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 some problems remain unexplored. For example, while various different anthropmorphic algorithms have been developed, 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.

[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

However, one exciting unexplored problem is that of AI safety: could anthopomorphic algorithms give humanity an edge in developing friendly AI? For example, 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. In any case, further research is required to determine the method's efficacy, and to explore the practical applications of these anthropomorphic algorithms.

[3] Murray Shanahan. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press, 2010; and Aaron Sloman. The Structure of the Space of Possible Minds. pages 35–42, 1984

Responsibility and its implications

As a masters student at Glasgow University, my current research is in developing a computational formalism of "responsibility". This formalism would be the first algorithmic definition of an agent's responsibility, and fits the above problems perfectly. The formalism is similar to current trust and comfort models, making it easy to integrate with existing frameworks into an agent with several anthropomorphic traits. Most interestingly, an intelligent agent with a concept

of responsibility is useful to analyse from the perspective of AI safety in a way that trust and comfort models are less suitable.

As a result of the wealth of literature on responsibility for human agents, much work can be done to teach artificial agents to act in responsible ways; either in developing machine learning algorithms which tune the parameters of an agent's feeling of responsibility, or in imposing a strict sense of repsonsibility on that agent. The notion that computational responsibility might improve agent corrigability[4] is an exciting prospect – perhaps a "rogue" AI could be tamed through its social senses, much like an ordinary person might be when acting out. Another possibility is that of avoiding Reward Hacking[5]. When anthropomorphic constructs like responsibility are combined with uncertainty in the return value of a reward function, we can more readily rely on the intelligent agent to operate "responsibly" such that the uncertainty is reduced by proper action. The responsibility formalism developed should associate responsibility not only with a goal to achieve, but a set of possible actions which represent the "responsible" path to achieving the goal. One might leverage this to steer an intelligent agent's actions toward the responsible, even when myriad more are available to it.

[4] Stuart Armstrong, Nate Soares, Benja Fallenstein, and Eliezer Yudkowsky. Corrigibility. *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016. URL `http://arxiv.org/abs/1606.06565`

I PROPOSE that the breadth and practicality of this work represents a substantial addition to the current literature on intelligent agents, and that the introduction of computational responsibility to the growing arsenal of anthopomorphic algorithms represents a turning point in the relevance of anthropomorphic algorithms to philosophical literature. At a stretch, anthropomorphic algorithms may even represent a new area in the study of artificial intelligence safety, which promises to advance literature for computing science, and presents interesting moral repsonsibility and machine ethics collaborations with philosophical research.

My suitability

GIVEN MY EXPERIENCE developing this computational responsibility formalism, I am uniquely equipped to begin the research to be done. The formalism I have designed has been purposefully created with a philosophical foundation in mind, drawing from work by P.F. Strawson[6] and Ben Colburn, but has equal foundation in computational disciplines such as machine learning and sociotechnical systems modelling[7]. In addition, the breadth of the work to be done makes it ideal for pursuit as a PhD project.

[6] P.F Strawson. Freedom and resentment. *Proceedings of the British Academy, Vol. 48*, 1960

[7] Ian Sommerville. Models for responsibility assignment. In *Responsibility and Dependable Systems*, chapter 10

My other experience in research falls into two fields.

The first, sociotechnical systems, involves analysis of complex systems of people and the technology they interact with; my work was to create a modelling system which allowed simulation of sociotechnical workflows without barriers to entry like understanding of a domain-specific modelling language. These models were then dynamically altered during runtime by a code fuzzer I designed and implemented, which injected human-like variance into the model, allowing for accurate simulations of human behaviour using simple modelling techniques. This research won an award for "Best Software Product" for my year.

A paper on this work is currently being developed, which can be found at http://bit.ly/2gi4GDo. My original dissertation can be found at http://bit.ly/2fYbcgy.

Other research I pursue lies in the field of experimental storytelling. Project Albert explores the possibility that improvisation of children's bedtime stories might be made easier and more accessible by use of design patterns, which generalise complex ideas by turning them into interrelated rules which are defined semantically. A full suite of design patterns have been developed, with example stories which follow the patterns. The intention is to provide a framework for parents to share stories with their children which have some sentimental value as a result of the improvisation and random aspect, as well as to provide a way for parents to connect through stories when they cannot necessarily afford books to read from.

Information on Project Albert can be found at http://projectalbert.net/.

An essay on the work so far and its efficacy is currently under development, which can be found at http://bit.ly/2fZvggr.