AI, thinking, and consciousness

Domonkos Kertész
doker24@student.sdu.dk

Syddansk Universitet, Campusvej 55, 5230 Odense M, Denmark

The source LaTeX files for this essay are available on my GitHub

Abstract. In this essay, the relationship between Artificial Intelligence (AI), human thinking, and consciousness will be explored, with reference to William J. Rapaport's *Philosophy of Computer Science*. Following the introduction of fundamental concepts, definitions, and the classification of AI, a summary of Rapaport's arguments will be presented.
A comparison will then be made between AI computation and human thinking, with a subsequent reflection on functionalism and the distinction between strong and weak AI. The conclusion of this essay will be marked by a discussion of future prospects and ethical implications.

Keywords: Artificial Intelligence · Computing · Human thinking.

1 Introduction

The question of whether Artificial Intelligence (AI) can think is a complex one, with answers that vary significantly depending on the expertise in the field of individual consulted. This is particularly true in the context of the current AI ecosystem, which is characterised by the prevalence of Large Language Models (LLMs). AI models that mimic human thinking and reasoning processes may give non-expert users the false impression that the AI is thinking.
In order to understand the thinking process of AI, it is necessary to understand human thinking. Simultaneously, this understanding facilitates a more accurate interpretation of human thought processes. [1]. The present essay aims to explore this question by drawing upon the framework of William J. Rapaport's *Philosophy of Computer Science*.

2 What is AI?

Artificial Intelligence (AI) is a field of study that lies at the intersection of computer science, cognitive science, and philosophy. It focuses on the development of systems that are capable of performing tasks that would require intelligence if they were carried out by humans. In this particular context, the term "intelligence" is understood to encompass a range of cognitive abilities including, but not limited to, reasoning, learning, planning, perception, and natural language

understanding. [2].
Despite the extensive use of the term, considerable debate continues regarding the precise definition of AI, reflecting both technological and philosophical uncertainties. Since the distinction between thinking and computation is unclear, it is impossible to say with certainty when an algorithm reaches the level at which it is considered to be thinking rather than computing.

2.1 Definitions of AI

"The science and engineering of making intelligent machines" Artificial Intelligence as a term was coined by John McCarthy in 1955.
The definition of AI depends on the philosophical perspective adopted. One influential classification was proposed by Russell and Norvig [2], distinguishes four approaches: thinking like humans, acting like humans, thinking rationally, and acting rationally. Systems that "think like humans" attempt to model human cognitive processes, whereas systems that "act like humans" aim to behave in a way indistinguishable from humans. This distinction can be evaluated by means of the Turing Test, for example. From a rationalist perspective, systems that "think rationally" employ formal logic to draw valid conclusions, while systems that "act rationally" seek optimal decisions that align with their objectives, despite the inherent uncertainties of human behaviour.
In the modern definition, an artificial system is intelligent if it is capable of processing information, adapting to its environment, and achieving goals in ways that are functionally similar to intelligent beings. This approach involves the abstraction of intelligence from biological constraints, with an emphasis placed on capabilities.

2.2 Narrow AI and General AI

Contemporary research in the field of AI is primarily concerned with narrow AI, also referred to as weak AI. Narrow AI systems are designed and trained to execute specific tasks, such as playing chess, recommending products, or recognizing faces, and frequently demonstrate superior performance to humans within their designated domain. Since these AI systems are made to solve a single specific problem, they lack the generality and flexibility that is typically associated with human intelligence.
The term general AI, or strong AI in contrast, refers to a hypothetical form of intelligence that would be capable of performing any intellectual task that a human being can. The possession of general AI by a system would enable not only the transfer of knowledge across domains but also the demonstration of self-awareness, understanding, and common sense reasoning. The achievement of general AI remains an open problem, and some theorists question whether it is even possible in principle.

2.3 Modern Approach

A substantial paradigm shift has transpired within the domain of AI research, marked by the emergence of machine learning (ML), a subfield that places emphasis on the development of systems capable of enhancing their performance through the utilisation of data, as opposed to the conventional approach of relying on explicitly programmed rules. [2]. ML approaches enable AI systems to infer patterns, make predictions, and adapt to new information through experience.
In the field of ML, the application of deep learning techniques, founded on multilayered artificial neural networks, has yielded significant advancements in areas such as computer vision, speech recognition, and natural language processing. ML has thus transformed AI from a primarily symbolic, logic-based discipline into a statistical and probabilistic enterprise. Rather than manually creating rules, researchers are now developing models that are capable of learning from extensive datasets.

2.4 Challenges in Defining AI

Despite these advancements, the definition of AI remains problematic. The notion of AI is subject to a moving goalposts phenomenon, where the tasks once considered as indicators of intelligence are progressively mastered by machines, resulting in these tasks no longer being regarded as definitive markers of intelligence. For instance, the once-impressive applications of optical character recognition and rudimentary arithmetic computation have now become considered standard practice.
Philosophical questions have the potential to introduce additional layers of complexity to the situation. The question arises as to whether the mere simulation of intelligent behaviour is sufficient for intelligence, or there is a need for genuine understanding to underpin such behaviour. The central question under consideration is whether intelligence is linked to consciousness, or it can exist independently. These questions underscore that AI is not merely a technical endeavour, but also a deeply philosophical one, raising issues about the nature of mind, understanding, and agency. The essay will expand on these questions later.

3 Rapaport on AI and Understanding

3.1 The Question of Understanding

In Chapter 18 of *Philosophy of Computer Science*, William J. Rapaport's primary concern is with the fundamental question of whether computers are truly capable of understanding anything, or they merely manipulate symbols without comprehending their meaning [1]. This fundamental question is at the core of philosophical discourse surrounding AI and consciousness, and is intimately linked to issues of syntax, semantics and intentionality.

Rapaport's theoretical framework distinguishes between two modes of processing: syntactic processing, defined as the manipulation of formal symbols according to established rules, and semantic understanding, understood as the comprehension of meaning. While computers have been demonstrated to excel at the former, they lack the latter. Therefore, it can be posited that artificial intelligence systems may manifest as intelligent entities externally, however they may not possess genuine comprehension.
This is particularly true of modern LLMs, which use a method called token prediction to predict the next word, punctuation, or sentence in a body of text in a probabilistic manner. This method of generating language is completely different from the way humans generate language, where words in a sentence are connected to each other, in a similar way to a mathematical tree. Yet, only observing inputs and outputs, communicating with an LLM is indistinguishable from communicating with a human.

3.2 Syntax and Semantics

A recurrent theme in Rapaport's discourse revolves around the interplay between syntax and semantics. It is evident that traditional computers function exclusively as syntactic engines, operating at the level of symbol manipulation. Rapaport engages with the renowned Chinese Room Argument proposed by John Searle. This argument claims that mere symbol manipulation, no matter how sophisticated, is insufficient for understanding [1]. In Searle's thought experiment, an individual who follows instructions to manipulate Chinese symbols without understanding Chinese, appears to mirror the way computers operate. his is precisely how token prediction works. LLMs generate their responses without understanding their meaning. Although current models can reason and explain their thought processes, they essentially use a mathematical method to generate text sequentially, which lacks the dimensionality of human thought.

3.3 Internal Representations

A fundamental concept that Rapaport explores is the significance of internal representations. In order for a system to be said to understand, it is not sufficient for it merely to produce appropriate outputs; it must also possess internal structures that represent its external environment or internal states in meaningful ways. Rapaport hypothesises that if a computer can effectively construct and utilise such internal representations, it may attain a form of understanding.
This concept entails a shift in focus from external behaviour alone to the internal architecture of cognitive systems. In this perspective, comprehension is derived not solely from output behaviour, but also from the manner in which information is internally processed, organised, and related to the external world. Examining LLMs with this concept in mind makes it clear that they work completely differently to the way humans think. Instead, they utilise mathematical methods such as probability matrices and functions. Understanding this is crucial when considering whether AI can think. Unlike human thinking, the processes

of AI systems can be broken down into their smallest mathematical building blocks, and we understand how they work. Currently, the human thought process remains a mystery, but would we also describe it as mathematical once we understand it?

3.4 Simulation versus Genuine Understanding

Rapaport further distinguishes between simulating understanding and genuinely possessing it. It is important to note that simulations can be behaviourally indistinguishable from genuine understanding without necessarily sharing their inner state. While a computer may simulate understanding by producing convincing outputs, the question of whether it truly comprehends remains contingent on its internal processes.
Rapaport's discussion prompts consideration of the criteria that should be employed in order to attribute understanding. The question arises as to whether behavioural indistinguishability should be considered sufficient, as proposed by Turing, or whether the search should be directed towards identifying internal properties that mirror human cognitive processes.

3.5 Broader Philosophical Context

Rapaport situates his discussion within the broader debates on the philosophy of mind and computer science. The treatment in question is characterised by its strong association with functionalism, a theoretical framework that predicates that mental states are defined by their causal roles rather than their material realization [1]. In the event of the hypothesis of functionalism being accurate, computational systems that replicate the causal structures underlying human cognition could, in principle, be capable of genuine understanding.
While Rapaport's analysis does not provide a definitive resolution to the question of whether computers can truly understand, it does offer a nuanced framework for approaching the issue. The author emphasises that understanding is a complex phenomenon involving not just observable behaviour, but also the presence of meaningful internal representations. This challenges simplistic accounts that equate intelligence solely with input-output performance.

4 Comparison and Philosophy

4.1 AI and Human Thinking: Differences and Similarities

While AI systems have been demonstrated to achieve feats of remarkable success, significant differences remain between AI based computation and human thinking. Conventional computing devices are designed to process information in a systematic, rule based manner, while human cognitive processes encompass a broader spectrum of reasoning, emotions, intuition, and context sensitive judgment.

A notable similarity is evident in their capacity for pattern recognition and adaptive learning. ML systems in the modern era have demonstrated the capacity to generalise from examples in a manner that bears a striking resemblance to the human inductive reasoning process. However, it should be noted that AI systems are devoid of the conscious awareness and subjective experience that accompany human thought processes.
Human thinking is inherently embodied and situated. This phenomenon emerges from an organism's interaction with a physical and social environment, whereas AI systems typically operate within narrowly defined parameters. Consequently, while functional outputs may occasionally converge, the underlying mechanisms and experiential qualities diverge considerably, as this essay has discussed it earlier.

4.2 Functionalism and AI

Functionalism offers an influential framework for understanding these issues. In accordance with the principles of functionalist theories, the causal role of mental states is regarded as the primary defining feature, superseding the notion of their physical substrate [2].
From this standpoint, the fundamental question is not the composition of the system, be it neuronal or silicon based, but rather the manifestation of the appropriate patterns of causal interaction among internal states, sensory inputs, and behavioural outputs.
When applied to the domain of AI, functionalism proposes that if an AI system were to implement the correct functional organisation, it could, in principle, possess genuine mental states, including understanding and consciousness. This perspective corresponds with the notion of strong AI, which suggests that meticulously engineered artificial systems may possess cognitive capabilities.
The counterargument, however, is that functional equivalence may not be sufficient. The qualitative, what it is like, experience of consciousness, may not emerge solely from functional organisation. This line of criticism is often associated with arguments from philosophers such as Thomas Nagel and David Chalmers, who emphasise the so called, hard problem, of consciousness.

4.3 Strong AI versus Weak AI

The distinction between strong and weak AI encapsulates the discourse surrounding the scope and significance of AI. Weak AI is defined as systems designed to simulate human cognitive abilities without claiming that they possess real understanding or consciousness. In this perspective, AI is regarded as a powerful instrument for achieving practical objectives. Modern LLMs fall into this category.
In contrast, the concept of strong AI states that with high enough complexity and computational power, computers possess the capacity not merely to simulate minds, but to actually instantiate them. Advocates of strong AI propose that machines with the appropriate software and hardware configurations would

possess thoughts, feelings, and genuine understanding comparable to human beings.
In Chapter 18, Rapaport puts forward a cautionary perspective on strong AI, proposing that under specific circumstances, internal representations and functional organisation may potentially lead to understanding. However, he also acknowledges the significant philosophical and empirical challenges involved in substantiating this claim [1].

4.4 Philosophical Reflections

The question of whether AI can truly think or understand raises significant issues about the nature of thought. From a purely behaviouralist perspective, as initially proposed by Alan Turing, observable behaviour alone might suffice [2]. However, argueably internal structure and subjective experience are equally essential.
The development of AI has given rise to a need for a reevaluation of traditional philosophical categories. It challenges dualist separations between mind and body, questions the human centric assumptions about intelligence, and invites a reevaluation of the boundaries between natural and artificial entities.
The philosophy of AI ultimately demonstrates the deep interconnection between computational theories of mind, and broader metaphysical and epistemological questions. As AI technology continues to advance, these philosophical reflections will become increasingly urgent, shaping not only theoretical debates but also practical decisions about the design and integration of intelligent systems in human society.

5 Future Perspectives and Ethical Considerations

The accelerated evolution of AI technologies gives rise to significant ethical concerns and considerations regarding their future implications and the moral obligations inherent in their conceptualisation and implementation. As AI systems become more sophisticated, concerns regarding safety, transparency, and the alignment of machine behaviour with human values are being raised.
A significant ethical concern relates to the realm of decision making in domains characterised by elevated stakes. AI systems are becoming increasingly prevalent in domains such as healthcare, criminal justice, and autonomous vehicles. In such contexts, potential for errors or biases could lead to severe ramifications, which raises significant concern. Ensuring fairness, accountability, and explainability in these systems is a critical challenge that demands ongoing philosophical and technical reflection.
Another concern relates to the potential emergence of autonomous agents whose actions are not fully predictable or controllable by their creators. This raises questions regarding moral responsibility and the attribution of agency, if an AI system acts in harmful ways, who should be held responsible, the developers, the users, or the system itself? As AI systems approach higher levels of autonomy

and sophistication, debates about the moral status of artificial agents become increasingly relevant. Should an artificial system be capable of achieving consciousness or genuine understanding, as has been proposed by proponents of strong AI, ethical considerations would be required to extend to the treatment of such entities.
It is imperative to consider the broader societal impacts of these measures. The advent of AI technologies has the potential to transform labor markets, intensify social inequalities, and reshape political structures. The notion of ethical foresight necessitates not only the implementation of technical solutions but also the facilitation of inclusive dialogues involving a diverse range of stakeholders, thereby guiding the responsible development of AI.
In view of the challenges identified, the philosophy of AI must be understood as a dynamic and evolving field. It is imperative that this philosophy critically engages with both the possibilities and the perils that AI brings to human society.

6 Closing thoughts

The question of whether AI can truly think or understand remains at the forefront of contemporary philosophy. While AI systems demonstrate remarkable functional abilities, significant differences persist between artificial and human cognition, particularly regarding consciousness, embodiment, and subjective experience. Functionalist theories offer a compelling framework for bridging this gap, yet the philosophical challenges posed by phenomena such as intentionality and conciousness remain unresolved.
The essay has argued, drawing on Rapaport's analysis, that internal representations and functional organisation are crucial to any serious account of understanding in AI. The ethical and societal implications of increasingly autonomous systems extend beyond theoretical debates, however, and careful reflection on the responsibilities involved in creating and deploying artificial agents is required.
As AI continues to evolve, philosophical inquiry will play an essential role in guiding its development. In this way, technological progress will be matched by ethical foresight and conceptual clarity.

References

1. Rapaport, W.J.: Philosophy of Computer Science: An Introduction to the Issues and the Literature. State University of New York Press (2005)
2. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson, 4 edn. (1995)