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] discusses scenarios 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.

Other similar problems exist, too. We are aware of some concrete problems in AI which experts can work on solving today[2] — one particularly interesting issue is that of Reward Hacking, where an agent might "cheat" its reward functions in order to achieve its goals. While this could be harmless on 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 algorithm.cataclymr

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.

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

[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`

[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 continues 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, and their application to the solution of the reward hacking problem, as well as the ethical concerns regarding the development truely anthropomorphic intellect.

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, sociotechnical systems, involves analysis of complex systems of people and 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. I believe that my experience researching complex systems of this nature natually lends itself to existential risk research.

A paper on this work is currently being developed, which can be found at http://bit.ly/2gi4GD0. My original dissertation can be found at http://bit.ly/2fYbcgy.

[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

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.