Anthropomorphic Algorithms for AI Safety

Tom Wallis

Anthopomorphic algorithms

for a long time, sociotechnical systems analysis within computing science has been developing formalisms of human-like traits, such as trust and comfort. These allow an intelligent agent to interact with other agents in its environment in measured, cautious ways; they might be used, for example, to decide whether it should accept information from another agent if its behaviour is becoming erratic (or to discard previous data which is no longer "trustworthy").

These anthopomorphic algorithms are undergoing continual improvement[1,2], but two problems remain untouched:

1. Various different anthropmorphic algorithms have been developed, but none have been combined into a system with several traits.

   For example, an algorithm might be designed where an agent's ratings of trust and comfort in a given scenario influence each other — not unlike a human's lack of trust in an agent making it less comfortable with certain situations.

2. Lots of philosophical questions arise when developing anthopomorphic agents. There are questions in roboethics, such as the morality of creating an intelligent agent which might exhibit racial bias after its training. There are also questions in machine ethics, involving ethical decisions that an agent might make, and how they can be affected by trust and comfort.

however, one exciting unexplored problem is that of AI safety: could anthopomorphic algorithms give humanity an edge in developing friendly AI? If we can limit its space of mind[3], perhaps we can limit potential damage from an artificial agent. Perhaps developing anthropomorphic agents allows us to reason better about how an artificial agent can be unfriendly, allowing us to better predict catastrophies related to the agent turning malicious. Perhaps such a method is provably inadequate for solving the problem of AI safety. In any case, further research is required to determine the method's efficacy.

[1] Seifeddine Kramdi. A modal approach to model computational trust. *PhD Thesis*, 2015. URL `https://tel.archives-ouvertes.fr/tel-01328169`

[2] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. An approach to computational social trust. 27:113–131, 2014. doi: 10.3233/AIC-130587

[3] Murray Shanahan. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds.* Oxford University Press, 2010

Implications and work for Intelligent Agents

My suitability