Obelisk is a research lab focused on brainlike AGI safety and the development of brainlike AGI more generally. They think that there is a good chance that the first AGIs will be neuromorphic. Aligning brainlike AI comes with a different set of opportunities and challenges than other development paths, and Obelisk is studying computational neuroscience and human goal formation to prepare and explore the options.

One threat model they think about is misaligned model-based RL agents, and two possible paths to alignment they're looking at are "Controlled AGI"[1] and "Social-instinct AGI"[2].

The main crux for their agenda is whether transformative AI will be brainlike, and whether it's possible to sufficiently align brainlike AI. They are relatively agnostic on the exact failure modes caused by unaligned brainlike AGI.

Interested in helping out?

Making progress on Obelisk's agenda requires advanced knowledge of neuroscience and related disciplines. A strong grounding in machine learning is an asset here, but being a top-level ML engineer or ML research scientist is not required. They hire both within and outside academia.

Related

- What approaches are AI alignment organizations working on?
- What safety problems are associated with whole brain emulation?
- How would we align an AGI whose learning algorithms / cognition look like human …

[1] Automatically assessing the AI's thoughts and enforcing conservatism in value extrapolation
[2] Building an AI which deeply feels something like an idealized form of kinship and love for humanity, by understanding the computational structure of these drives in humans.