

Anthropomorphic Algorithms for AI Safety

Tom Wallis

AI Safety and Existential Risk

A GENERAL ARTIFICIAL INTELLIGENCE, should it be constructed, would be very dangerous. This is rather well documented. For example, a recent paper discussing an artificial intelligence's corrigibility¹ 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.

OTHER PROBLEMS exist, too. We are aware of some concrete problems in AI which experts can work on solving today² — 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. Unprecedented action which *technically* achieves goals but inadvertently causes other problems (or immediate harm) could be cataclysmic when enacted by a sufficiently intelligent agent.

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 are algorithms which simulate a social human trait in an artificial agent. Examples of anthropomorphic algorithms already widely researched would be ones termed “computational trust”, versions of which are now several years old³, but rarely researched outside of sociological and sociotechnical research. I believe that an application of these algorithms lies in researching and implementing philosophical theories of AI safety.

Anthropomorphic Algorithms and Philosophy: Solving Problems

WHILE THERE ARE computer science researchers attempting to solve the noted AI safety problems algorithmically, philosophical work frequently produces fewer practical results due to its often metaphysical nature.

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 Stephen Paul Marsh. Formalising Trust as a Computational Concept. *Computing*, Doctor of(April):184, 1994. DOI: 10.2165/00128413-199409230-00010

However, this doesn't have to be the case. Michael Devitt's work in experimental semantics⁴ is a shining example of philosophical research which is backed by data and concrete, repeatable examples. Computational responsibility, and anthropomorphic algorithms in general, afford other avenues to test philosophical theory by.

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.

THE PROBLEM OF AI safety — while one of existential risk — is also one requiring collaboration between various fields, including philosophy, computer science, psychology, political science⁵ and others; this collaboration can be difficult to organise, particularly when lacking a common frame of reference such as jargon or theory. 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.

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. Strawson⁶ and Thomas Scanlon⁷, and an equal foundation in computational disciplines such as machine learning and sociotechnical systems modelling⁸. Therefore, its design lends itself to the testing of philosophical theories while relating to computer science research on artificial intelligence rather nicely. The formalism is inspired in its design by Marsh's seminal model of computational trust⁹, which

⁴ Michael Devitt. Experimental semantics. *Philosophy and Phenomenological Research*, 82(2):418–435, 2011

⁵ An interesting political question related to general artificial intelligence presents itself: should we have an intelligence roughly equal to that of humans, this introduces the social quandary of rights. Intelligence seems to be the important factor in the allocation of rights. Some countries award rights to animals like dolphins and intelligent apes, but less intelligent animals are less often afforded this thought.

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

⁷ Thomas M Scanlon. Justice, responsibility, and the demands of equality. 2006

⁸ Ian Sommerville. Models for responsibility assignment. In *Responsibility and Dependable Systems*, chapter 10

⁹ Stephen Paul Marsh. Formalising Trust as a Computational Concept. *Computing*, Doctor of(April):184, 1994. DOI: 10.2165/00128413-199409230-00010

itself was inspired by mathematics, psychology and sociology — an intention of its design was its inter-disciplinary potential as a research tool. The formalism of responsibility, then, follows in Marsh's ideological footsteps, but with a philosophy-centric focus.

ASIDE FROM ANTHROPOMORPHIC ALGORITHMS, I have other research experience relevant to the AI safety work proposed.

One example of this is the experimental literature project I run extra-curricularly, *Project Albert*¹⁰. Project Albert is an application of systems design techniques to children's bedtime story improvisation. I am particularly interested in experimental techniques for humanities research, much in the vein of Michael Devitt's experimental semantics¹¹ — I look forward to applying this mindset to philosophy research also.

Another relevant research project would be my sociotechnical systems modelling project for my honours year¹². This project applied a novel modelling approach for sociotechnical systems to a new system for injecting variance into those models — both of which were my own design¹³. The project then used these new approaches to insert human-like mistakes in the artificial agents modelled. Sociotechnical modelling, as well as and the representation of human traits in these models, naturally lend themselves to the proposed research.

TAKING THIS EXPERIENCE into account, and having an intricate understanding of the development and design of anthropomorphic algorithms, I am certain I am an ideal candidate to undertake the essential AI safety research pertaining to this novel technique. Not only have I designed myself the responsibility formalism which gives rise to a new method to tackle problems of reward hacking and corrigibility, but I have also undertaken award-winning research representing human traits in the past, and currently research experimental humanities techniques which provide an ideal background to begin experimental philosophy research. I am excited to begin the philosophical work, and doubtless that this new work it permits will be not only vast, but deeply fascinating.

Thank you for your consideration.

Cordially,
Tom Wallis.

¹⁰ Information on Project Albert can be found at <http://projectalbert.net/>.

¹¹ An essay on the work so far and its efficacy is currently in early development, and can be found at <http://bit.ly/2fZvggr>.

¹² This work won the prize for "BEST SOFTWARE PRODUCT" for my year's honours projects.

¹³ A paper on this work is currently being developed, and can be found at <http://bit.ly/2gi4GDo>.