Tejas Kamtam

305749402

Topic: The AI Alignment Problem

References (not sure that I will use these for sure, but some starters):

- Strickland, E., August, 2023, OpenAI's Moonshot: Solving the AI Alignment Problem; IEEE Spectrum, https://spectrum.ieee.org/the-alignment-problem-openai
- Ngo, R., September, 2020, AGI safety from first principles: Alignment; AI Alignment Forum, https://www.alignmentforum.org/posts/PvA2gFMAaHCHfMXrw/agi-safety-from-first-principles-alignment
- Bostrom, N., 2016, Superintelligence: Paths, Dangers, Strategies; Oxford, England, Oxford University Press, 390 p.
- Ord, T., 2020, The Precipice: Existential Risk and the Future of Humanity; New York City, N.Y., U.S., Hachette Books, 480 p.
- Christiano, P., 2019, Current work in AI alignment; Center for Effective Altruism, EA Global: Bay Area, San Francisco, C.A., U.S., https://youtu.be/-vsYtevJ2bc?si=8XDpTGLsdIK4ne8d

The AI Alignment problem is perhaps the most critical issue of the current century. With the rate of AI advancement year-on-year, Artificial General Intelligence (AGI) is right around the corner and will likely pose an existential threat to humanity if misaligned. The alignment problem is the issue of certifiably ensuring AI/AGI understand and innately behave with human values in mind when making decisions toward their objectives. However, this problem is far more insidious than it may appear. Questions of "How do we define human values?," "How do we know we're not being deceived?," and "Is the AI just cooperating temporarily to achieve some nefarious task in the future?" are just a sample of the many genuine concerns that must be answered before the advent of transformative AGI.

To this end, many scientists and researchers have already begun to answer some of these questions - usually resulting in more questions. Paul Christiano (OpenAI), Jan Leike (Anthropic), and Eliezer Yudkowsky (MIRI) are a few of the top names in the AI safety landscape, each ensuring their respective organizations aid the development of safe and robust AI. Their research has divided the alignment problem into two major topics: inner and outer alignment (usually referred to by their contrapositive, "misalignment"). Outer alignment relates to agents making decisions/actions that are perceivably aligned with human values. Inner alignment, on the other hand, considers whether the intentions and thought process (usually termed "chain-of-thought") of the agent that came to its decision/action are truly aligned. These two topics have resulted in a handful of AI safety subfields to tackle the overarching problem. Specifically, capabilities and control evaluations (currently conducted by many startups including MIRI[1], Redwood Research, and FAR AI; personal communication, 2024) to determine whether an AI is outer misaligned and mechanistic interpretability research to lift the veil over the black box shrouding our understanding of LLMs to determine inner alignment (a few larger organizations pursuing this problem include Google DeepMind and Anthropic; personal communication, 2024). However, there is still much to be discussed: timelines, the possibilities of superintelligence, challenges in the field outside of research, setbacks, etc.

[1] Machine Intelligence Research Institute