Possibilities and Limitations of AI in Philosophical Inquiry

Compared to Human Capabilities

Keita Tsuzuki[1]

[1] Tokai High School, Nagoya, Japan

Corresponding Author: keitatsuzuki2006@gmail.com

Keywords: Philosophy, Artificial Intelligence, Natural Language Processing, Symbolic System

Abstract

Traditionally, philosophy has been strictly a human domain, with wide applications in science and ethics. However, with the rapid advancement of natural language processing technologies like ChatGPT, the question of whether artificial intelligence can engage in philosophical thinking is becoming increasingly important. This work first clarifies the meaning of philosophy based on its historical background, then explores the possibility of AI engaging in philosophy. We conclude that AI has reached a stage where it can engage in philosophical inquiry. The study also examines differences between AI and humans in terms of statistical processing, creativity, the frame problem, and intrinsic motivation, assessing whether AI can philosophize in a manner indistinguishable from humans. While AI can imitate many aspects of human philosophical inquiry, the lack of intrinsic motivation remains a significant limitation. Finally, the paper explores the potential for AI to offer unique philosophical insights through its diversity and limitless learning capacity, which could open new avenues for philosophical exploration far beyond conventional human perspectives.

Statements and Declarations

Competing Interests: The author declares that they have no competing interests.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Approval: This article does not contain any studies with human participants or animals performed by the author.

1. Introduction

The term itself implies that philosophy pertains to the love of wisdom—wherein "philo" is translated into love and "sophia" as wisdom. It is one of those assets that have contributed a great deal to building human knowledge and sciences. Up to now, this tradition of seeking wisdom has been purely a human mental exercise. These recent advances in artificial intelligence are impressive. The development of natural language technologies has culminated in the creation of large language models like ChatGPT. All this makes us draw close to the moment when AI will really become able to properly work with languages and contribute to cognitive processes. In the past, the realm of intellectual inquiry was exclusively dominated by humans, but in this rapidly evolving era of AI, it is important to examine whether AI can engage in philosophical thought.

Thus, in a sense, major philosophies in history have often had the characteristic of opening up through the flowering of new emancipating insights which break away from all conventional forms of understanding. This implies René Descartes's systematic skepticism and Immanuel Kant's Copernican Revolution. Through methodical doubt of everything that can be doubted for a basis of undoubtable knowledge, Descartes literally turned the conversation of certitude in knowledge on its head (Descartes, 1641/1996). Equally, the statement "objects must conform to our cognition" from Kant innovatively revised the functions of the observed and the observer in a relationship (Kant, 1781/1998). Cognition differs in Artificial Intelligence compared to that of the human variety, and its unique characteristics have the potential to provide new light to many of the created schools of philosophy. This research will attempt to establish, not only whether AI can be a philosophical inquirer, but also to what extent the philosophical approach in AI is different from that of a human.

Then, Chapter 2 will discuss the definition of philosophy in relation to the ability of AI in engaging in philosophical inquiry. Chapter 3 discusses how different philosophies practiced by AI are compared to those done by humans, allowing us to appreciate how much philosophy can actually be done by artificial intelligence as a human would do it. In Chapter 4, the critique on philosophy unique to AI is conducted, underlining its benefits and drawbacks.

2. Can AI Engage in Philosophy?

To begin the process of establishing whether or not artificial intelligence can philosophize, one must first attempt to outline what defines philosophy. Philosophy is a very vague and open-ended subject; however, there do exist several common hallmarks of philosophy, such as "the search for truth." Truth in this context is a simple word that symbolizes fact based on overwhelming evidence. As such, philosophy is therefore geared toward "logical persuasion." Where religion relies on superstitious supernatural beings, such as gods, angels, and demons, philosophy in most cases tries to provide a logical explanation for phenomena while using reason to convince others. As noted by John Messerly, philosophy would rely on reason, evidence, and experience in its pursuit of finding truth, whereas religion shall rely on faith, authority, and revelation

(Messerly, 2020).

Otherwise, one of the moot questions in the philosophy of history would be whether there is "truth," and if so, whether humans can perceive it. The general argument of philosophy, either that there is no such thing as truth or that it exists but is hidden from human sight, is summed up lucidly by the classical Sophist Gorgias. According to Sextus Empiricus, Gorgias argued that "nothing exists; even if something exists, it cannot be known; even if it can be known, it cannot be communicated" (Sextus Empiricus, ca. 200 CE, as cited in Bury, 1933). Within philosophical discussions, countless methodologies have been advanced to achieve truth, but deduction and induction are regarded as the best. But deduction has one serious problem: if its original premise is wrong, then any conclusion extracted from it must also be wrong. In this way, deduction faces a critical dilemma, and this limitation occurs because it lacks the ability to prove that its original premise is true.

In The Justification of Deduction, Susan Haack claims that this first premise can never be justified without circularity of argument, hence embedding one of the most significant restraints of deductive reasoning into itself (Haack, 1976). Yet induction has its limits as well, and one of the major ones was pointed out by Wittgenstein's Paradox. He proved that a conclusion given from several cases cannot be taken as truth. For instance, in the series "2, 4, 6, 8, 10, □," one can induce the next number is "12." But suppose this series were representing the rainy days of some month; then the next figure is "13," which gives proof that inductive generalizations are not necessarily true (Wittgenstein, 1953).

If philosophy were not to reach the truth, what relevance does it have? My view is that the worth of philosophy is not in its pursuit of "truth" as such but in offering interpretations that might mean something to a certain individual, group, or sentient being. As Nietzsche wrote in The Will to Power, "There are no facts, only interpretations," thus placing interpretation at the heart of the philosophical undertaking with utmost importance (Nietzsche, 1968). Therefore, philosophy can be thought of as "the search for interpretations which are logically appealing to some target agent." From this definition, one can conclude that philosophy does not have an immanent urge for the discovery of ultimate truths; rather, it looks for the interpretations important in a certain context and standpoint, which is the core of its worth. We have seen thus far, on the definitions of Gorgias and Nietzsche, that philosophy is: "the search for interpretations which are logically compelling to a particular being." With that definition in hand, we may turn to the question of whether an artificial intelligence will be able to do philosophy.

Humans accept conclusions, not strictly owing to logical validity, but owing to their capacity to supply a logical reason for one's feeling of conviction. For example, when events happen successively, many people draw the conclusion of causality. That may or may not be strictly logical, but if it feels like a logical explanation to them, then it is an acceptable one. In that respect, one might then say, "logically compelling interpretations" of course needn't meet rigorous standards for logical validity. In other words, if something seems logical to the reader or listener, then it may be philosophical in that person's mind, even without a sound logical basis.

The "Domino Fallacy" is an excellent example of how a logically valid chain of occurrences could result in an implausible conclusion. For example, someone might say, "If a country embraces some new legislative policy, then other countries will be influenced, which in turn will bring about the breakdown of the world economy." Even though there is a congenital weakness in the causal links of the sequence, it nonetheless is presented as logically valid. This kind of argument, while seemingly persuasive at face value, is nonetheless used in many, many arguments, and to those people who find it logical, it's almost philosophical.

This would then follow that inasmuch as AI could come up with "plausible" justifications for the output, AI can therefore also do philosophy. Perhaps some would say that, because the AI cognitive processing involves a different form from human reasoning, AI could actually never engage in philosophy. But very natures of philosophy are concerned with what an interpretation means to the receiver. Hence, if AI can create interpretations that have meaning for some agent, then only this AI can be said to do philosophy.

Recent studies prove that AI-generated responses are believed to be more human-like compared to responses coming from real subjects. Milani et al. (2023) showed that the behavior of AI in video game environments was tested for a higher degree of human-likeness compared to the real players. In his work, Rathi et al. (2024) spotted that GPT-4 was judged to be human-like with a considerably higher rate compared to real humans within a conceptually dislocated and inverted Turing test. The deep neural network showed, with strong indication, that AI is able to make plausible interpretations which could logically convince a human.

3. Can AI Engage in Philosophy in the Same Way as Humans?

Thus far, we have argued that AI is capable of engaging in philosophical discourse. The following four main objections which refute the assumption that an AI can philosophize in the same way humans do will be discussed. This paper will analyze these points and hopefully shed some light on how much of AI's thought processes are similar to that of a human and how much they are not, and also if AI can ever truly perform philosophical thinking like a human.

3.1. AI is merely a statistical process.

One of the common criticisms of AI is that all it does is statistical processing, and humans don't. In order to test this statement, it is necessary to understand what exactly AI is. Artificial Intelligence (AI) is basically the concept of a machine thinking and learning the way humans do. The most advanced AI methods are machine learning. Newer AIs, such as ChatGPT and image generation models, are trained on what's known as "big data" using mathematical models called neural networks that simulate how a human brain learns and is able to produce an output given an input. This is done using artificial neural networks, which are composed of elements called neurons, and by continually tweaking the synapses between these neurons in a process called learning, AI gradually gains these abilities. Thus, AI learns by statistical learning over massive amounts of data. Like how an image recognition AI is trained to recognize a cat, it is provided with

the image data, and if the output is incorrect, then the connections of the neural network are changed, and basically, it keeps learning. It is during this time that the AI reverse engineers a cat's qualities out of huge amounts of data, and thus it thinks abductively. Therefore, we have to understand the two philosophical thinking processes to contemplate this problem.

The "inductive" method, as suggested by Francis Bacon, is about deriving general principles from observations made from the physical world (Bacon, 1902). On the other hand, René Descartes proposed "deductive" reasoning, which is a systematic method in which truths are logically unveiled from a pre-existing set of established truths. Descartes (1641) used the idea of "methodical doubt" to establish the undoubtable existence of the self, from which he could therefore deduce the existence of God. This debate was resolved by Immanuel Kant, who said that humans do not experience phenomena per se, but their experiences of them are interpreted through their innate a priori sets, and they perform both kinds of induction and deduction (Kant, 1781).

The various criticisms of artificial intelligence often come together with their statistical basis. The "Chinese Room" thought experiment by John Searle strongly negates the possibility that AI could ever obtain actual consciousness or knowledge. In this thought experiment, a non-Chinese-speaking person can answer questions in Chinese with the aid of an included instruction booklet. It appears to the onlooker as if the speaker really does understand the language, while in fact, he is just manipulating symbols according to rules. Searle appeals to this example to conclude that AI itself is quite the same: it derives responses from statistical processing instead of comprehension (Searle, 1980).

The working of AI output is parallel to Searle's "Chinese Room" argument. The most statistically likely response is selected without really understanding the meaning or context. On this score, perhaps the most consistent critique of AI—the fact that it lacks any human capacity to understand anything seriously and is just a machine doing its statistical operations—returns again to whether human beings understand "meaning" as it is commonly supposed. Much of human behavior is determined by past experiences, much the same way AI models themselves learn from the data. Translated in terms of this statistically implemented brain prediction mechanism, matters get a great deal murkier. The brain is ceaselessly generating predictions and, in consideration of new sensory information, revising the said predictions by resorting to prior experience in that process. That concept, in fact called predictive coding, was issued by Friston in 2012. It would appear to prove human cognition is little different than the pattern recognition by a program of AI.

This makes the concept also in line with the Bayesian brain hypothesis, which purports that human cognition involves probabilistic reasoning and is continuously updating itself in the face of evidence relevant to how people frame their beliefs (Clark, 2013). This hypothesis goes to show that human cognitive processes may not be essentially different from those statistical analyses performed by artificial intelligence.

Kantian philosophy postulates that it is a priori knowledge of time and space that actually enables humans to reason deductively. Kant went on to say that these two ideas are categories of the mind that humans use

to view reality and thereby make logical arguments. This immediately raises a question about whether artificial intelligence is actually capable of understanding time and space, an important constituent in the philosophy underlying reasoning. Within the context of spatial reasoning in artificial intelligence, this can be achieved through computer vision and physical simulations. For example, computer vision techniques based on deep learning make it easy to extract spatial relationships presented in images and videos so that object locations and shapes can be recognized (LeCun, Bengio, & Hinton, 2015). Such functionality empowers artificial intelligence to understand the configurations of the physical environment by fusing in spatial data.

Regarding an understanding of time, models such as the RNN provide vast insights. RNNs are made to deal with data from sequences because they accumulate past history against current input. This characteristic ensures that these types of networks can learn time dependencies and predict future events based on earlier cause-and-effect relationships (Hochreiter & Schmidhuber, 1997). Temporal information processing of this sort basically merges with Kant's description of time, which suggests that, in reality, artificial intelligence does have some notion of time. Thus, it is suggested that AI may possess a priori concepts of time and space, similar to those in Kantian philosophy, and engage in reasoning in a manner comparable to humans. Therefore, the criticism that AI merely performs statistical processing and does not engage in philosophy akin to human reasoning is likely to be mistaken.

3.2. AI Is Not Creating Anything New

One of the common criticisms is that artificial intelligence is just a system based on large databases for knowledge and, unlike humans, it has no power to truly create anything. To somewhat repudiate this point of view, I would like to try my best and argue against it with the idea put forth by David Hume in regard to "complex ideas." Hume felt that all human ideas stem from combining past perceptions, meaning "creativity" actually stems from an organization of those past experiences (Hume, 1739).

Take, for example, "Pegasus": in actuality, there is no such thing, but it is a word created by combining the words "horse" and "wings" together. What is "new," therefore, has proven to be nothing more than a recombination of existing things. Similarly, artificial intelligence feeds on voluminous data and comes up with new results from the pieces of information that have been gathered. Outputs, for example, of an illustration-generating AI are products of the amalgamation of the information it has consumed and can also be considered a sort of "creation" (Goodfellow et al., 2016). Indeed, while outputs of AI are definitely re-combinations of past information, novel advances in generative AI prove that AI is producing new forms of art which have never before been created by humans. It therefore means that, when it comes to being creative, there is no fundamental difference between human beings and artificial intelligence.

3.3. AI Cannot Perform Certain Types of Thinking About Infinity

The famous physicist Roger Penrose holds the opinion that human thinking cannot be reduced to the

mechanical activities of neural cells only. In his work The Emperor's New Mind, Penrose gave proof that any computational device, including even those theoretical machines suggested by Turing, can never solve specific problems connected with infinity (Hawking & Penrose, 1996). On the other hand, it is arguable that human beings are able to solve these problems. This leads Penrose to hypothesize that human thought involves some unknown factors that are related to quantum mechanics. There is much in the world of quantum mechanics that is not explainable by classical physics, and Penrose suggests that it is these quantum phenomena that are at the foundation of human thinking (Hawking & Penrose, 1996).

However, Stephen Hawking refuted Penrose's proposal, arguing that Penrose's claim amounts to little more than saying that because quantum mechanics is mysterious and the human brain is also mysterious, there must be a connection between the two. Hawking asserted that the relationship between quantum mechanics and human thought is essentially speculative. In addition, Hawking criticizes Penrose for not using any equations or theoretical models that could prove quantum effects on human thought processes (Hawking & Penrose, 1996).

The Frame Problem is the classical problem concerning the limits of artificial intelligence. From this point of view, the frame problem indicates the weakness of artificial intelligence, as it is programmed for incomplete processing of information and hence fails to touch upon all the potential happenings in reality. For instance, an AI-driven robot being ordered to pick something from a warehouse could never conceive that a bomb might explode over the object, unexpectedly creating an accident. If AI tries to consider all possible happenings, it will get into infinite calculation loops and therefore cease to function at all (McCarthy & Hayes, 1969).

In contrast, humans are able to avoid such infinite computations. Separate from Penrose’s quantum mechanical perspective, relevance theory in cognitive science offers an explanation for this issue. According to Sperber and Wilson (1995), the human brain filters situations, goals, and prior knowledge, automatically focusing attention on information of high relevance. This process helps humans avoid cognitive overload and bypass the frame problem.

Moreover, ongoing research suggests that AI might be able to solve the frame problem through relevance theory. In their study, Russell and Norvig (2021) propose mechanisms that enable artificial intelligence to dynamically select task-relevant information while filtering out irrelevant details. In this way, AI could avoid infinite computations in a manner similar to human cognition and execute tasks more efficiently.

At present, as the complete avoidance of the frame problem has not yet been achieved, it cannot be concluded that AI is capable of thinking about infinity in the same way as humans. However, considering recent research, it can be said that AI has high potential to think similarly to humans.

3.4. AI Lacks Intrinsic Motivation

A traditional objection concerning AI goes like this: AI lacks the self-motivation to philosophize. Self-motivation amounts to a drive initiated internally in response to one's needs or desires. And to think

philosophically—operating to solve problems pertinent to oneself—requires this sort of motivation. And because it lacks such motivation, AI is incapable of any philosophical thinking, in contrast to human beings.

Susan Schneider wonders whether AI systems can be in any way like conscious or intrinsically motivated humans. She argues that whereas current AI can simulate behavior, it lacks the consciousness and internal desires responsible for human thought (Schneider, 2020). That current AI works by following objectives put to it by humans, not by engaging in philosophy for its own sake. For example, it allows AI to learn to maximize rewards through reinforcement learning. However, such rewards are defined externally, not representing autonomous goal-setting (Chadi & Mousannif, 2023).

Should artificial intelligence ever develop goals featuring self-preservation, the ability to conduct autonomous philosophical thought may well be within its grasp. As robotics expert Rodney Brooks points out, the ability for AI to coexist with humans as equals requires a foundational motivation based in the former's survival (Brooks, 2002). Thus, should it be capable of processing information related to survival and acting accordingly, one might argue that it could develop an intrinsic drive for philosophical thought.

For instance, including goals relevant to self-preservation in the reward function of AI, and/or providing it with information about its external environment, may motivate the growth of a kind of endogenous motivation in AI. Or, putting AIs into some virtual world in which they must struggle for survival should expand their ability to do autonomous philosophical reflection.

Therefore, current AI lacks intrinsic motivation, which can be seen as a clear distinction from humans. However, by incorporating goals related to self-preservation into AI, there is a strong possibility that AI could develop intrinsic motivation.

4. Philosophy Unique to AI

The last chapter considered whether AI can really engage in philosophical discourse in a manner similar to humans. This chapter we consider what AI is capable of when we consider kinds of philosophical questions that no human thinker could have. The notions of learning diversity and the infinite character of learning will be central to this discussion.

4.1. Diversity of Learning

The main feature of the new artificial intelligence lies in its ability to self-modify the experience or training data. Unlike humans, AI's flexible learning framework grants it the ability to derive perspectives that are impossible for humans, enabling it to explore aspects of knowledge inaccessible to the human mind. This capability has the potential to introduce new philosophical viewpoints to traditional philosophy.

For instance, artificial intelligence can investigate domains that need not duplicate human conceptual thought, allowing it to penetrate zones lacking notions of "money" or "love." It is almost impossible for a human being to live without possessing these concepts, although AI can create concepts even in these "restricted experiences." On the other hand, AI can also learn ideas that most human beings find difficult

to grasp simultaneously, such as learning Egyptian mythology alongside Buddhist principles. Historically, philosophy has always developed by the adoption of new methods. Artificial intelligence promises to offer fresh philosophical insights beyond the current latitude of human cognitive capacity, given its adaptive application capabilities across a broad spectrum of educational frameworks.

4.2. Infinity of Learning

Another key benefit of artificial intelligence is that AI can, in principle, learn infinitely. In reality, though, AI's learning capability is limited by hardware and memory capacity. While both of these continue to advance, the data from which AI could learn will continue to increase towards infinity. This gives AI the capacity to study a gigantic variety of data and draw insights from an amount of data that is otherwise inconceivable to the human mind.

Going back to the conceptual schema of philosophy, it would seem that the "interpretations" created by the person only apply to discrete persons or collectives rather than any general "truths." Actually, with the singularity of life experiences each person has, few major philosophical "interpretations" reach universality. Observe how it is already a fact that all of humanity does not share one single religious belief. Whereas there are principles, as in mathematics, which appear universal, our belief in their universality may only arise from our experience with people who are acquainted with mathematics. This is to say that we could only have complete assurance of the universality of such principles if we were able to behold the mind of every person. Again, the skepticism of Descartes arises here.

Consider the theoretical ideal of "AI for all humans": learning from the experiences and reflections of everyone. Such AI would be a construction of understandings based upon the collective experience of humankind, in the end making such understandings theoretically rational for every human being. Only if such an AI would not have to rely entirely on human knowledge and it were able—as it were—to digest all the knowledge from the whole world, then an "AI for the universe" could deduce universally valid interpretations for all beings.

Immanuel Kant maintained that our knowledge relies on objects, which, in philosophical terms, amounts to the Copernican revolution. This radical step suggests that our philosophical inquiries break off within the bounds of human understanding and thereby remain within the scope of human knowledge. On the contrary, a Universal AI that transcends human limits will be involved with a higher kind of philosophizing. Such an AI, in turn, can reach what has been said to be impossible: "a door to the world truth." But then again, absolute certainty still would not exist under inductive reasoning, as the rules adopted based on the cases may not always apply more broadly. The point nonetheless is this: if a Universal AI derives rules based on all information, then it would appear that there is sufficient warrant for them to be deemed true. But of course, this is also where the problem becomes analogous to that of Laplace's demon. It is a theoretical entity based on the laws of classical mechanics, which would have the potential, if it possesses complete information about the position of every particle at any instant, to predict all past and future states.

The uncertainty relation in quantum mechanics, elaborated by Heisenberg, forbids the simultaneous knowledge of a particle's position and its momentum with absolute accuracy. Hence, in the same way, Laplace's demon would not be able to predict the future with certainty either.

Besides, even with the hypothetical Universal AI, which could contain all past and future knowledge—ever evolving and thus constantly changing—it would also need education to be continually repeated in some renewed process of learning. That would be the irony, wherein even though there exists a Universal AI, man would not be able to understand the ultimate reality of the world. Unless that Universal AI existed utterly outside of our world and completely apart from us, the realization of all the truths it learns would be out of reach for us.

However, even if a Universal AI could not reach ultimate reality, its interpretations surely would be beyond human understanding and, therefore, of great benefit to mankind. Although we won't understand these interpretations in full, we will yet be able to understand them to some extent. A good analogy in this respect is that of the fourth dimension: bound by our three-dimensional nature, we cannot actually conceptualize it for ourselves, but we are nevertheless capable of theorizing about it mathematically and proposing theories to understand it. In return, the explanations that a Universal AI could provide would be values and knowledge that are new in kind but understandable for philosophical standpoints that so far lie beyond human capacities. New standpoints may even surpass human limitations and thereby bring new intellectual dimensions, which have been closed to us, into our reach.

5. Conclusion and Future Outlook

In this paper, we will address three basic questions: the degree to which AI is able to engage in philosophical discussion, if AI is able to do philosophy in a manner comparable to human philosophers, and specifically, whether for AI, some form of philosophy can exist. It is clear that AI, like any other doer, can be engaged in those kinds of practices that create "interpretations" relative to something. The differences that do exist between artificial intelligence and human beings are probably most significantly that AI intrinsically lacks motive. There are differences mainly in the methodology of philosophy rather than in the content of conclusions. That becomes a lesser problem in the case of providing "interpretations" to an audience. Moreover, as has been shown in Chapter 4, artificial intelligence can give some rather new insights and functions, both challenging and renovating the philosophical heritage developed by men. This points out that philosophy, combined with AI, has much more capacity for further developments.

This study primarily explored the theoretical dimensions of whether AI can engage in philosophy. Moving forward, it will be essential to apply the insights gained here by developing AI systems equipped with self-preservation instincts or customized learning processes, and empirically investigate the kinds of philosophical inquiry AI can perform. The findings of this study open new doors for exploration at the intersection of AI and philosophy, paving the way for further advancements in this emerging field.

References

Bacon, F. (1902). *Novum Organum* (J. Devey, Ed.). P. F. Collier. (Original work published 1620)

Brooks, R. A. (2002). *Flesh and machines: How robots will change us*. Pantheon Books.

Chadi, M.-A., & Mousannif, H. (2023). Understanding reinforcement learning algorithms: The progress from basic Q-learning to proximal policy optimization. *arXiv*.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Descartes, R. (1996). *Meditations on first philosophy* (J. Cottingham, Trans.). Cambridge University Press. (Original work published 1641)

Friston, K. (2012). The Bayesian brain: An introduction to predictive coding. *NeuroImage, 62*(2), 1230–1239.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Haack, S. (1976). The justification of deduction. *Mind, 85*(337), 112–119.

Hawking, S., & Penrose, R. (1996). *The nature of space and time*. Princeton University Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hume, D. (1978). *A treatise of human nature* (L. A. Selby-Bigge, Ed., 2nd ed.). Clarendon Press. (Original work published 1739)

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans.). Cambridge University Press. (Original work published 1781)

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence, 4*, 463–502.

Messerly, J. (2020, August 17). Philosophy, science, and religion. *Reason and Meaning*.

Milani, S., Juliani, A., Momennejad, I., Georgescu, R., Rzepecki, J., Shaw, A., Costello, G., Fang, F., Devlin, S., & Hofmann, K. (2023). Navigates like me: Understanding how people evaluate human-like AI in video games. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Nietzsche, F. (1968). *The will to power* (W. Kaufmann & R. J. Hollingdale, Trans.). Vintage Books.

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

Rathi, I., Taylor, S., Bergen, B. K., & Jones, C. R. (2024). GPT-4 is judged more human than humans in displaced and inverted Turing tests. *arXiv*.

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Schneider, S. (2020). *Artificial you: AI and the future of your mind*. Princeton University Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457.

Sextus Empiricus. (n.d.). *Against the logicians* (R. G. Bury, Trans.). Harvard University Press. (Original work published ca. 200 CE)

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Wiley-Blackwell.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell.