Anthropomorphic Algorithms for AI Safety

Tom Wallis

AI Safety and Existential Risk

A general artificial intelligence, should it be constructed, would be very dangerous. There are many reasons for this. For example, a recent paper discusing an artificial intelligence's corrigibility[1] demonstrates an issue where a general artificial intelligence may resist human control. If this agent was dangerous — which is probable — then the problem of corrigibility becomes very important indeed.

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

Other problems exist, too. We are aware of some concrete problems in AI which experts can work on solving today[2] — one particularly interesting example is that of Reward Hacking, where an agent might "cheat" its reward functions in order to achieve its goals. While this could be harmless or inconvenient at best (a cleaning robot, say, which shuts its eyes and believes no mess exists because it can't see any), it could be devastating at worst. Unprecidented action which *technically* achieves goals but inadvertantly causes other problems (or immediate harm) could be cataclysmic when done by a sufficiently intelligent agent.

[2] 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`

In my research this year on computational responsibility formalisms — algorithms which imbue an intelligent agent with a sense of "responsibility" as it chooses actions to achieve its goals — I believe I have found an interesting opportunity to solve these problems using what I term "anthropomorphic algorithms"

Anthropomorphic algorithms and Philosophy: solving problems

While there are computer science researchers attempting to solve these problems using algorithmic techniques, philosophical work can see less implementation as a result of its often metaphysical nature.

However, this doesn't have to be the case. Michael Devitt's work in experimental semantics[3] is a shining example of philosophical research which is backed by data and concrete, repeatable examples. Computational responsibility, and anthropomorphic algorithms in general, afford philosophical theory another avenue to test theories by.

[3] Michael Devitt. Experimental semantics. *Philosophy and Phenomenological Research*, 82(2):418–435, 2011

Indeed, anthropomorphic algorithms provide other problems for philosophy to solve; the claim of imbuing a computer with human traits is a contentious one, and as computer science, sociology and psychology continue to refine formalisms of ordinarily human traits, the necessity of philosophical literature on the topic increases proportionally.

Unfortunately, the problem of AI safety — while one of existential risk — is also one requiring collaboration between various fields, including philosophy, computer science, psychology, political science and others. Fortunately, anthropomorphic algorithms provide a framework for interdisciplinary research between all of these fields.

I propose that a significant body of philosophical literature stands to be written on the subject. Particularly, I am excited to investigate the impact of anthropomorphic algorithms on the corrigibility of intelligent agents, as well as their application to the solution of the reward hacking problem. Introducing human-like traits to an artificial agent may make it controllable via indirect means. It could also be used to demonstrate artificial agents which are capable of reward hacking, but unwilling to act on this ability due to an ingrained sense of responsibility. I am also interested in researching the problems in roboethics and moral responsibility arising from the introduction of anthropomorphic agents.

An interesting political question related to general artificial intelligence is that, should we have an intelligence of roughy equal to that of humans, this introduces the social quandry of rights. Intelligence seems to be the important factor in allocation of rights: some countries award rights to animals like dolphins and intelligent apes, but less intelligent animals are less often afforded this thought.

My suitability

Given my experience developing the first computational responsibility formalism, I am uniquely equipped to begin the proposed research. The formalism I have designed has been purposefully created with a philosophical foundation in mind, drawing from work by P.F. Strawson[4] and Thomas Scanlon[5], and an equal foundation in computational disciplines such as machine learning and sociotechnical systems modelling[6].

My other experience in research falls into two fields.

The first, *Project Albert*, a research project in experimental storytelling. Not unlike Michael Devitt's work on experimental semantics, Project Albert seeks to employ systems design techniques to the improvisation of childrens' bedtime stories. Aside from the societal and practical benefits, this was an experiment with the intention of taking a humanities subject — here literature — and applying a practical experimental technique to it. I hope to apply a similar mindset to philosophical work on artificial intelligence. I believe this holds

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

[5] Thomas M Scanlon. Justice, responsibility, and the demands of equality. 2006

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

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

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

great value for the field: it is imperative that philosophical work on AI safety is carried out, and that this research be made testable and practical will ease interdisciplinary research with another vital field for the area, computing science (in which I am also well versed).

THE SECOND research experience to note is work done in sociotechnical systems modelling: the analysis and simulation of complex systems of people and the technology they interact with. Naturally, artificial agents (particularly anthropomorphic ones) interacting in the day-to-day world of humans is a prime example of a sociotechnical system on our cultural horizon. My research was on developing a new modelling framework for sociotechnical systems, which injected human-like variance to a model of a system so as to properly simulate what happens when introducing unpredictable human error in a model. While my masters in computing science lends itself to understanding the nuances of AI development itself, my experience modelling people with technology — and modelling complex systems with human traits — lends itself to a great deal of understanding of the impact of intelligent agents in the wider society, meaning I have a broad range of perspectives to employ in philosophical research on the topic.
With this experience in mind, I am certain I am the ideal candidate to carry out the proposed work.

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

This work won the prize for "best software product" for my year's honours projects.