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 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. Unprecedented action which *technically* achieves goals but inadvertently causes other problems (or immediate harm) could be cataclysmic when enacted 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 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 decades old[3], 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.

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

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.

However, this doesn't have to be the case. Michael Devitt's work in experimental semantics[4] is a shining example of philosophical research which is proven through 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.

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

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[5] 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.

[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.

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, and have a working paper detailing a technique which allows just this. The method presented is unusual, in that it constrains behaviour through analysis with that of humans — in applying anthropomorphic algorithms — rather than tweaking aspects of the AI engineering itself. 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, for example, responsibility.

Proposed Work

Specific work I intend to work on during a DPhil at Oxford University involves developing further anthropomorphic formalisms, performing interdisciplinary research in order to assert the applicability of a formalism to the problem of AI safety, and further developing the technique as it pertains to both AI Safety and theories of mind. The formalisms developed would be used to further the relevant philosophical and computational technique of my own design, which presents new ways to analyse the space of possible minds and

approach AI safety through behaviour constraint.

Problems to be solved in this work involve the application of the theory to a broader range of theories of mind, as well as asserting a-priori its application to AI Safety and its validity as an empirical philosophical tool.

Finally, the research might involve an analysis of the implications of the technique's abstraction of behaviour over the space of possible minds. This would provide a stronger argument backing up the theory, alongside a technique by corollary for creating abstractions over the space of possible minds, thereby avoiding problems with the space's indeterminate structure. As the space of possible minds is an important and common concept, gaining analytical value from the space without specifying its structure is certain to have utility far beyond what is developed over the course of the DPhil, and would constitute a significant advancement in the field on its own.

Personal Statement

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[6] and Thomas Scanlon[7], and an equal foundation in computational disciplines such as machine learning and sociotechnical systems modelling[8]. 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 trust[9], which 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.

[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

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 semantics[11] — I look forward to applying this mindset to philosophy research also, particularly as the proposed research is fundamentally interdisciplinary.

[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.

Another relevant research project would be my sociotechnical systems modelling project for my honours year[12]. 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 design[13]. 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.

[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.

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. The shift in my current research to a philosophical focus indicates that the locus of my own attention should shift accordingly, and Oxford University — with a rich AI Safety and interdisciplinary research community — provides the most fertile intellectual ground for this work to flourish. Indeed, the achievement of complete drafts of a working paper in philosophy[14], having undertaken only half of the research so far, demonstrates both my aptitude for philosophical study and the very real potential for prolific development of a new angle in AI Safety and theories of mind.

I am excited to begin the philosophical work, and doubtless that the opportunities it permits will be not only vast, but deeply fascinating, and of great impact.

Thank you for your consideration.

Cordially,
Tom Wallis.

[14] Tom Wallis. On anthropomorphic algorithms. Paper currently in progress, recent draft can be found at https://github.com/probablytom/PhD-Project-Proposal/blob/master/ai_safety/cs/proposal.pdf., 2016