Philosophy and Large Language Models

Bob Fu, 2025

Summary

The project is motivated by a long-standing philosophical concern: can AI systems think, and how do we determine? Philosophers such as Hubert Dreyfus, John Searle, and Ned Block have provided critiques of early AI paradigms, particularly symbolic AI, by arguing that rule-based systems do not amount to real intelligence. As the paradigm shifts towards connectionism, represented by modern machine learning, these philosophical challenges have evolved.

Purpose

My summer research focused on the intersection of philosophy, cognitive science, and artificial intelligence (AI), exploring how the current paradigm of AI relates to philosophical issues about mind, language, and knowledge. The primary goal was to conduct background research and identify a promising direction for my philosophy honors thesis.

Methodology

The research utilized a combination of literature review and theoretical analysis. First, I conducted an in-depth examination of both classical and contemporary philosophical works relating to AI, ranging from Alan Turing's theory on the possibility of machine intelligence to philosophers who denied such possibility, such as Hubert Dreyfus, John Searle, and Ned Block. Additionally, I reviewed philosophers who believe smarter-than-human AI will be a possibility in the near future and have written about how controlling advanced machine intelligence will be difficult.

In parallel, I reviewed the technical literature on AI, from the foundations of neural networks to transformers, the architecture underlying LLMs. I specifically explored how these models operate, focusing on key concepts such as tokenization, embedding, backpropagation, the attention mechanism, and reinforcement learning from human feedback that enable these models to process and generate coherent language. In addition, I explored the major directions in technical safety research, such as mechanistic interpretability and human value/preference alignment. I aimed to explain these concepts in accessible terms while maintaining philosophical rigor in analyzing their significance.

Findings

It seems that some classic philosophical criticisms of AI have become less relevant now that the AI research paradigm has shifted away from building Strong AI, in large part due to the weight of those criticisms. Very few researchers believe that strong AI will be realized, and most have the more modest goal of realizing artificial general intelligence (AGI), which is envisioned to be inferior to strong AI, without consciousness, embodiment, or moral agency.

There is a distinct lack of engagement from philosophers about ongoing developments of AI and emerging empirical evidence. Most contemporary philosophers are polarized in that while some maintain machine intelligence impossible (e.g., Noam Chomsky), others have advocated for the dangers of advanced machine intelligence(e.g., Nick Bostrom). I am aligned against these two positions, and I plan to argue for the middle ground between them in my honors thesis through a philosophical analysis of LLMs.

LLMs are philosophically interesting for two reasons. First, they exhibited unprecedented competence in several general domains, in which strong performance has been historically considered to be exclusive to human intelligence. Second, the success of LLM was haphazard because its development was theoretically unexpected. Virtually no one expected such a degree of competence. Taken together, I think there is a mystery about LLM property and behavior that requires explanation and further philosophical engagement. Since no existing literature adequately addresses this mystery, it would be valuable to lay the groundwork for future research.

Faculty Mentor: Scott Sehon
Funded by the Surdna Undergraduate Research Fellowship