics." Stanford Encyclopedia of Philosophy. January 18, 2018. https://plato.stanford.edu/entries/platonism-mathematics/.

Megill, Jason. "The Lucas-Penrose Argument about Gödel's Theorem." *Internet Encyclopedia of Philosphy*. https://www.iep.utm.edu/lp-argue/#SH2b

Miller, Alexander. "Realism." Stanford Encyclopedia of Philosophy. October 02, 2014. https://plato.stanford.edu/entries/realism/.

Nagel, Ernest, and James R. Newman. *Gödel's Proof*. N.Y.: New York University Press, 1958.

"New Evidence for Cyclic Universe Claimed by Roger Penrose and Colleagues – Physics World." Physics World. August 28, 2018. https://physicsworld.com/a/new-evidence-for-cyclic-universe-claimed-by-roger-penrose-and-colleagues/.

Palmer, T. N. "Lorenz, Gödel and Penrose: New Perspectives on Determinism and Causality in Fundamental Physics." *Contemporary Physics* 55, no. 3. March 2014. doi:10.1080/00107514.2014.908624.

Paulson, Steve. "Roger Penrose On Why Consciousness Does Not Compute - Issue 47: Consciousness." *Nautilus*. May 04, 2017. http://nautil.us/issue/47/consciousness/roger-penrose-on-why-consciousness-does-not-compute.

Pinker, Steven. *How the Mind Works*. London: Penguin Books, 2015.

Putnam, Hilary. "Brains and Behavior." 1961.

O'Connor, J. J., and E. F. Robertson. "Roger Penrose biography."

Raatikainen, Panu. "Gödel's Incompleteness Theorems." Stanford Encyclopedia of Philosophy. January 20, 2015. Accessed March 21, 2019. https://plato.stanford.edu/entries/goedel-incompleteness/.

Reichenbach M., Cohen R.S. (1978) The Königsberg Conference on the Epistemology of the Exact Sciences [1930f]. In: Reichenbach M., Cohen R.S. (eds) Hans Reichenbach Selected Writings 1909–1953. Vienna Circle Collection, vol 4a. Springer, Dordrecht.

Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Upper Saddle River: Pearson, 2016.

Ryan, Kevin J. "Elon Musk (and 350 Experts) Predict Exactly When Artificial Intelligence Will Overtake Human Intelligence." Inc.com. June 06, 2017. https://www.inc.com/kevin-j-ryan/elon-musk-and-350-experts-revealed-when-ai-will-overtake-humans.html.

Sabinasz, Daniel. "Gödel's Incompleteness Theorem And Its Implications For Artificial Intelligence." Deep Ideas. September 12, 2017. http://www.deepideas.net/godels-incompleteness-theorem-and-its-implications-for-artificial-intelligence/.

Sbitnev, Valeriy I. "Quantum Consciousness in Warm, Wet and Noisy Brain." *Modern Physics Letters B* 30, no. 28. July 15, 2016. doi:10.1142/s0217984916503292.

Seth, Anil. “Your brain hallucinates your conscious reality,” filmed in April 2017. TED video, 17:01, https://www.ted.com/talks/anil_seth_how_your_brain_hallucinates_your_conscious_reality?language=en.

Smith, Peter. *An Introduction to Gödel's Theorems*. Cambridge: Cambridge University Press, 2014.

Volk, Steve. "Down the Quantum Rabbit Hole." Discover. http://discovermagazine.com/bonus/quantum.

Wang, Hao. *Reflections on Kurt Gödel*. Cambridge, Mass. MIT, 1987.

Zach, Richard. "Hilbert's Program." Stanford Encyclopedia of Philosophy. January 06, 2015. https://plato.stanford.edu/entries/hilbert-program/.

Zarkadakēs, George. *In Our Own Image: Savior or Destroyer?: The History and Future of Artificial Intelligence*. New York, NY: Pegasus Books LLC, 2016.