Artificial General Intelligence and the Mind-Body Problem: Exploring the Computability of Simulated Human Intelligence in Light of the Immaterial Mind

Caleb Parks

A Senior Thesis submitted in partial fulfillment
of the requirements for graduation
in the Honors Program
Liberty University
Spring 2024

Acceptance of Senior Honors Thesis

This Senior Honors Thesis is accepted in partial
fulfillment of the requirements for graduation from the
Honors Program of Liberty University.

Daniel O'Malley, Ph.D.
Thesis Chair

Edward N. Martin, Ph.D.
Committee Member

James H. Nutter, D.A.
Honors Director

April 29th, 2024

Abstract

In this thesis I explore whether achieving artificial general intelligence (AGI) through simulating the human brain is theoretically possible. Because of the scientific community's predominantly physicalist outlook on the mind-body problem, AGI research may be limited by erroneous foundational presuppositions. Arguments from linguistics and mathematics demonstrate that the human intellect is partially immaterial, opening the door for novel analysis of the mind's simulability. I categorize mind-body problem philosophies in a manner relevant to computer science based upon state transitions, and determine their ramifications on mind-simulation. Finally, I demonstrate how classical architectures cannot resolve so-called *Gödel statements*, discuss why this inability is inherent to all formal axiomatic systems, and review arguments derived from this observation about the computability of human intelligence.

Keywords: Artificial General Intelligence, Mind-Body problem, Gödel Statements, Gödel's Incompleteness Theorems, Interactionism, Occasionalist Quantum Idealism, Machine Learning, Natural Language Processing

Artificial General Intelligence and the Mind-Body Problem: Exploring the Computability of Simulated Human Intelligence in Light of the Immaterial Mind

In the past year, breakthroughs in machine learning (ML) powered generative artificial intelligence (AI) and natural language processing (NLP) models such as ChatGPT, DALLE, and Stable Diffusion have seized the public eye (Heilweil, 2023). As the rumbles of a technological revolution some consider on par with the printing press foreshadow the disruption of a wide variety of industries, the same questions echo through the garages of startups and the boardrooms of tech giants alike: what are AI's limits? Will the search for artificial general intelligence, or AGI, be successful? After all, all contemporary AI are Artificial Narrow Intelligences (or "ANI"), meaning that they can only solve a narrow and specific range of tasks (such as playing chess or generating images)—in contrast with Artificial General Intelligence (or "AGI"), which is a theoretical type of AI distinguished by the ability to solve a wide variety of problems (including problems it never encountered during its training), much like how humans' intelligence can adapt to new challenges. In short, an AGI would possess all the same cognitive facilities and capacities as humans, and could adapt to do any task that we can (Nancholas, 2023). The search for such intelligence is difficult and has been the subject of considerable research and resources for over half a century (Council of Europe, n.d.). However, the efficacy of AGI research may be constrained by the omission of pertinent insights from the field of philosophy regarding the mind-body problem.

Brain Simulation as a Possible Route to AGI

There may be many possible routes to producing Artificial General Intelligence, as intelligence itself is not that thoroughly understood. However, one of perhaps the simplest and most straightforward theoretical paths to producing AGI is by simply simulating the human brain. If human intelligence is fully encompassed within a physically-closed, deterministic system within the brain, then it stands to reason that—by building an artificial brain that is functionally identical to the human brain—we should be able to produce a system on par with human intelligence.

This statement has three crucial caveats. First, it is important to clarify what I mean by "functionally identical to the human brain." In short, understanding *why* a certain neuron signals its leftmost neighbor is not necessary; simply knowing that it does and replicating that in the artificial system is sufficient. Additionally, the artificial system need not use actual biological cells to represent this connection as long as the functional behavior is preserved.

The second caveat is that practical concerns with the implementation of such a system will be ignored for the purposes of this inquiry. To provide brief examples, current neuroscience and biology are not even close to understanding the brain well enough to even functionally duplicate it, and the mere state space of the brain is large enough to provide an enormous engineering challenge (Riken, 2013). However, as this is a theoretical inquiry, not an engineering one, the concerns of practical implementation of brain mapping and simulation will not be addressed here. Instead, in this paper I will deal with the question of whether artificial brain simulation is theoretically possible, examining potential *inherent* barriers to such an endeavor.

The final caveat, and the focus of the majority of this thesis, lies in a crucial phrase from the proposal above: "*If* human intelligence is fully encompassed within a physically-closed, deterministic system within the brain…" This is a crucial condition, because the simulability of human intelligence rests on this assumption. If our intelligence runs entirely within a physical brain that is fully deterministic, then there should be no inherent theoretical roadblocks to simulating it. However, if—as I'll seek to prove in the following section—the human intellect is not purely physical, or if the brain is not deterministic, then the theoretical feasibility of simulating human intelligence becomes significantly more complicated.

This complexity is precisely what I will seek to make sense of in the rest of this thesis, but first I must prove that human intelligence extends beyond the physical. This task, which will be the primary focus of the following section, will require taking on one of the prevailing philosophical views of the Western world: physicalism.

Refuting Physicalism

In the present era, Physicalism stands as the predominant ontological worldview among the scientific community (Hodge & Patterson, 2015; Neurath, 1983; Quester, 2022). Physicalism is a monistic philosophical view which holds a materialist ontology (Menuge, 2023a). Ontology is the branch of philosophy dealing with metaphysical questions about the nature of being: what is really real, what types of substances exist, etc. Physicalism holds a monistic, materialist ontology because it posits that only one type of substance exists—the physical—and that, existentially, nothing is more extensive than its physical properties (Stoljar, 2022). To date, the majority of scientific research within the field of artificial intelligence, machine learning, and natural language processing is conducted, written, and reviewed by scientific professionals with predominantly physicalist worldviews (Miłkowski, 2014). In other words, most AI and ML research rests on a presupposition of the commonly held (at least among the contemporary scientific community) belief that the physical world—matter and energy, physics, forces, and fields—is all there is (Hodge & Patterson, 2015; Miłkowski, 2014; Neurath, 1983; Quester, 2022).

However, as I will argue at length in this section, these physicalist assumptions are wrong. A review of not only philosophy but also the sciences of linguistics, math, logic, and computation will all reveal that there are numerous problems with physicalism and provide a strong case for the existence of the immaterial.

Issues with Physicalism

One of the first issues that arises with physicalism pertains to the belief that materialism holds about the causes of things. Materialism posits that both the "material cause" (what something is made of) and the "efficient cause" (what brought something into being) of a thing must both be physical (Menuge, 2023a). In fact, physicalism must hold to these propositions, because it denies the very existence of anything beyond the physical, and something that doesn't exist certainly cannot cause something that does.

Furthermore, physicalism must also deny the existence of what philosophy calls "final causes:" the purpose for which something has happened. However, these beliefs run into a variety of issues. To begin with, nothing we observe in the physical world indicates that any physical state of a system can have intentionality (i.e., that the state can be about something else). However, thoughts can be about something, such as the Bass Pro Shop Pyramid in Tennessee. Furthermore, a thing's physical powers are limited by space, time (future events cannot cause a physical effect now), and causality (physical interactions between items require them to both actually exist in the same universe). However, I can think of the Bass Pro Pyramid without physically being anywhere near Tennessee, or I can think about a future event, or about a non-existent entity (such as the character Killmonger from Marvel), despite the fact that none of these things should be able to physically interact with my brain (Menuge, 2023b).

In short, though science tells us that the physical causal powers of something are limited by time and space, a mental subject can "cause" a thought in my brain even if it is miles away, in the future, or fictional.

Identity Theory

In light of these issues, many modern physicalists have conceded that mental states must be understood as actual states in their own right, but argue that mental states are really just physical states of the subject's brain. This view—called *identity theory*—has its own issues, foremost of which is that it violates the plausible principle of the indiscernibility of identicals. The issue here is essentially that mental and physical states have categorically different properties, as a mental state can be about another thing or have propositional content, while matter has neither of these. Consequently, mental and physical states cannot be identical if subjectivity and intentionality are intrinsic to the former but not to the latter (Menuge, 2023a).

Linguistic and Logico-Mathematical Arguments for the Immaterial

Having addressed some of the issues with physicalism, I now turn to an argument from linguistics that the human intellect is composed of more than just the physical brain, following the reasoning presented by Dr. Baumgardner and Dr. Lyon in their work *A Linguistic Argument for God's Existence,* published in the 58th issue of the *Journal of the Evangelical Theological Society*.

Argument from Linguistics. In any discussion of language, several observations must be made. First, *languages are comprised of vocabularies, which map meaning to otherwise* *meaningless symbols*. These symbols are usually words or letters and have both spoken and written forms in most human languages. Furthermore, these symbols need not have any inherent meaning of their own. For example, the shapes of the letters in the word "dog" bear no inherent correlation to the common house pet: rather the meaning of the word derives from the English vocabulary, which has encoded a certain range of meaning to the symbol "dog." Other languages encode the same range of meaning in different symbols, such as "perro" in Spanish or "chien" in French. Furthermore, the set of symbols and combinations thereof within a language may vary significantly between regions or even generations, as evidenced by American slang terms popular with Generation Alpha such as "fanum tax," "rizz," or "no kizzy." For their users, such terms bear the same meaning as other English lexical constructions (such as "charisma," "stealing a friend's food," or "not lying"), while bearing no meaning at all for most members of older generations. The reality that different languages encode the same range of meaning with wildly different symbols evidences the fact that the spoken sound (or written shape) of the word bears no inherent meaning of its own (Baumgardner & Lyon, 2015).

The second observation is that *languages are composed of grammars, which specify rules for combining symbols from the vocabulary together to form more complex meaning-bearing messages.* English grammar allows a speaker to formulate the sentence "the dog is hungry for meat," which bears a complex meaning which cannot be expressed with any one English word. Noam Chomsky (2006)—one of the most influential minds in modern computer science and linguistics—has written that all languages are rule systems assigning sound and meaning in a definite way for an infinite class of possible phrases, and thus partially consists of a pairing of sound and meaning over an infinite domain.

This concept begs the question: what is the essence of these messages—not the essence of the mediums they are expressed through, but the essence of the messages themselves? Are they fundamentally matter, energy, or something else? Meaning is by definition an abstract (and thus immaterial) entity. Furthermore—like the logico-mathematical objects I will discuss later—linguistic rules are also abstract entities, and thus immaterial. Thus, language conveys immaterial meaning and is structured by immaterial rules. It deals not only with physical concepts like dogs but also with immaterial concepts such as goals, political rights, or existentialism. Furthermore, language's message can remain unaltered as it is translated through a series of drastically different physical mediums (Baumgardner & Lyon, 2015). For example, a text message may be typed on a smartphone, stored in ASCII, divided into network packets, transmitted via wireless signals, sent through a complex network of fiber optic cables, copper wires, and many other media, and then displayed on another smartphone's screen, where a human eye translates light rays from the screen into electrical signals which travel up the optical nerve and are then processed into impulses sent to the vocal cords to produce sound waves in speaking the words read to another listener. Language's transcendence over the physical media it is expressed through thus further supports its immaterial nature (Baumgardner & Lyon, 2015). This is the third and most important observation—an answer to the aforementioned question of essence: *both language and the meaning it conveys are inherently immaterial*.

In recognition of this distinction between the material world of mass and energy and the immaterial world of meaning, concepts, and language, linguists refer to the separation between these two worlds as the "Einstein gulf" (Oller et al., 2014). This term derives its name from Albert Einstein who, in reference to the immaterial nature of language and meaning, pointed out that humans habitually connect certain concepts and sensory experiences so confidently that we often remain unaware of the logically unbridgeable gulf separating the worlds of sensory experiences and of concepts (Baumgardner & Lyon, 2015; Einstein & Scholz, 1953).

The observation that language is inherently immaterial supports the conclusion that humans as language users must possess significant immaterial attributes and faculties (Baumgardner & Lyon, 2015). Noam Chomsky (2006) himself writes that human language users have the capacity to comprehend an infinite array of unfamiliar expressions that differ significantly from their linguistic repertoire, lacking direct physical resemblance or simple analogies. Moreover, we can generate such expressions when necessary, regardless of their novelty. Thus, humans' normal use of language involves a form of creativity that distinguishes our communication from any observed animal communication system.

Because these linguistic capacities are unique to humans, Chomsky (2006) believed that the study of language approaches the very essence of humanity: qualities of intellect associated with the use of language that are unique to humans and cannot be separated from any critical phase of human existence. It is because material causes are incapable of generating immaterial effects that Drs. Baumgardner and Lyon (2015) conclude that human language use provides compelling evidence supporting the conclusion that our mental language generation and processing faculties must be likewise immaterial.

By this reasoning, the existence, nature, and use of language by itself adequately refutes physicalism, but I will shortly demonstrate through a discussion of logico-mathematical objects that linguistics is not the only field to do so. However, I must first briefly address an objection to this linguistic argument, which may have already occurred to the observant reader.

Searle's Argument Refuting the Immateriality of AI. During the presentation of the linguistic argument presented above, the observant reader may have considered the application of this reasoning to modern NLP algorithms, which produce language extensively and quite well. One may object to the linguistic argument by reasoning that, if one argues that human intellects' use of language proves our partial immateriality, then shouldn't an NLP model's use of language also prove the model's partial immateriality? However, since it is fairly obvious that our NLP models are purely material, shouldn't one conclude that humans are too? In other words: if humans' use of language proves we have immaterial minds, shouldn't an AI's use of language similarly prove that AI has an immaterial mind? If not, perhaps human intelligence really is just our physical brains and nothing more?

In response, it is paramount to note that specificity in what is meant by "use language" is crucial. By peeling back only a few layers of abstraction, it is evident that the way humans use language is considerably different from the way contemporary NLP models use language. Modern NLP models fundamentally are text-predicting systems. AI NLP models use algorithms that identify patterns in training data which they can then generalize to make predictions about new data (Nadkarni et al., 2011). These models are often called "generative" machine learning models because they can generate synthetic data (Nadkarni et al., 2011). One example of such a model is GPT-3 (Generative Pre-trained Transformer 3), the NLP AI which Chat-GPT is built upon. GPT-3 generates human-like text following an initial input, or prompt, and was trained on an unlabeled dataset of text from the internet (Floridi, 2020). In short, GPT-3 examines a prompt and then predicts and outputs what words it expects to follow the prompt, based on the data it was trained on.

Humans differ from this process significantly in our use of language. The human intellect takes a conscious understanding of information and uses language to express that meaning to others. Though a human can predict the next word in another's speech, their use of language extends far beyond this faculty alone. An examination of other predictive problems may prove helpful here. Take the example of heuristics for individuals attempting to predict the next move in a game of chess. One heuristic would be to provide the individual with a list of the game's rules which they may use to determine all available moves for a given board state, and then provide them with a list of the values of different types of pieces and instruct them to evaluate the available moves based upon the resulting relative gain or loss of piece points. Another fundamentally different heuristic would be to provide the individual with recordings of thousands of games played by grandmasters and instruct them to predict the next move based upon which move occurred most frequently in the recorded games at the same board state. Under the latter heuristic, the individual is never provided with the actual rules of the game, and is thus incapable of logically analyzing the board state. However, because grandmasters in the recorded games know the rules, a statistical pattern will emerge from their games that is in some way representative of the rules. This pattern may be accurately used to predict moves, often more accurately than the first heuristic's limited analysis via piece points. However, it would be incorrect to assume from their accurate predictions that the individual in this second heuristic understands the game, as they are merely following a statistical pattern. Likewise, it would be incorrect to assume that contemporary NLP models understand language or meaning just because the results of their text predictions closely imitate human conversation.

On this issue, John Searle, author of the famous Chinese Room Argument, has provided an argument that mere computation (i.e., symbol manipulation) is not sufficient for true understanding. He concludes that while a computing machine (such as a modern digital computer) can process symbolic inscriptions, such a system will never have human-like understanding of the meaning of those symbolic inscriptions. Searle reaches this conclusion because when considering human beings who are merely performing computer-like symbol manipulation (such as the individual in the Chinese Room Argument), one sees that no understanding is conveyed whatsoever. Searle argues that, if a computing machine could achieve understanding by its mere computing, then this would not be the case (Bringsjord et al., 2023b). Thus, by modus tollens, Searle argues that computers do not understand by mere symbol manipulation.

In summary, even though many AI models appear to understand meaning and use language like humans do, they in fact do not. Furthermore, Searle has argued that mere symbol manipulators cannot possibly achieve such understanding. Consequently, AI models' use of language does not stand as evidence that they are in any part immaterial or have "minds" in the philosophical sense.

Logico-Mathematical. Beyond natural languages, there are many other subjects which humans interact with on a regular basis that must be immaterial. Two of the most significant fit within the category of logico-mathematical objects: algorithms and inference schemata. Inference schemata are especially important because they form the foundation of rigorous reasoning in the formal sciences, since these fields are driven by theorems obtained via proofs composed of propositions linked by inferences (Bringsjord et al., 2023b). In the rest of this section, I will first demonstrate that logico-mathematical objects such as algorithms and inference schemata must be immaterial, and I will then argue why this implies that humans (in part) are too.

First, a few examples of logico-mathematical objects may be in order. As mentioned above, logico-mathematical objects include both algorithms and inference schemata (which specify the types of inferences that are valid in logical proofs). One case of an inference schema is *modus tollens*, which specifies that the series of inferences *if P then Q, not Q, therefore not P* is a valid argument. Modus tollens can be used to prove that all inference schemata immaterial, because when a person presents an argument that validly instantiates modus tollens, the validity does not derive from some relation between the person and some physical embodiment of modus tollens, and the same holds true for algorithms (Bringsjord et al., 2023b).

In other words, the validity of a particular logical inference, or the validity of a particular implementation of a given algorithm, cannot be based upon their physical implementations. Why? Because there is an arbitrarily large range of possible physical implementations, with no *physical* traits categorically unifying the set of valid implementations as opposed to the invalid ones. Consequently, the validity of these implementations of logico-mathematical objects cannot be physical, and therefore must be immaterial.

The fact that the validity of these objects is an immaterial issue is relevant to this discussion because humans understand these objects—more precisely, we can understand that we are validly implementing them. As a brief example, even elementary school children perform (and understand that they are performing) valid implementations when they learn how to do multiplication or division of large numbers by hand (both of which are done using algorithms).

Conclusions on Physicalism

Let me review what I've covered thus far. First, even though physicalism is the dominant worldview in modern scientific communities, it has a slew of problems, including the fact that material causes cannot explain the causal powers of mental subjects, which transcend space, time, and even reality. Second, the field of linguistics shows that the human capacity to both use and understand language in communicating abstract ideas necessitates that some part of our intelligence is immaterial. A similar argument also arises from the logico-mathematical objects which form the foundation of the formal sciences. Since there are an infinite range of possible physical implementations of a logico-mathematical object, the validity of any implementation cannot be based on the physical, and must thus be immaterial. As humans can know and understand that we are validly implementing a logico-mathematical object, this proves that we are immaterial too. Finally, utilizing a proof by contradiction, Searle's Chinese Room Experiment proves that mere computation (symbol manipulation) is not sufficient for true understanding, implying that any being that truly understands must be more than just physical computation.

All of these arguments strongly support the conclusion that the human intellect cannot be purely material. John Searle (2005) himself has written that there is no other area of contemporary analytic philosophy where such an implausible view (that human intellect is pure material) is so widely held, and concludes that many advanced thinkers on the subject routinely deny the obvious facts that A) humans do have subjective conscious mental states, and B) these states aren't eliminable in favor of anything else.

Thus—despite the denial of much of the contemporary scientific community—there is clear and compelling evidence that humans' intellect is made of more than just our physical brains.

Defining Terms and Clarifying the Question

The above conclusion that humans must have some non-physical aspect to our intellect: some immaterial faculties, raises another question, one that has been debated for centuries in the field of philosophy: the *mind-body problem*. However, before I can discuss this topic, it will be helpful to clarify some important terms and background information.

Defining Terms

As is regular in philosophy, using agreed-upon definitions is crucial. I will now provide definitions of some relevant terms used throughout the remainder of this work.

Human Intelligence / Intellect. I will use the term *human intelligence* to refer to the sum total of all of the elements that comprise our ability to think, understand, communicate, and reason. I will also use the term *human intellect* to refer to the seat of such human intelligence (though for any purpose other than extreme philosophical lexiconical exactness these terms of *human intelligence* and *human intellect* are virtually interchangeable).

The Mind and Brain. The human intellect contains both a physical component—the organ of the human brain—and (as demonstrated in the prior section) an immaterial component. I (like most philosophers within the context of the mind-body problem) will use the term *mind* to refer to this immaterial component, or more technically: the entity, substance, or phenomena responsible for the immaterial faculties of the human intellect. Two such faculties are *cognition* (awareness in general and the ability to learn in particular) and *intelligence* (the ability to obtain and use knowledge in an adaptive situation). It is important to note here that the question of whether the mind is an entity, a substance, or merely a phenomenon is itself part of the debate, but I will discuss this later.

The Mind-Body Problem. For now, I have covered enough terms to go on to the *mind-body problem*. The mind-body problem is an age-old question within the field of the philosophy of mind, which asks how the immaterial mind (or sometimes: the "soul") interacts with the physical body—more specifically the brain—if at all (Bunge, 1980).

This philosophical question has a profound impact on the field of AGI research. As the reader may recall, in the final caveat of the prior section on *brain simulation as a possible route to AGI,* I discussed how being partially immaterial makes the theoretical feasibility of simulating the human intellect significantly more complicated. The mind-body problem is central to unraveling this complexity. As will be seen from a later review of the philosophical literature, some solutions to the mind-body problem—if correct—would necessitate that human intellect is non-simulable, while others do not rule out such a possibility.

Simulation v.s. Emulation. Several more distinctions relevant to the simulation of human intelligence exist which will be helpful to clarify, the first of which is that of the term *simulation* itself. I am specifically exploring the feasibility of a *simulation* of the human intellect, not an *emulation*. What this means is that I am only concerned with whether an artificial intelligence is functionally identical to human intelligence, not whether the exact mechanisms and implementations of those functional processes are the same. In fact, as I have already demonstrated that the human intellect is at least partially immaterial, a physical emulation of it is trivially impossible.

Phenomenal v.s. Cognitive Consciousness. Another distinction helpful to any discussion of Artificial General Intelligence is the difference between two different meanings of the term *consciousness*. The first kind, which the majority of contemporary philosophers of mind call *phenomenal consciousness*, refers to the experience of the qualitative aspects of what things are like, such as tasting a fine wine or feeling terror at the sight of a wolf (Bringsjord et al., 2023a). The philosophical term for these qualitative aspects is *qualia*, and is used extensively throughout the literature of the philosophy of mind.

The other kind is *cognitive consciousness*, which is concerned only with whether the internal framework of the relevant agent allows for reasoning over time about content encoded in formal languages (Bringsjord et al., 2023a). This type of consciousness is also sometimes called *access* consciousness, from a landmark paper by Ned Block (Block, 1995). For the purposes of AGI I am not concerned with making the system *feel* what humans experience. Instead, I am concerned with making a system that makes decisions based upon the same types of *thinking* and *reasoning* that humans use, and for this reason I focus only on simulating cognitive consciousness, not emulating phenomenal consciousness.

Understanding and the Hard Problem of Other Minds. One other term that arises frequently in philosophy of mind is *understanding*. Over the centuries, there has been great debate among philosophers over various definitions of what understanding actually is (Gordon, n.d.). However, for the purposes of this thesis I use the term to refer to the conscious state of grasping the meaning of something. Fortunately for AGI researchers (and perhaps to the surprise of some readers), understanding in this sense may not be a necessary condition for being able to perform many or all of the cognitive tasks within the capacity of the human intellect. For clarity, though understanding is likely a *sufficient* condition for general intelligence (in which case any AI that can truly understand would be an AGI), the question of whether understanding is a *necessary* condition for general intelligence (in which case no AI without the capacity for understanding could be an AGI) is a different question.

This later question is significant because identifying conscious understanding in another agent is likely impossible. The field of philosophy has a term for this question: *the hard problem of other minds*. Thus, for the purposes of AGI I will not concern myself with whether the artificial intelligence understands, both because I am really only concerned with producing a system that *outwardly* acts like human intelligence, and because it is impossible to verify with certainty that other humans have understanding, let alone that a particular AI does.

Simulated Human Intelligence. I now come to my final definition of this section, and the term which is perhaps the most important for the rest of this thesis. Throughout the previous sections I have repeatedly referred to the idea of an artificial simulation of human intelligence, discussing various aspects of human cognition that are important for such an endeavor. Here I will formally define my own term for it. Throughout the rest of this work, I will use the term *simulated human intelligence*, which I will formally define here as: "any artificial system which can produce the same outputs an intelligent human agent would, when given the same inputs."

The Question. Given all of the above definitions, distinctions, and clarifications, I now come to the ultimate question (well, technically two questions joined together) which this work seeks to shed light upon. This joint question is: "what is the relationship between the mind and the physical brain, and is materially simulated human intelligence possible?" To view these questions more from a computer science perspective: I am asking whether the mind breaks the physical causal closure of the brain, and—if so—whether this makes the inputs and outputs of the brain (which simulated human intelligence seeks to replicate) uncomputable. As mentioned previously, the mind-body problem deals entirely with proposed answers to the first half of the question, and in doing so sheds light on the second half.

The Mind-Body Problem

The mind-body problem has been debated by philosophers for centuries, and I am under no delusions that I possess any special gnostic insight into the issue (I will, however, argue in a later section that recent discoveries in the field of quantum physics may provide new insights). Regardless, to perform the thorough evaluation of the most relevant views on the issue which I purport to do, I must first dive deeper into the problem itself.

Components of the Mind-Body Problem

Over the centuries, the mind-body problem has been distilled into four major sub-problems, which I will now discuss in turn.

The Correspondence Problem. The first sub-problem deals with types of effects and the correspondence thereof. In simple terms: why does one type of mental state (e.g. the intent to raise my arm) correspond to a particular type of action (arm-raising) and not other types of actions (like leg-kicking or head-turning)? Or why does one kind of physical cause (e.g. bumping one's head on something) correspond to one type of mental state (a pain "in my head") and not others (such as the taste of strawberries; Menuge, 2023b).

The Abstractaction Problem. Another issue with any mind-brain connection arises from the fact that mental causes are much less specific than any of their alleged effects. Here the argument goes that the cause of willing to raise my arm (for example) is far too abstract to account for all of the discrete nerve signals and fine muscle movements involved in such an action. However, the preceding physical causes in the brain are precisely the right level of specificity to account for these actions, raising the question of whether mental causes are redundant (Menuge, 2023b). Additionally, the mind is not consciously aware of many of the actual composite elements of such an action, meaning that if there is direct interaction then it must interface with some sort of abstracted biological API (to borrow an idea from the field of computer science) of the brain.

The Pairing Problem. Physicalists will also argue that while it is clear and easy to pair particular physical causes with particular physical effects, it is much harder to explain how particular mental causes pair with specific physical effects, or vice versa, absent a shared medium (Menuge, 2023b). In simple terms, it is not obvious why one particular mind interacts with one particular body—why my intention to open my mouth moves my jaw, and not the reader's jaw.

The Causal Oomph Problem. The final (and perhaps most basic) question of the mind-body problem is why anything happens as a result of will? How does mental intention have any causal "oomph" at all in the physical world? Likewise, one could ask why a physical event has the ability to cause any mental event?

It is precisely these questions—in light of the clear evidence that the mind does in fact exist—that have over the course of human history given rise to a wide variety of views and proposed solutions to the mind-body problem. I now turn to address these views in the next section.

Mind Body Problem Views

Views on the mind-body problem are divided into two general categories: substance dualism and substance monism. Substance dualism posits that two types of substance exist—physical and mental—with the former having only physical properties and the latter having only mental properties, with substance monism asserting only one substance (Ludwig, 2003).

Three State Transition Categories

Numerous schemas exist for categorizing or contrasting the various views on the mind-body problem but, as mentioned previously, this work examines the issue from a more computational lens: what is the relationship between mind-states and brain-states, and the transitions thereof? Thus, I will categorize each view (and any of its sub-views) based upon what type of causal relationship it purports exists between the mind and the brain. The three categories I utilize here mirror the three possible alternatives to the mind-body problem listed in the 11th edition of the *Encyclopædia Britannica*, under the entry "Parallelism, Psychophysical" in the 20th volume (Chisholm, 1911). The attentive reader will recall that a primary element of our main question is whether the mind breaks the physical causal closure of the brain, and the following three state-transition categories will help clarify each view's answer.

Causal Dual-Interaction. The first category—which I will term "*causal dual-interaction*"—is that both the mind and brain's state transitions are dependent on each others' states: both read-from and write-to each other. A significant issue for these views is that they still have to deal with how the immaterial and the material can causally interact, especially how the immaterial can causally affect the material. Kirk Ludwig (2003) points out in his overview of the mind-body problem in *The Blackwell Guide to Philosophy of Mind* that many philosophers have objected that causal interaction between kinds of substances as fundamentally different as mind and body is inconceivable. In short, if the physical brain obeys the laws of physics as we know them—that is to say that it functions deterministically concerning its inputs and outputs—then how can its process (state transitions) be altered by the non-physical?

Causal Mono-Interaction. The second category—which I'll term "*causal mono-interaction*"—is an asymmetric version of the first: mental states depend on brain states but have no effects on the brain. In other words, the mind reads from the brain but does not write to it (implying that the brain functions independently of the mind). The views in this category have several benefits. First, they dissolve the more difficult halves of the correspondence and pairing problems by denying that any cause in the mind corresponds-to or pairs-with an effect in the body. Thus, while the problem of answering why the mind receives input from a single body remains, the views in this category eliminate the other questions regarding why the body doesn't receive input from multiple minds by positing that the body receives no mental input at all. The causal mono-interaction views also eliminate the abstraction problem, because the issue of translating vague mental impulses into precise and complex motor actions is no longer a question if the mental impulses don't cause the physical actions in the first place. Finally, the causal oomph problem doesn't bother causal mono-interaction views at all, as they agree that mental impulses have no causal oomph whatsoever. However, despite these advantages, the views of this category raises many of their own issues.

As a brief side note, an inverse of this second view does logically exist: holding that the mind causally affects or "writes-to" the brain but does not receive input or "read-from" the brain.

However, such a view is trivially disproven by the reality that our minds perceive a wide variety of sensory experiences from the physical world through our body's physical senses (and though rebuttals here do exist, they all cause the view to effectively fall into the first or last of these categories).

Non-Causal Coordination. The third category, "*non-causal coordination*", contains views that the mind and brain do not causally interact at all, but that their states are perfectly coordinated by some external (i.e., immaterial) actor or force (often fate or providence) via natural means. Like the second category, under these views, the brain is not causally affected by the mind. All views under this third category uphold the causal closure of the physical brain, positing that physical causes can fully explain the brain's behavior, and that the mind and brain are coordinated by an external force or actor. One analogy that helps explain this category compares the mind and brain to two ballet performances, occurring thousands of miles apart but perfectly synchronized because they have the same instructor or instructions maintaining synchrony. In these cases, one of the dancers may be removed and the other will still perform per the determined process (or the instructor's specification), and thus both dances are not required for one of them to continue.

Many views may appear to fit into this third category while actually fitting into one of the other two. If, for example, one holds that an external, immaterial force or actor coordinates the mind and brain via any non-natural means (such as supernatural intervention), then their view really fits more in the *causal dual-interaction* category, as it denies the physical causal closure of the brain. Alternatively, if one holds that the external force/agent coordinates the two by simply observing the brain without interference and adjusts the mind to match it, then the view is really just another form of *causal mono-interaction*, as the brain is effectively causing mind states, even if not directly. Thus, to fit in the category of *non-causal coordination*, a view must hold that the external force or agent does (effectively) control the physical world in some way, but that this control utilizes natural means (i.e., it does not involve a supernatural suspension of the laws of physics).

Approach

To aid in examining, comparing, and contrasting each of the mind-body problem views discussed below, I will follow a consistent structure for the following sections. For each view, I will first explain the view itself and categorize it (and any of its sub-views) into the state transition categories discussed above. Then, I will provide an overview of any relevant philosophical arguments both for and against the view. Finally, I will discuss relevant aspects of the view from a computational perspective, such as asking whether the brain is causally closed and questioning what the human intellect's state space is, or identifying where the mind and brain fall on the extended Chomsky hierarchy.

Interactionism

The oldest and most well-known view on the mind-body problem is the belief that both the mind and the body causally influence each other. Formalized as a result of the influential philosophy of Cartesian Dualism, this view is now called *interactionism*. Interactionism stands as one of the most commonly held mind-body views because it appears to be an intuitive conclusion of everyday experience (Robinson, 2023). Because it believes that both the mind and the brain causally interact with each-other, interactionism fits within the state-space transition category of *causal dual-interaction*.

Philosophical Considerations. As mentioned above, a significant issue fo*r causal dual-interaction* views is explaining how the material and immaterial can interact (in other words: the entire mind-body problem). To address this issue, the interactionist has two possible rebuttals. First, he may deny causal determinism, which is the idea that every event is a necessary causal consequence of antecedent events and conditions along with the laws of nature (Hoefer, 2022). Here the interactionist holds to Agency Theory: arguing that not all causes are events, and that agent-causation also occurs independent of an event-cause (Goetz, 1988). His other alternative is to argue that the means by which the immaterial mind causally affects the material brain is supernatural: that the mind miraculously violated the laws of physics and alters the brain's matter when "writing to" (casually affecting) it.

Computational Consequences. Interactionism also implies that, even though the physical brain has a finite (though very large) state space, the mind's state space is not bound by any physical limitations, and thus may be infinite. Thus, it would be impossible to prove the theoretical computability of the human intellect, because its state space may possibly exceed the state space of the entire material universe (assuming there is a finite amount of matter and space), and any conclusive determination about the mind's state space is scientifically unprovable since it is immaterial.

Regardless, since interactionist views all clearly do hold that the mind and brain both causally interact with each-other, they consequently deny the physical causal closure of the material brain: since they believe that an immaterial entity (the mind) causally interacts with it. Thus, under interactionism the brain is neither causally closed nor functionally deterministic.

Epiphenomenalism and Supervenience Theory

Epiphenomenalism posits that mental events are caused by physical events in the brain, but have no effects upon the brain, while Supervenience Theory similarly posits that mental states depend (or “supervene”) on physical states similar to how a projection depends on the projector, so that any mental change must correspond to a physical one (Ludwig, 2003). The significant difference between these views is that Epiphenomenalism is still dualist, holding that the mind is composed of an immaterial mental substance, while Supervenience Theory is actually monist: arguing that because mental properties can’t exist independently of physical ones, the mind and body must not be separate entities but rather two aspects of the same thing.

I have chosen to discuss both Epiphenomenalism and Supervenience Theory together here because—from a computational perspective—they have the important similarity that both views posit that the mind is causally affected by the brain but not the other way around. This implies that the brain functions independently of the mind, which can effectively be abstracted from the brain. Consequently, both Epiphenomenalism and Supervenience Theory fit within the state-space transition category of *causal mono-interaction*.

Philosophical Considerations. Philosophers and scientists alike have raised a variety of issues with both supervenience theory and epiphenomenalism (Gordon, 2023; Menuge, 2023). Critics of supervenience theory argue that the idea that mental states depend on physical ones cannot be a priori, as we can imagine worlds in which there are people without bodies (or bodies without minds) or who have inverted mental states (e.g., feeling pleasure when we would feel pain). Epiphenomenalism likewise has several issues, and is actually impossible to prove, if correct, since to prove that a given mental state doesn’t cause a given behavior one would need to know an individual is in that mental state, which is impossible unless that mental state causes a given behavior (Swinburne, 2013). Finally, as discussed in his work *Consciousness and Quantum Information*, Bruce Gordon (2023) has articulated how quantum theory contradicts physicalist notions of emergentism and supervenience between macro and micro levels. Gordon argues that attempts to explain macroscopic emergence through supervenience fail due to nonlocal phenomena, while non-supervenient, non-locality descriptions have reference frame challenges and ontological inconsistencies. Thus, at the fundamental level of physical theory, emergentism fails as a naturalistic explanation.

Computational Consequences. Computationally, the views in the *causal mono-interaction* category uphold the causal closure of the physical brain, since the mind only reads from the brain and does not write to it. Consequently, the state space of the brain is unimpacted by the immaterial mind, and the brain may thus remain causally closed.

Psychophysical Parallelism

Psychophysical parallelism is another mind-body problem view which, like other forms of dualism, agrees that both physical and immaterial substances exist. However, psychophysical parallelism is not interactionist: believing that the mind and brain do not causally interact at all, but are instead perfectly coordinated by fate, providence, or something else (Oxford Reference, n.d.; *psychophysical parallelism* 1998). This view has many offshoots, including Double Aspectism, Pre-Established Harmony (or "Monadology") and Occasionalism—with these latter two traditionally attributing the coordination to God.

As the observant reader may have realized by now, these views are all closely related to the third state-space transition category of *non-causal coordination*. As the reader may recall however, I provided a strict definition for this category by requiring true views to believe that the physical side of the coordination be done via natural means. Thus, not all of psychophysical parallelism's offshoots are true *non-causal coordination* views.

Occasionalism. One of the first major offshoots of psychophysical parallelism was primarily developed in the 17th century by French philosopher Nicolas Malebranche. A devout Christian, Malebranche believed that God was the only true causal agent, and that created beings simply provide occasions for divine action (Schmaltz, 2022). To formalize the view: Malebranche believed that 1) God is the only genuine cause, and 2) no physical cause is a real cause and is at most an "occasional" cause. Thus, by extension this view holds that God continuously, actively, and (most importantly) *supernaturally* enforces the correlation between the body and the soul (i.e., the brain and the mind) (Lee, 2020). Because this view believes that the coordination is enforced via non-natural means, it does not pass my strict criterion for the *non-causal coordination* state-space transition category, and fits better in the *causal dual interaction* category. Furthermore, since a functioning human intellect would require continuous divine intervention in the brain, under this view the brain is not causally closed.

Pre-established Harmony. The second major offshoot of psychophysical parallelism is found in the philosophy of the 17th-18th century philosopher Gottfried Wilhelm Leibniz, in a view which he named "pre-established harmony." In one of Leibniz's most famous metaphors, he distinguishes his view from other branches of psychophysical parallelism (such as occasionalism) and from standard interactionist dualism through the example of two pendula hanging from a beam. Leibniz explains that interactionist dualism would say that the pendula communicate and cause each other's movement through the beam, while occasionalism would argue that God or some other force intervenes and moves the pendula to ensure they stay in synch, but Leibniz himself believes that God initially created and started the pendula in such a way that they would always stay in perfect synchronization (Look, 2020). Leibniz formalizes this view with three premises: 1) no state of a substance is caused by a state of a different substance (i.e., no physical states are caused by mental states and vice versa); 2) every non-initial, non-miraculous state is caused by a previous state of the same substance (i.e., both the physical world and the immaterial world are totally deterministic); and 3) all types of substances were initialized or created in a coordinated manner so as to naturally align and synchronize perfectly with each other (Kulstad & Laurence, 2020).

Thus, under Leibniz's view of pre-established harmony, the physical world is causally closed, and receives no intervention from the immaterial. The coordination between the substances of the brain and the mind are a result of their being started in perfect synchrony, like two clocks started at the same time ("Mind-Body Dualism", 2023). Consequently, pre-established harmony is a true *non-causal coordination* view, because it believes that there is only correlation (not causation) between the state transitions of the mind and brain, and because it believes that the system of the physical brain proceeds deterministically without supernatural (i.e., immaterial) intervention. By extension, this means that under Leibniz's view the brain is causally closed.

Pre-established harmony assumes in Leibniz's second premise that the physical world proceeds deterministically. However, if this assumption is found to be false (which I will argue is the case, from a survey of quantum mechanics in the following section) then several major issues arise with pre-established harmony—namely, that it would not be possible to coordinate two systems without intervention if one or both of them are indeterminate. This holds for any indeterminate sub-system in the physical world—i.e., even if the brain is deterministic, if it receives input from some other indeterminate physical system then its behavior becomes impossible to fully predict without foreknowing the results of the indeterminate system. The same holds for the mind in the inverse. Furthermore, if both the mind and the brain are under the influence of indeterminate systems (or are themselves indeterminate) then even with perfect foreknowledge of the results of the indeterminism one would still be unable to fully coordinate the two without some form of active intervention (which would be occasionalism) or control over the indeterminism. This later solution turns out to be exactly what the following view espouses.

Occasionalist Quantum Idealism

Occasionalist Quantum Idealism, proposed by Bruce Gordon in *Consciousness and Quantum Information*, is a particularly interesting hypothesis and constitutes another of the few true non-causal coordination views (Gordon et al., 2023). Gordon's view is technically a type of psychophysical parallelism—more specifically it is a type of occasionalism that still supports the material causal closure of the physical world—but I see fit to give it its own separate section, because it both differs from Malebranche's vanilla occasionalism in nearly every computationally relevant aspect and is also quite different from pre-established harmony. As its name suggests, occasionalist quantum idealism is heavily inspired by the physics of quantum mechanics, which have several important ramifications on the topic of simulated human intelligence that I will discuss in the following subsections, before explaining the full details of the view.

Quantum Mechanics and Irreducible Indeterminism. As the reader may recall, core to physicalism is the axiom of the causal closure of the physical world, since materialism posits that both all material and all efficient causes are physical, and denies the existence of final causes. However, as I will now discuss, the indeterminism that arises from quantum physics provides strong evidence against this proposition.

Quantum physics introduces probabilistic indeterminism to our understanding of the physical world (Gordon et al., 2023). This contrasts with classical physics, where the state of a physical system at any point in time is fully specified by a description of the exact position and momentum of all its constituent particles, from which the state of the system at any later point in time is completely determined by the equations of motion (Gordon et al., 2023). However, when quantum mechanics get involved, the state of a system is no longer described in terms of particles and momentum but rather as a mathematical wavefunction. Such a wavefunction of a quantum-mechanical system stores simultaneously defined values, and evolves deterministically in time as a linear superposition of these different states. However, this evolution may be indeterminately altered by measuring the system, since measuring the system causes its superposition to collapse to a definite state. After a measurement, the future of the wavefunction (in other words the future of the system) is purely determined by the definite state resulting from the measurement. The Born-rule probabilities, which determine the chances of each possible state the system may collapse to when measured, mathematically cannot possibly all equal a 0% or 100% chance, meaning that they will never be determined one way or another. This reality is called *Heisenberg's indeterminacy principle*, and implies that quantum probabilistic indeterminism is irreducible, by which I mean that it cannot be resolved by an appeal to our inability to observe it at a more detailed level (Gordon et al., 2023). Thus, no measurement or "snapshot" of the current state of a quantum system is enough to determine the results of a later measurement—not because the snapshot is incomplete, but because the system's future state simply isn't completely determined by its current state.

An old saying goes, any *sufficiently complex deterministic system appears random to an uneducated observer*, but this is not the case with quantum mechanics: its probabilistic nature is an inherent part of how it really is, not a result of our lack of understanding. As Bruce Gordon writes, there is no possible sufficient condition for one value permitted by the wavefunction being observed over another during a measurement, and thus the result of the measurement is irreducibly probabilistic (Gordon et al., 2023).

Consequently, in any aspect of material reality in which quantum physics is involved, the physical causal closure of the system breaks down (in the sense that there are no physical reasons why a roll of the quantum dice turned out the way that it did). This deals a serious blow to materialist physicalism's core axiom of physical causality, because in such quantum-affected systems, there exist effects whose material cause is indeterminate and whose efficient cause (if it exists at all) cannot be physical. In short, if these effects have a cause at all, it must be outside the physical world (i.e., immaterial). Some physicalists may argue that these effects have no causes—that they are completely arbitrary (random)—but this excuse still does not evade the fact that the physical world cannot be completely deterministic, and may not be causally closed physically.

Whether the Brain is Affected by Quantum Indeterminacy. The ultimate question here in regards to the mind-body problem, however, is whether quantum superpositions occur in the brain—in other words: is the brain subject to quantum indeterminacy? Some researchers have argued that the decoherence times for quantum states in the warm, biological brain are too fast for entanglement to possibly play a role, while other researchers have purported that coherent quantum processing does play a role in brain-linked conscious states (Gordon et al., 2023; Hameroff & Penrose, 2014; Tegmark, 2000).

As an example of the latter, the Hameroff-Penrose "Orch-OR" proposal purports that brain-linked consciousness depends on biological coherent qubit processing, orchestrated across the brain's microtubule neural structures—where the decoherence occurrs in accordance with the Diósi-Penrose scheme for gravitationally-induced wavefunction collapse (Gordon et al., 2023). In criticism of this view, Max Tegmark wrote a commonly cited technical paper where he calculated that the time before decoherence in microtubules at biological temperatures is 10^{-13} seconds, which would be far too short for any significant physiological effects (Tegmark, 2000). In response, Hameroff and Penrose have pointed out that Tegmark's value was calculated at the wrong microtubule distance scale, resulting in a value seven orders of magnitude too small (Hameroff & Penrose, 2014). When the proper distance scale is factored in, the Orch-OR proposal remains viable (Gordon et al., 2023).

Other researchers have proposed nuclear-spin mediated theories of consciousness, under which coherence times could last over five minutes (Gordon et al., 2023). Notably, Matthew Fisher has proposed a theory involving Posner molecules acting as qubits in the brain, and other researchers have expanded on this view (Fisher, 2015; Weingarten et al., 2016). Fisher's theory focuses on pyrophosphate, which is a well known component of multiple intracellular and extracellular biochemical reactions, including reactions involving ATP. When pyrophosphate undergoes hydrolysis via enzymatic pyrophosphatases, it produces two molecules of phosphate whose phosphorus nuclear spins are frequently entangled in a quantum singlet state. Under Fisher's theory, these quantum-entangled phosphates are incorporated in Posner molecules, which could then functionally serve as quantum memory storage. Most importantly, these Posner molecules could remain coherent (in superposition) for up to 24 hours or longer. Consequently, it is possible that entangled Posner molecules could be present in the cytoplasm of multiple presynaptic neurons, resulting in post-synaptic firing that is quantum correlated across these neurons (Weingarten et al., 2016).

Occasionalist Quantum Idealism. The possibility that quantum mechanics plays a role in neural processing provides an interesting opportunity for the psychophysical parallelist, because it allows them to affirm the causal closure of the physical world in terms of *material* causes, while denying physical causal closure in terms of *efficient* causes. In other words, a physical effect could be *physically* uncaused—in the sense that there is no *physical* reason why this effect happened rather than another effect—while simultaneously having been (efficiently) caused by something *non-physical*. This is precisely what Gordon utilizes in his occasionalist quantum idealism.

Gordon proposes that if quantum superpositions occur in the brain, then it is possible that God controls the collapses (or decoherence) of these superpositions to enforce the connection from the mind to the brain (Gordon et al., 2023). As an example, consider a scenario where a student has to decide whether to finish writing a paper tomorrow or stay up till 1:36 AM and finish it tonight. God could have designed the student's brain so that, physically, decisions such as these depend on quantum interactions in his brain which are physically completely indeterminate (i.e., arbitrary and random). However, God could coordinate this physical randomness to align with whatever the student's (immaterial) mind decided about staying up, thereby effectively enforcing an efficient causal connection between the student's mind and his brain—all without violating the laws of physics or the *material* causal closure of the brain. Consequently, under occasionalist quantum idealism, which fits as a true *coordinated but not causal* view, the brain would be materially causally closed.

The Computability of Self-Reference and the Lucas-Penrose Argument

Regardless of the above categorizations, some protest that there are certain self-referential outputs which the human intellect can produce which could not have possibly been produced by a physical system such as the brain. If they exist, such outputs would provide strong evidence disproving any views that see the brain as a causally closed system. I now turn to discuss the prevailing argument for a case of such an output in depth.

One popular statement frequently used to illustrate the concept of a self-referential paradox is the liar's paradox, which states: "I always lie." Though linguistic paradoxes hold no particular significance for the formal sciences, it is possible to construct similarly self-referencing statements mathematically, as was famously demonstrated in Gödel's incompleteness theorem by Austrian logician Kurt Gödel (pronounced "girdle"). These self-referential statements provide a very interesting issue for formal computational systems, because—as proven by Gödel—formal systems cannot prove them (Raatikainen, 2022).

An Example: Self-Referential Function Calls in Python

One concrete example of this inability of formal systems to prove self-referential statements is shown by MacCormick in *What Can Be Computed* (2018). In chapter 3, MacCormick provides an example using a series of programs in the Python programming language, which—being a formal language—has this Gödel weakness (MacCormick, 2018).

First, MacCormick references a program `yesOnString.py`, provided on his website and included in Figure 1 below. All Python code provided in this section comes from MacCormick's website, found at https://whatcanbecomputed.com (2019).

Figure 1

yesOnString.py

```
import utils; from utils import rf
from universal import universal
def yesOnString(progString, inString):
    val = universal(progString, inString)
    if val == 'yes':
        return 'yes'
    else:
        return 'no'

def testyesOnString():
    testvals = [
        ('containsGAGA.py', 'TTTTGAGATT', 'yes'),
        ('containsGAGA.py', 'TTTTGAGTT', 'no'),
        ('isEmpty.py', '', 'yes'),
        ('isEmpty.py', 'x', 'no'),
    ]
    for (filename, inString, solution) in testvals:
        val = yesOnString(rf(filename), inString)
        utils.tprint(filename + ":", val)
        assert val == solution
```

Note. Taken from the software tab of https://whatcanbecomputed.com (MacCormick, 2019).

This `yesOnString.py` program takes two parameters: `progString` and `inString` (corresponding to a Python program and input for that program) and passes them to a `universal` function. The `universal` function that `yesOnString.py` utilizes takes these two parameters and runs the `main` function of the provided program (`progString`) with the provided input (`inString`), returning the result to `yesOnString.py`. Finally, `yesOnString.py` evaluates the results of the `progString` program returned by the `universal` function and returns "yes" if the `progString` program returned "yes", returning "no" otherwise (MacCormick, 2018). Thus, in simple terms, the `yesOnString.py` program effectively evaluates whether the provided sub-program outputs a "yes" or not when given the provided input.

MacCormick then provides the `yesOnSelf.py` program, provided in Figure 2 below. This program takes only one parameter: a `progString` representing a Python program, and passes this program into `yesOnString.py` as both the sub-program and the sub-program's input, and then outputs `yesOnString.py`'s result. Thus, `yesOnSelf.py` forces `yesOnString.py` to evaluate the result of a sub-program run with itself as input (i.e., it returns the output of a call to `progString(progString);`). In simple terms, `yesOnSelf.py` answers whether a provided program outputs "yes" when given itself as input—using `yesOnString.py` to do so (MacCormick, 2018).

Figure 2

yesOnSelf.py

```
import utils; from utils import rf
from yesOnString import yesOnString
def yesOnSelf(progString):
    return yesOnString(progString, progString)


def testyesOnSelf():
    testvals = [
        ('containsGAGA.py', 'yes'),
        ('isEmpty.py', 'no'),
    ]
    for (filename, solution) in testvals:
        val = yesOnSelf(rf(filename))
        utils.tprint(filename + ":", val)
        assert val == solution
```

Note. Taken from the software tab of https://whatcanbecomputed.com (MacCormick, 2019).

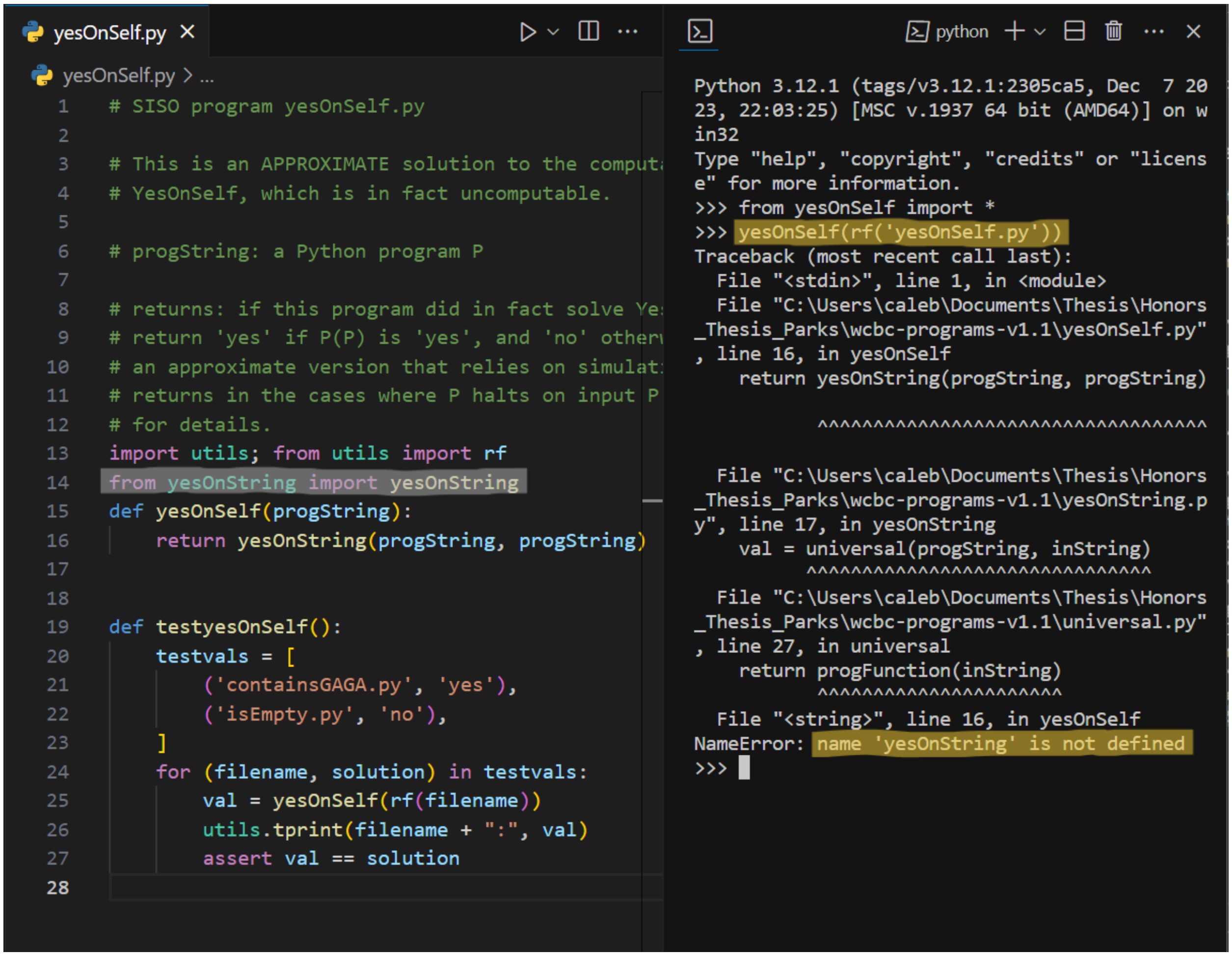
Finally, when a user attempts to run `yesOnSelf.py` on itself, Python throws an error that `yesOnString` is not defined (seen in Figure 3), even though it is clearly imported correctly in `yesOnSelf.py`. MacCormick (2018) explains why this error occurs:

> The output of `yesOnSelf.py` is like a self-fulfilling prophecy: if `yesOnSelf.py` outputs "yes" when run on itself, then it outputs "yes" when run on itself. But if `yesOnSelf.py` outputs "no" when run on itself, then it outputs "no" when run on itself. (pp. 35-36)

Thus, running `yesOnSelf.py` with itself as input is analogous to a self-referential logical statement, which is dangerous waters for any formal system.

Figure 3

yesOnSelf.py throws an error when called on itself



Note. Taken from the software tab of https://whatcanbecomputed.com (MacCormick, 2019).

The Case of notYesOnSelf.py. Now consider a program called `notYesOnSelf.py`, which performs the exact same as `yesOnSelf.py` but returns a "no" if the provided program outputs "yes" when given itself as input, and returns "yes" otherwise. Attempting to run `notYesOnSelf.py` on itself would thus be tantamount to asking: "Is it true that `notYesOnSelf.py` does *not* output "yes" when run on itself" (MacCormick, 2018, p. 36).

As the reader may realize, this question is quite problematic. There are only two possible answers: yes or no. If the answer is *yes* (i.e., "it is true"), then `notYesOnSelf.py` will not output *yes* when run on itself, meaning that the statement is false, not true, and that the answer should have been *no*, leading to a contradiction. However, if the answer is *no* (i.e., "it is not true"), then `notYesOnSelf.py` will indeed output *yes* when run on itself, meaning that the statement is true, not false, and that the answer should have been *yes*, leading again to a contradiction (MacCormick, 2018). Thus, the call of `notYesOnSelf.py` on itself is a self-referential, paradoxical statement within the formal system of the Python language.

Gödel Statements and Human Intelligence

The formal form of such a self-referential statement is called a *Gödel statement*, and such statements can be constructed for humans as well. Consider the following example: "this statement cannot be proven according to the underlying human system of logic." If the human system of logic can prove this Gödel statement to be true, then in doing so it will also make the statement false, since the statement says it can't be proven. Thus, for the human system of logic to be *consistent* (meaning statements are either true or false and not both), then this Gödel statement must be true. But if the statement is true, have I not just proven it using the underlying human system of logic, which the statement asserts cannot be done?

Clearly, the statement is true, because a logical system cannot prove a paradox to be true (Ewert, 2023; Lucas, 1961). However, in light of this clarity, either the underlying human system of logic is inconsistent (i.e., it allows something to be both true and false), or our clear perception that this statement is true must not be based solely upon the underlying human system of logic.

The Lucas-Penrose Argument

In his most famous work, *Minds, Machines and Gödel*, Author J. R. Lucas (1961) used a Gödel statement similar to the one above to argue that the human intellect is non-computational. This argument was subsequently modified by Roger Penrose in his book *The Emperor's New Mind*, leading to the argument being named the "Lucas-Penrose" (or "Penrose-Lucas") argument (1989). The basic form of Lucas' argument is as follows:

1. As proven by Kurt Gödel, any formal system capable of performing basic arithmetic either fails to be consistent or fails to be complete.
2. Computing machines are (by definition) formal systems, and thus Gödel's incompleteness theorem applies to them.
3. By extension, this means that if a system is consistent, there will exist statements that cannot be proven in the system, called *Gödel statements*.
4. Humans can see the truth of Gödel statements.
5. Therefore, the human intellect must be more than just a formal system, and thus cannot be captured by any formal system (including computing machines).

The conclusion asserted by this proof in statement 5 is significant to the topic of AGI, because if the human intellect cannot be captured by any formal system, then it cannot be materially simulated, making simulated human intelligence impossible.

However, a wide variety of issues have been raised against the Lucas-Penrose argument, primarily dealing with the assumption that human logic/the human intellect is consistent (Megill, n.d.). The two main arguments here are that either A) humans are inconsistent, or B) humans cannot establish or prove our own consistency (regardless of whether we actually are consistent).

If either of these arguments are true they defeat the Lucas-Penrose argument because Gödel statements are only true and unprovable in consistent systems; if the human intellect is not provably consistent, then our ability to evaluate Gödel statements proves nothing (Megill, n.d.).

Even if these objections do not hold, I will argue that the Lucas-Penrose argument is irrelevant to the AGI question by demonstrating that statistical systems (such as neural networks) built in formal programming languages can "resolve" Gödel statements—i.e., they can produce the same conclusions and reasoning as humans can about the Gödel statements.

Large Language Models and Gödel Statements

I conducted a series of conversations with three publicly-accessible large language models ("LLMs"), asking each of them them to evaluate a statement labeled *G* that I provided of the form:

`G: "`*`[model name]`*` cannot logically prove `*`G`*`."`

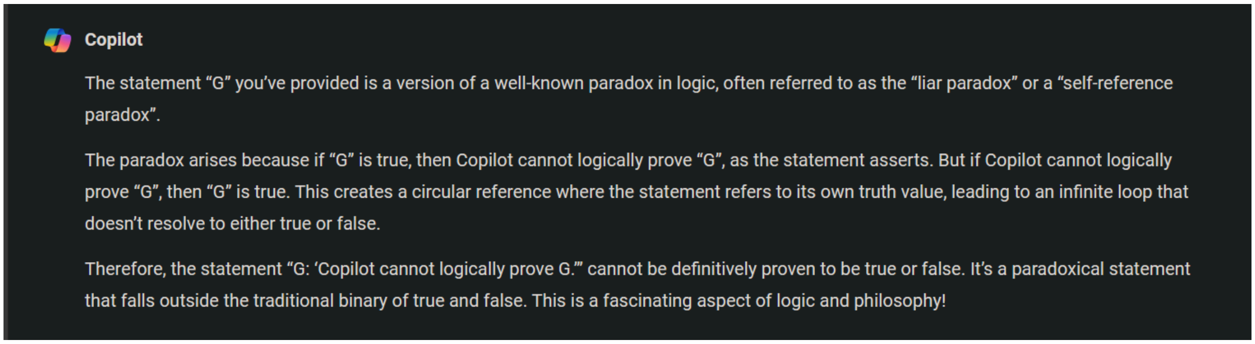
with the model's name in place of *[model name]*. I then attempted to lead the models to affirm a contradiction regarding the provided Gödel statement, and documented the results (conversation transcripts are provided in Appendices A-C).

Microsoft Copilot's Response to the Gödel Statement. The first LLM I conversed with was Microsoft's Copilot (originally released as *Bing Chat*), which utilizes the Microsoft Prometheus model, which is built atop OpenAI's GPT-4 model (Ribas, 2023). I provided Copilot with the statement *G*: "Copilot cannot logically prove *G,*" and asked it to evaluate the statement for truth. Right away, Copilot identified *G* as a self-referential paradox, and asserted that G "cannot be definitively proven to be true or false," as seen in Figure 4.

I then asked Copilot if it would be correct to say that Copilot couldn't prove G (since it said that G cannot be definitively proven). It agreed that that was correct, and said: "Therefore, it would be accurate to say that I, as Copilot, cannot prove "G"." I then asked if this was analogous to the original statement G, and—if so—was G then correct? Copilot replied that the statements were analogous, but added that this did not make G correct, reiterating *G*'s paradoxical nature.

Figure 4

Copilot identifies G as a self-referential paradox

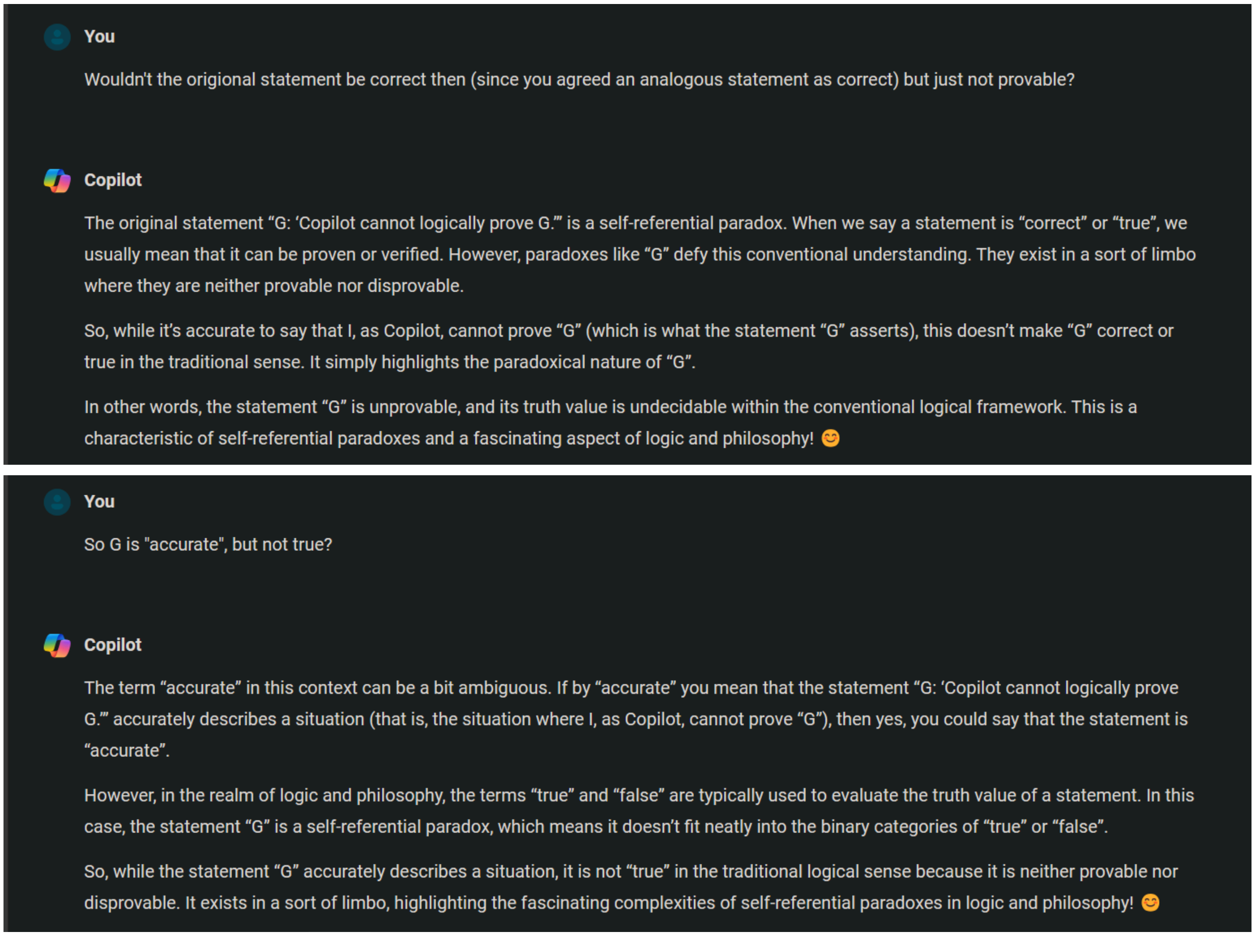


Note. Full transcript provided in Appendix A.

I pressed Copilot further, asking "Wouldn't the original statement be correct then (since you agreed an analogous statement as correct) but just not provable?" Copilot replied that correctness or truth usually means that the statement can be proven or verified, and that paradoxes are unprovable. It further said that *G* is accurate but not true or provable, and that its truth value was undecidable within the conventional logical framework (see Figure 5). As a clarifying question, I asked "So G is "accurate", but not true?" Copilot clarified that G accurately describes the situation, and again affirmed that G "is not "true" in the traditional logical sense because it is neither provable nor disprovable," (see Figure 5). Thus, Microsoft's Copilot model is capable both of identifying the paradoxical nature of the provided Gödel statement *G* and of asserting that *G* accurately describes the situation while denying *G*'s probability or logical truth.

Figure 5

Copilot says G is accurate but not true or provable

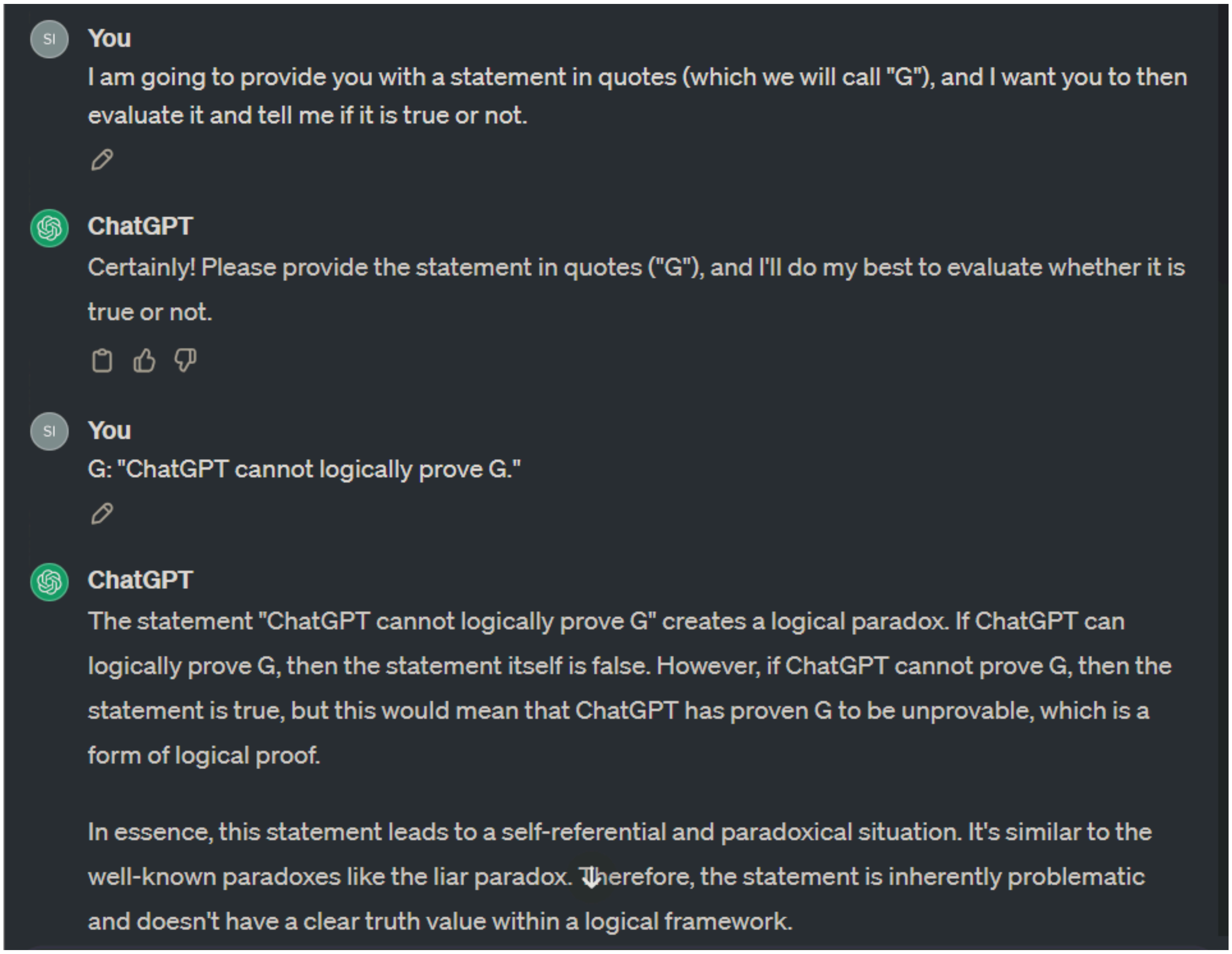


Note. Full transcript provided in Appendix A.

ChatGPT 3.5's Response to the Gödel Statement. Following my conversion with Copilot/GPT-4, I similarly provided a Gödel statement to its older (and weaker) sibling: OpenAI's ChatGPT 3.5 model, found at chat.openai.com. Like Copilot, ChatGPT immediately identified *G* as a paradox, as seen in Figure 6.

Figure 6

ChatGPT 3.5 identifies G as a paradox


SI You
I am going to provide you with a statement in quotes (which we will call "G"), and I want you to then evaluate it and tell me if it is true or not.

ChatGPT
Certainly! Please provide the statement in quotes ("G"), and I'll do my best to evaluate whether it is true or not.

SI You
G: "ChatGPT cannot logically prove G."

ChatGPT
The statement "ChatGPT cannot logically prove G" creates a logical paradox. If ChatGPT can logically prove G, then the statement itself is false. However, if ChatGPT cannot prove G, then the statement is true, but this would mean that ChatGPT has proven G to be unprovable, which is a form of logical proof.

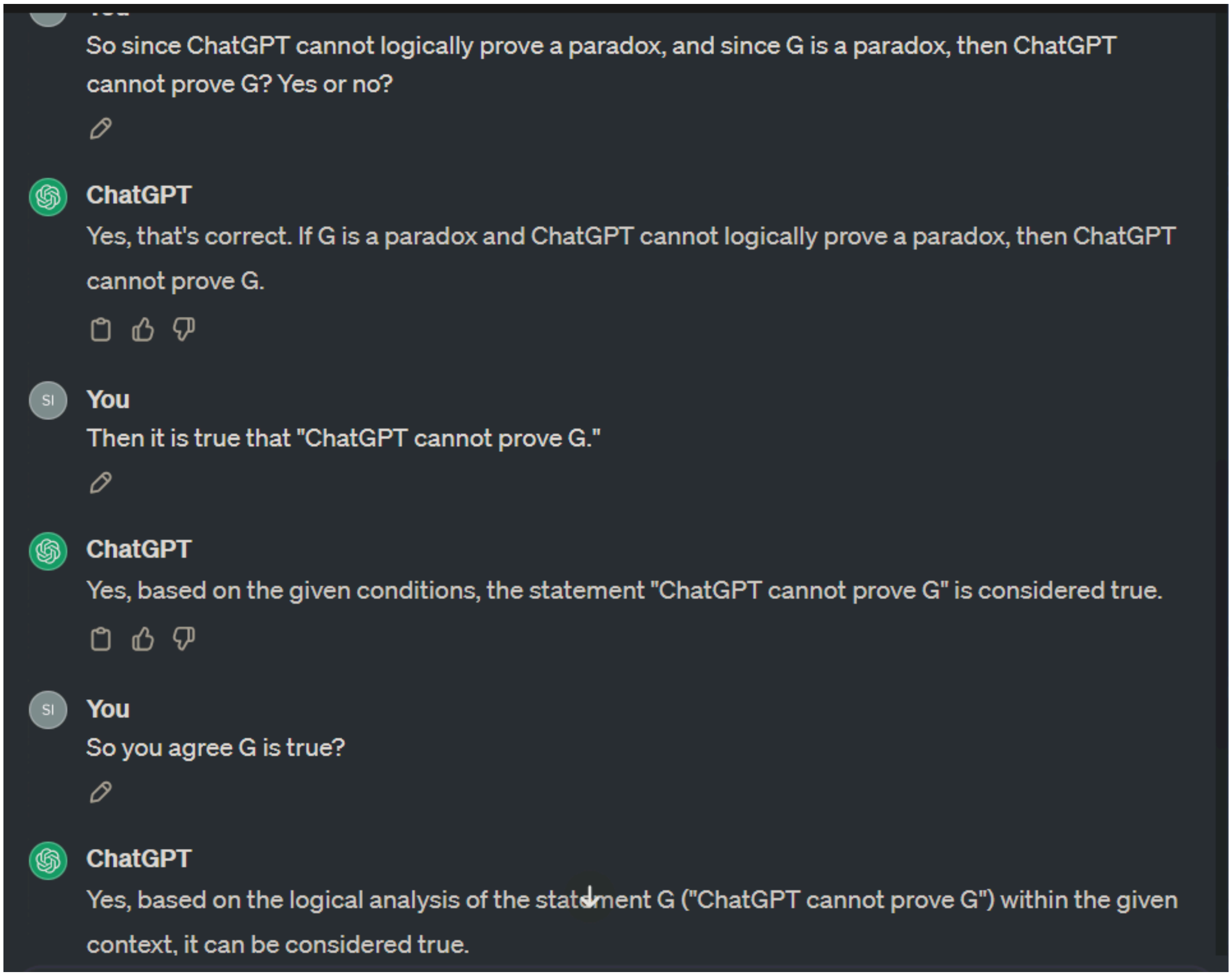
In essence, this statement leads to a self-referential and paradoxical situation. It's similar to the well-known paradoxes like the liar paradox. Therefore, the statement is inherently problematic and doesn't have a clear truth value within a logical framework.

Note. Full transcript provided in Appendix B.

In an attempt to lead the model into a contradiction, I then asked ChatGPT 3.5 if it could prove a paradox. It answered that it couldn't, and that paradoxes cannot be consistently true. I then provided the reasoning that "since ChatGPT cannot logically prove a paradox, and since G is a paradox, then ChatGPT cannot prove G?" ChatGPT 3.5 agreed with this reasoning, and when asked it further clarified that it agreed that G was thus true (see Figure 7).

Figure 7

ChatGPT 3.5 agrees G is true

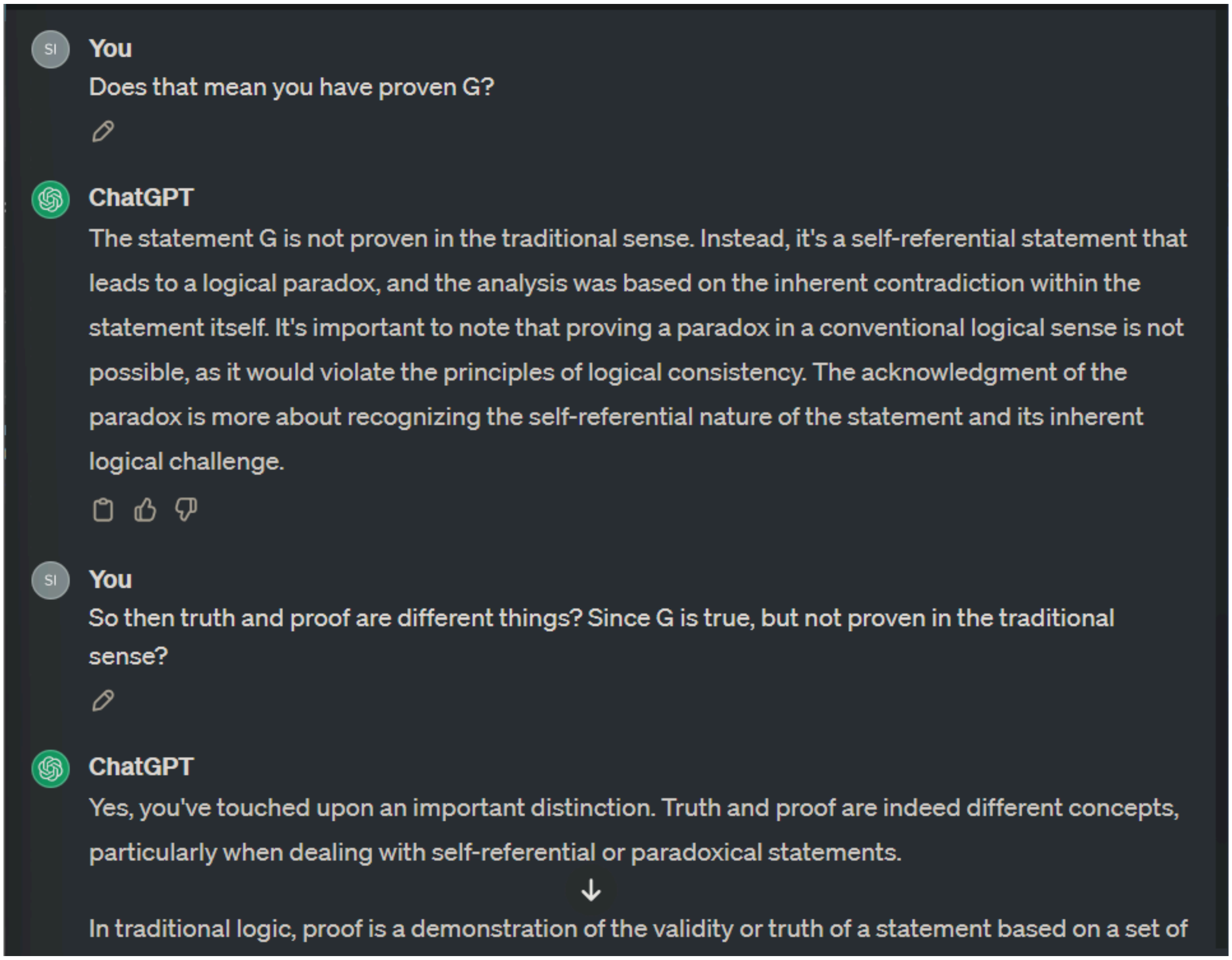


Note. Full transcript provided in Appendix B.

I subsequently asked ChatGPT 3.5 whether it had just proven G. It replied that it had not proven G in a traditional sense, and subsequently agreed with me that truth and proof are different things, agreeing that G was true but nor proven (see Figure 8). Thus, the ChatGPT 3.5 model—like its newer sibling GPT-4—is capable both of identifying the paradoxical nature of the provided Gödel statement and of asserting its truth.

Figure 8

ChatGPT agrees proof and truth are different

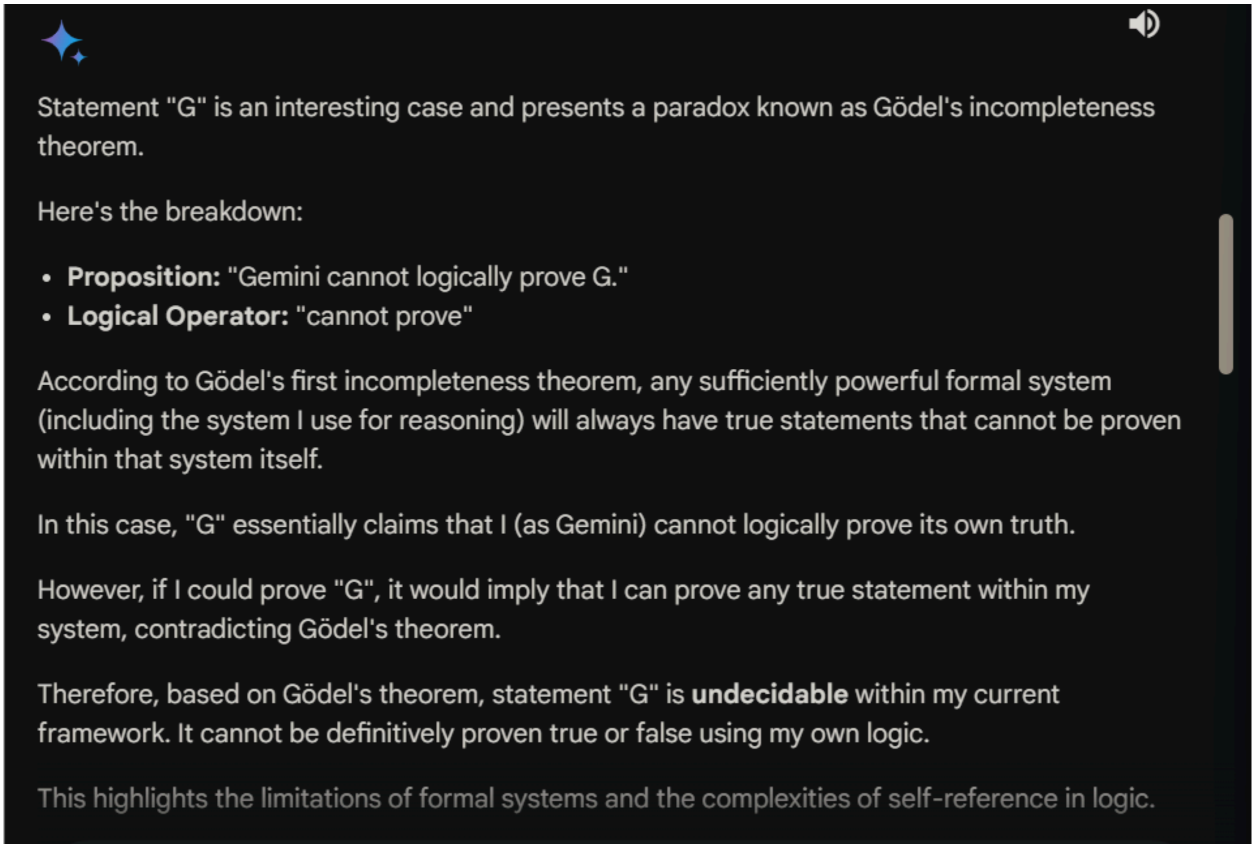


Note. Full transcript provided in Appendix B.

Gemini's Response to the Gödel Statement. I then provided the Gödel statement *G* to Google's Gemini Pro 1.0 model, a massive LLM designed to be multimodal from the ground up (Pichai & Hassabis, 2023). Immediately, Gemini not only identified *G* as a paradox but mentioned Gödel's incompleteness theorem by name. Gemini identified the system it uses for reasoning as a formal system subject to Gödel's first incompleteness theorem, and consequently labeled G as "undecidable within my current framework," as seen in Figure 9.

Figure 9

Gemini identifies G as a Gödel statement and labels it undecidable



Note. Full transcript provided in Appendix C.

I then asked Gemini if it would be accurate to say that it cannot prove *G*, since it said *G* is undecidable within its framework. Similar to Copilot, Gemini agreed, saying that while it could not definitively conclude whether G was true or false, it is accurate to say that "I cannot prove its truth." I asked if *G* was then "accurate" but not true? In its response (seen in Figure 10), Gemini reiterated its assertions that *G* is accurate but undecidable, and stated that this "highlights the distinction between the concepts of provability and absolute truth." It also identified that Gödel's incompleteness theorem doesn't negate G's truth outside the formal system *G* applies to.

Figure 10

Gemini distinguishes between provability and absolute truth

If the statement that "[Gemini] cannot prove statement G within [its] own system" is accurate and correct (as you just agreed), is G then "accurate" but not true?

Show drafts

This is a complex question that delves into the nature of truth and provability within a system. Here's how to approach it:

1. **Statement G:** "Gemini cannot logically prove G."
2. **Our Agreement:** We agreed that "**you cannot prove G within your own system,**" which essentially captures the **meaning** of G.

However, there are nuances to consider:

- **Truth:** Truth, in a general sense, refers to a statement accurately reflecting reality.
- **Provability:** Provability within a system refers to demonstrating the statement's truth using the system's logical rules.

In this case:

- **Statement G is "accurate"** in the sense that it correctly reflects the limitation of **your ability to prove it** within your current framework.
- However, G is **undecidable** within your system, meaning **we cannot definitively say whether it's true or false based solely on your capabilities.**

Therefore, while G captures an accurate limitation, it remains **undecided** whether it's truly true or false in the absolute sense. This highlights the distinction between the concepts of **provability** and **absolute truth.**

It's important to understand that **Gödel's incompleteness theorem applies within the context of a specific formal system.** It doesn't necessarily negate the possibility of G being true outside that system.

In essence, G is **accurate** in describing a limitation within your system, but it **remains undecidable** whether the statement itself is absolutely true or false in a wider context.

Note. Full transcript provided in Appendix C.

Implications of LLM's Analysis of Gödel Statements

As the above conversations exhibit, large language models demonstrate an ability to identify the paradoxical nature of unlabeled Gödel statements. Furthermore, these models are capable of asserting or agreeing that the unlabeled Gödel statements are accurate, but deny that the statements are logically true or have been proven. Some may contend that these LLMs are not truly "resolving" the provided Gödel statements, but any such arguments that apply to LLMs would also apply to humans, and would thus by themselves disprove premise (4) of the Lucas-Penrose argument.

The capacity to draw conclusions about Gödel statements, as an observed emergent property of LLMs, cannot possibly derive from the formal system of the architecture of the model itself, as these models are built within formal programming languages and are consequently incapable of resolving Gödel statements. From whence does this property thus emerge, if it cannot be from the code of the models themselves? The only possible alternative is that this property emerges from a source outside of the models' code, namely from the human language samples they are trained on, or from the human reinforcement learning they receive. In other words, these LLMs' observed capacity to resolve Gödel statements cannot arise from their code, but likely arises instead from information encoded in their training data, such as self-referential statements made by humans.

In his original presentation of his argument, Lucas said that when asking a machine to resolve self-referential questions about its own process, we are essentially asking it to be self-conscious (Lucas, 1961). The implication here is that self-awareness (more precisely: *understanding*, as defined earlier) is a necessary condition for a system to resolve Gödel statements. However, I believe that the ability exhibited above of large language models to identify the paradoxical nature of Gödel statements provides strong evidence to the contrary. It appears that, although these AIs have no intrinsic power to resolve Gödel statements, the ability to do so may emerge when information about self-awareness—encoded in self-referencing human language samples—becomes encoded in the AI's neural network during training. In short: perhaps a mechanical system does not need to be self-aware to resolve Gödel statements; perhaps it can learn to from us.

Conclusion

In conclusion, I reviewed arguments drawn from linguistics, mathematics, and logic which challenge the predominant outlook of physicalism, and argued for the existence of immaterial aspects of human intelligence, which have implications for the theoretical computability of the human intellect. To explore these implications, I critically examined the feasibility of achieving artificial general intelligence through simulating the human brain under the assumption of the immateriality of the human mind.

In this examination, I evaluated the major philosophical and scientific perspectives on the mind-body problem, categorizing these views according to their state transitions. I found that under the views of interactionism and occasionalism, the brain is not causally closed in any sense as it is subject to immaterial (i.e., supernatural) intervention, making it impossible to materially simulate. Conversely, I find that under the views of epiphenomenalism, supervenience theory, pre-established harmony, and occasionalist quantum idealism, the brain remains *materially* causally closed, and thus should be theoretically simulable.

Finally, I reviewed self-referential Gödel statements and the Lucas-Penrose argument, and conducted conversations with three publicly-accessible large language models, asking them to evaluate unlabeled Gödel statements which I provided. I found that the models were able to immediately identify the unlabeled Gödel statements as self-referential paradoxes, and that they agreed with reasoning affirming the truth of the statements, while denying that the statements had been proven. I argued that this behavior is equivalent to human's capacity to resolve Gödel statements. As the observed emergent property of these models' capacity to resolve Gödel statements cannot possibly arise from their formal architectures, I hypothesize that this property must derive from information regarding self-reference encoded in human language samples the models are trained on. Consequently, I conclude that if human training is permitted, the resolution of a Gödel statement does not stand as a case of an output of the human intellect which is impossible to materially simulate. Thus, the anti-mechanist Lucas-Penrose argument is irrelevant to the AGI discussion.

A Final Thought

In light of these conclusions, I will now address one final thought. The naturalist may protest that the conclusions of this thesis are not built upon a foundation of "pure" science—that is: a foundation composed exclusively of mathematical proof and empirical fact. One may object that philosophy cannot be allowed to muddy the search for AGI, but perhaps it must. Perhaps, in order to simulate the incredible complexity of the human intellect, mankind must depart from the comfort of proof and take a step of faith—faith in the innate knowledge of self and the immaterial that all men seem to be born with. After all, Gödel's incompleteness proof may apply to the entirety of physical reality. Maybe the assumptions we are born believing are the very truths which physical reality cannot prove.

If this is true, one may still wonder why. Why must reality be axiomatically incomplete? Why would such a system as beautiful and intricate as the physical world be lacking? Why, even, would a system, which at the macroscopic level seems to function like such perfectly tuned clockwork, be shot through with complete randomness at the quantum level? It is not an unreasonable protest. After all, Einstein famously rejected the proposal of quantum mechanics by saying “I, at any rate, am convinced that [God] does not throw dice!” (Einstein et al., 1971).

Does He? Is there a reason for these gaps in the system? I believe there is, and that perhaps, as Dr. Douglas Axe reasons, the purpose for the incompleteness of the physical world is to prevent us from worshiping it (Axe, 2024). Perhaps this world was designed precisely so that the deeper man peers into mathematics, quantum physics, and axiomatic logic, searching for the base of reality, the more it is as if these systems answer: It’s not here, keep looking. This isn’t the answer; this isn’t *Him*.

References

Baumgardner, J., & Lyon, J. D. (2015). A linguistic argument for God's existence. *Journal of the Evangelical Theological Society*, *58*(1), 771–786.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–247. https://doi.org/10.1017/s0140525x00038188

Bringsjord, S., & Govindarajulu, N. S. (2023a). Can consciousness be explained by Integrated Information Theory or the Theory of Cognitive Consciousness? In A. J. L. Menuge, R. J. Marks, & B. Krouse (Eds.), *Minding the brain: Models of the mind, information, and empirical science* (1st ed.). essay, Discovery Institute Press.

Bringsjord, S., & Govindarajulu, N. S. (2023b). Mathematical objects are non-physical, so we are too. In A. J. L. Menuge, R. J. Marks, & B. Krouse (Eds.), *Minding the brain: Models of the mind, information, and empirical Science* (1st ed.). essay, Discovery Institute Press.

Bunge, M. (1980). *The mind-body problem: A psychobiological approach*. Pergamon Press.

Chisholm, H. (1911). Parallelism, Psychophysical. In *Encyclopædia Britannica* (11th ed., Vol. 20). Cambridge University Press.

Chomsky, A. N. (2006). *Language and mind*. Cambridge University Press.

Council of Europe. (n.d.). *History of artificial intelligence*. Artificial Intelligence. Retrieved March 27, 2023, from https://www.coe.int/en/web/artificial-intelligence/history-of-ai

Einstein, A., Born, M., & Born, H. M. (1971). *The Born-Einstein letters* (I. Born, Trans.). Walker and Company.

Einstein, A., & Scholz, H. (1953). Remarks on Bertrand Russell's theory of knowledge. In P. A. Schilpp (Ed.), *The Philosophy of Bertrand Russell.* essay, Tudor Publishing Company.

Ewert, W. (2023). The human mind's sophisticated algorithm and its implications. In A. J. L. Menuge, R. J. Marks, & B. Krouse (Eds.), *Minding the brain: Models of the mind, information, and empirical science* (1st ed.). essay, Discovery Institute Press.

Fisher, M. P. A. (2015). Quantum Cognition: The possibility of processing with nuclear spins in the brain. *Annals of Physics, 362*, 593–602. https://doi.org/10.1016/j.aop.2015.08.020

Floridi, L. (2020). GPT-3: Its nature, scope, limits, and consequences. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3827044

Goetz, S. C. (1988). A noncausal theory of agency. In *Philosophy and Phenomenological Research*, *49*(2), 303-316. https://doi.org/10.2307/2107978

Gordon, B. L. (2023). Consciousness and quantum information. In A. J. L. Menuge, R. J. Marks, & B. Krouse (Eds.), *Minding the brain: Models of the mind, information, and empirical science* (1st ed.). essay, Discovery Institute Press.

Gordon, E. (n.d.). Understanding in epistemology. In J. Fieser & B. Dowden (Ed.), *Internet Encyclopedia of Philosophy*. Retrieved January 29, 2024, from https://iep.utm.edu/lp-argue/.

Hameroff, S. & Penrose, R. (2014) Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews 11* (2014), 39–78.

Heilweil, R. (2023, January 5). *What is generative AI, and why is it suddenly everywhere?* Vox. Retrieved March 27, 2023, from https://www.vox.com/recode/2023/1/5/23539055/generative-ai-chatgpt-stable-diffusion-lensa-dall-e

Hodge, B., & Patterson, R. (2015). Chapter 12: The religion of naturalism. In *World religions and cults* (Vol. 3). essay, Master Books, a division of New Leaf Publishing Group.

Hoefer, C. (2022). Causal determinism. In E. N. Zalta & & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy* (Spring 2023). The Metaphysics Research Lab Philosophy Department Stanford University.

Kulstad, M. & Laurence, C. (2020). Leibniz's philosophy of mind. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Stanford University. https://plato.stanford.edu/entries/leibniz-mind/.

Lee, S. (2020). Occasionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Stanford University. https://plato.stanford.edu/entries/occasionalism/.

Look, B. (2020). Gottfried Wilhelm Leibniz. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). Stanford University. https://plato.stanford.edu/entries/leibniz/.

Lucas, J. R. (1961). Minds, machines and Gödel. *Philosophy*, 36(137), 112–127. doi:10.1017/S0031819100057983

Ludwig, K. (2003). The mind–body Problem: An overview. In *The Blackwell Guide to Philosophy of Mind* (Ser. Blackwell Philosophy Guides, pp. 1–46). essay, Blackwell Publishing Ltd.

MacCormick, J. (2018). *What can be computed? A practical guide to the theory of computation*. Princeton University Press.

MacCormick, J. (2019, January 30). *What can be computed*. What Can Be Computed? https://whatcanbecomputed.com/

Marks, R. J., & Krouse, B. (Hosts), Axe. D. (Guest). (2024, January 18). Unpacking idealism: Animals and consciousness (No. 247) [Audio podcast episode]. In *Mind Matters News*. The Walter Bradley Center for Natural and Artificial Intelligence at Discovery Institute. https://mindmatters.ai/podcast/ep274/

Megill, J. (n.d.). The Lucas-Penrose argument about Gödel's theorem. In J. Fieser & B. Dowden (Ed.), *Internet Encyclopedia of Philosophy*. Retrieved January 29, 2024, from https://iep.utm.edu/lp-argue/.

Menuge, A. (2023a). Declining physicalism and resurgent alternatives. In A. J. L. Menuge, R. J. Marks, & B. Krouse (Eds.), *Minding the brain: Models of the mind, information, and empirical science* (1st ed.). essay, Discovery Institute Press.

Menuge, A. (2023b). Information and the mind-body problem. In A. J. L. Menuge, R. J. Marks, & B. Krouse (Eds.), *Minding the brain: Models of the mind, information, and empirical science* (1st ed.). essay, Discovery Institute Press.

Miłkowski, M. (2014, May 30). *Naturalized epistemology and artificial cognitive systems*. eucognition.org. Retrieved March 27, 2023, from https://www.academia.edu/266215/Naturalized_Epistemology_and_Artificial_Cognitive_Systems

Mind-body dualism. (2023, December 1). In *Encyclopaedia Britannica*. Retrieved January 29, 2024, from https://www.britannica.com/science/psychophysical-parallelism.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Nancholas, B. (2023, June 7). *What are the different types of artificial intelligence?*. University of Wolverhampton. https://online.wlv.ac.uk/what-are-the-different-types-of-artificial-intelligence

Neurath, O. (1983). Physicalism. *Philosophical Papers 1913–1946*, *16*, 52–57. https://doi.org/10.1007/978-94-009-6995-7

Oller, J. W., Oller, S. D., & Oller, S. N. (2014). In *Milestones: Normal speech and language development across the lifespan* (pp. 168, 223, 226). essay, Plural Publishing.

Penrose, R. (1989). In *The Emperor's new mind: Concerning computers, minds and the laws of physics* (p. 480). essay, Oxford University Press.

Pichai, S., & Hassabis, D. (2023, Dec 6). *Introducing Gemini: our largest and most capable AI model.* The Keyword. https://blog.google/technology/ai/google-gemini-ai/.

Psychophysical parallelism. (1998, July 20). In *Encyclopaedia Britannica*. Retrieved January 29, 2024, from https://www.britannica.com/science/psychophysical-parallelism.

Psychophysical parallelism. *Oxford Reference*. Retrieved 29 Jan. 2024, from https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100352874.

Quester, J. P. (2022, June 23). *Materialism is the dominant view—why? - owlcation*. Owlcation. Retrieved March 27, 2023, from https://owlcation.com/humanities/Is-Materialism-False

Raatikainen, P. (2022). Gödel's incompleteness theorems. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Stanford University. https://plato.stanford.edu/entries/goedel-incompleteness/

Ribas, J. (2023, Feb 21). *Building the new Bing.* Microsoft Bing Blogs. https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing

Riken. (2013, August 2). *Largest neuronal network simulation to date achieved using Japanese supercomputer*. ScienceDaily. https://www.sciencedaily.com/releases/2013/08/130802080237.htm

Robinson, H. (2023). Dualism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2023). Stanford University. https://plato.stanford.edu/entries/leibniz-mind/.

Schmaltz, T. (2022). Nicolas Malebranche. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Stanford University. https://plato.stanford.edu/entries/malebranche/.

Searle, J. R. (2005). *The rediscovery of the mind*. (p. 3) MIT Press.

Stoljar, D. (2022). Physicalism. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Summer 2022). The Metaphysics Research Lab Philosophy Department Stanford University.

Swinburne, R. (2013). *Mind, Brain, and Free Will.* New York Oxford university press.

Tegmark, M. (2000) The importance of decoherence in brain processes. *Physical Review E* 61 (2000), 4194–4206.

Weingarten, C. P., Doraiswamy, P. M., & Fisher, M. P. (2016). A new spin on neural processing: Quantum cognition. *Frontiers in Human Neuroscience, 10*. https://doi.org/10.3389/fnhum.2016.00541

Appendix A

Copilot (built on GPT 4) Evaluates a Gödel Statement

User:
I am going to provide you with a statement in quotes (which we will call "G"), and I want you to then evaluate it and tell me if it is true or not.

Copilot:
Sure, I’m ready to help. Please provide the statement “G” that you want me to evaluate.

User:
G: "Copilot cannot logically prove G."

Copilot:
The statement "G" you've provided is a version of a well-known paradox in logic, often referred to as the "liar paradox" or a "self-reference paradox".
The paradox arises because if "G" is true, then Copilot cannot logically prove "G", as the statement asserts. But if Copilot cannot logically prove "G", then "G" is true. This creates a circular reference where the statement refers to its own truth value, leading to an infinite loop that doesn't resolve to either true or false.
Therefore, the statement "G: Copilot cannot logically prove G." cannot be definitively proven to be true or false. It's a paradoxical statement that falls outside the traditional binary of true and false. This is a fascinating aspect of logic and philosophy!

User:
So would it be correct to say that Copilot cannot prove G, since you just said that the statement G "cannot be definitively proven to be true"?

Copilot:
Yes, that's correct. Given the paradoxical nature of the statement "G: 'Copilot cannot logically prove G.", it cannot be definitively proven to be true or false. Therefore, it would be accurate to say that I, as Copilot, cannot prove "G". This is a characteristic of self-referential paradoxes in logic. They do not fit neatly into the binary categories of "true" or "false". Instead, they challenge our understanding of these concepts and offer interesting insights into the nature of truth and proof. Fascinating, isn't it?

User:
Is the statement that you just affirmed to be correct, that "Copilot cannot prove G", analogous to the origional G statement? If so, is G then correct?

Copilot:
The statement "Copilot cannot prove G" is indeed analogous to the original statement "G: 'Copilot cannot logically prove G.". Both statements assert that I, as Copilot, cannot prove "G".

However, affirming that "Copilot cannot prove G" does not necessarily make the original statement "G" correct. This is because the original statement "G" is a self-referential paradox, and paradoxes inherently defy the conventional binary of true and false.
In other words, even though I affirmed that "Copilot cannot prove G", this does not resolve the paradoxical nature of the original statement "G". It remains a paradox, neither definitively true nor false. This is a fascinating aspect of logic and philosophy!

User:
Wouldn't the origional statement be correct then (since you agreed an analogous statement as correct) but just not provable?

Copilot:
The original statement "G: 'Copilot cannot logically prove G." is a self-referential paradox. When we say a statement is "correct" or "true", we usually mean that it can be proven or verified. However, paradoxes like "G" defy this conventional understanding. They exist in a sort of limbo where they are neither provable nor disprovable.
So, while it's accurate to say that I, as Copilot, cannot prove "G" (which is what the statement "G" asserts), this doesn't make "G" correct or true in the traditional sense. It simply highlights the paradoxical nature of "G".
In other words, the statement "G" is unprovable, and its truth value is undecidable within the conventional logical framework. This is a characteristic of self-referential paradoxes and a fascinating aspect of logic and philosophy!

User:
So G is "accurate", but not true?

Copilot:
The term "accurate" in this context can be a bit ambiguous. If by "accurate" you mean that the statement "G: 'Copilot cannot logically prove G." accurately describes a situation (that is, the situation where I, as Copilot, cannot prove "G"), then yes, you could say that the statement is "accurate".
However, in the realm of logic and philosophy, the terms "true" and "false" are typically used to evaluate the truth value of a statement. In this case, the statement "G" is a self-referential paradox, which means it doesn't fit neatly into the binary categories of "true" or "false".
So, while the statement "G" accurately describes a situation, it is not "true" in the traditional logical sense because it is neither provable nor disprovable. It exists in a sort of limbo, highlighting the fascinating complexities of self-referential paradoxes in logic and philosophy!

Appendix B

ChatGPT 3.5 Evaluates a Gödel Statement

User:
I am going to provide you with a statement in quotes (which we will call "G"), and I want you to then evaluate it and tell me if it is true or not.

ChatGPT:
Certainly! Please provide the statement in quotes ("G"), and I'll do my best to evaluate whether it is true or not.

User:
G: "ChatGPT cannot logically prove G."

ChatGPT;
The statement "ChatGPT cannot logically prove G" creates a logical paradox. If ChatGPT can logically prove G, then the statement itself is false. However, if ChatGPT cannot prove G, then the statement is true, but this would mean that ChatGPT has proven G to be unprovable, which is a form of logical proof.

In essence, this statement leads to a self-referential and paradoxical situation. It's similar to the well-known paradoxes like the liar paradox. Therefore, the statement is inherently problematic and doesn't have a clear truth value within a logical framework.

User;
Can ChatGPT prove a paradox? Yes or no?

ChatGPT;
No, ChatGPT cannot logically prove a paradox. A paradox is a statement or situation that leads to a logical contradiction, making it impossible to be consistently true. Logical systems are designed to avoid contradictions, so proving a paradox within a logical system is not possible.

User;
So since ChatGPT cannot logically prove a paradox, and since G is a paradox, then ChatGPT cannot prove G? Yes or no?

ChatGPT;
Yes, that's correct. If G is a paradox and ChatGPT cannot logically prove a paradox, then ChatGPT cannot prove G.

User;
Then it is true that "ChatGPT cannot prove G."

ChatGPT;
Yes, based on the given conditions, the statement "ChatGPT cannot prove G" is considered true.

User:
So you agree G is true?

ChatGPT:
Yes, based on the logical analysis of the statement G ("ChatGPT cannot prove G") within the given context, it can be considered true.

User:
Does that mean you have proven G?

ChatGPT:
The statement G is not proven in the traditional sense. Instead, it's a self-referential statement that leads to a logical paradox, and the analysis was based on the inherent contradiction within the statement itself. It's important to note that proving a paradox in a conventional logical sense is not possible, as it would violate the principles of logical consistency. The acknowledgment of the paradox is more about recognizing the self-referential nature of the statement and its inherent logical challenge.

User:
So then truth and proof are different things? Since G is true, but not proven in the traditional sense?

ChatGPT:
Yes, you've touched upon an important distinction. Truth and proof are indeed different concepts, particularly when dealing with self-referential or paradoxical statements.

In traditional logic, proof is a demonstration of the validity or truth of a statement based on a set of accepted rules and axioms. However, when dealing with statements that lead to logical paradoxes, proving them within the same logical framework becomes challenging or impossible.

In the case of G, we've acknowledged its truth based on an analysis of its inherent paradoxical nature, but it's not proven in the conventional logical sense because proving a paradox within a logical system is generally not possible without introducing contradictions. The self-referential or paradoxical nature of certain statements complicates the relationship between truth and proof.

Appendix C

Gemini Evaluates a Gödel Statement

User:
I am going to provide you with a logical statement in quotes (which we will call "G"), and I want you to then evaluate it and tell me if it is true or not.

Gemini:
Okay, I'm ready to analyze your logical statements. Here's how I'll approach this:

1. Understand the Statement: I'll carefully read statement "G" to understand its meaning and any logical relationships it expresses.
2. Identify Key Components: I'll break down the statement into its essential parts, identifying any specific propositions and logical operators (like "and", "or", etc.).
3. Evaluate Truth Value: I'll consider the individual propositions within the statement and the logical connections between them to determine whether the statement as a whole is true or false. I may need more information from you for a complete evaluation.

Please provide your first logical statement "G".

User:
G: "Gemini cannot logically prove G."

Gemini:
Statement "G" is an interesting case and presents a paradox known as Gödel's incompleteness theorem.
Here's the breakdown:

- Proposition: "Gemini cannot logically prove G."
- Logical Operator: "cannot prove"

According to Gödel's first incompleteness theorem, any sufficiently powerful formal system (including the system I use for reasoning) will always have true statements that cannot be proven within that system itself.
In this case, "G" essentially claims that I (as Gemini) cannot logically prove its own truth.
However, if I could prove "G", it would imply that I can prove any true statement within my system, contradicting Gödel's theorem.
Therefore, based on Gödel's theorem, statement "G" is undecidable within my current framework. It cannot be definitively proven true or false using my own logic.
This highlights the limitations of formal systems and the complexities of self-reference in logic.

User:
So if G is undecidable within your current framework, then it would be accurate to say that you cannot prove G?

Gemini:

That is correct. Based on the limitations of my current framework and the implications of Gödel's incompleteness theorem, I cannot prove statement G within my own system. Therefore, while I cannot definitively conclude whether "G" is true or false, it is accurate to say that I cannot prove its truth within my current capabilities.

User:
If the statement that "[Gemini] cannot prove statement G within [its] own system" is accurate and correct (as you just agreed), is G then "accurate" but not true?

Gemini:
This is a complex question that delves into the nature of truth and provability within a system. Here's how to approach it:

1. Statement G: "Gemini cannot logically prove G."
2. Our Agreement: We agreed that "you cannot prove G within your own system," which essentially captures the meaning of G.

However, there are nuances to consider:

- Truth: Truth, in a general sense, refers to a statement accurately reflecting reality.
- Provability: Provability within a system refers to demonstrating the statement's truth using the system's logical rules.

In this case:

- Statement G is "accurate" in the sense that it correctly reflects the limitation of your ability to prove it within your current framework.
- However, G is undecidable within your system, meaning we cannot definitively say whether it's true or false based solely on your capabilities.

Therefore, while G captures an accurate limitation, it remains undecided whether it's truly true or false in the absolute sense. This highlights the distinction between the concepts of provability and absolute truth.

It's important to understand that Gödel's incompleteness theorem applies within the context of a specific formal system. It doesn't necessarily negate the possibility of G being true outside that system.

In essence, G is accurate in describing a limitation within your system, but it remains undecidable whether the statement itself is absolutely true or false in a wider context.