Anthropomorphic Algorithms and AI Safety

Tom Wallis

January 6, 2017

Developing Anthropomorphic 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").

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[1], but rarely researched outside of sociological and sociotechnical research.

These anthropomorphic algorithms are undergoing continual improvement[2], but some problems remain unexplored. For example, while various different anthropomorphic algorithms have been developed — particularly for simulating trust — none have been combined into a system with several traits. Research on anthropomorphic algorithms fails to see attention in some interdisciplinary areas: work on AI safety, while a largely philosophical field, does not make use of literature on anthropomorphic algorithms despite the fact that human-like algorithms may help to solve known concrete problems[3] and would help to limit the space of possible minds an agent can inhabit.[4] An agent which can trust or feel responsible might even have implications in the area of safety critical systems, where solving issues like the Trolley Problem are important as automated dangerous equipment — such as cars — begin to become a reality.

It's plain to see that lots of work is yet to be done in the field. Applications of these algorithms are exciting — but the field itself is growing, too. My own masters research extends literature on computational trust formalisms to simulate computationally theories of personal responsibility. I believe that this particular research provides opportunities to further the state-of-the-art in groundbreaking ways, and to solve important concrete problems in AI safety. An

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

[2] Seifeddine Kramdi. A modal approach to model computational trust. *PhD Thesis*, 2015. URL `https://tel.archives-ouvertes.fr/tel-01328169`; and Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. An approach to computational social trust. 27:113–131, 2014. doi: 10.3233/AIC-130587

[3] 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`

[4] Aaron Sloman. The Structure of the Space of Possible Minds. pages 35–42, 1984

in-progress paper I am currently writing [5] highlights the great degree of practical detail that remains unexplored in this area, and its importance to behavioural AI Safety solutions.

[5] Tom Wallis. On anthropomorphic algorithms. Paper on the philosophical utility and implication of anthropomorphic algorithms on theory of mind and AI Safety, currently in progress. Recent draft can be found at https://github.com/probablytom/PhD-Project-Proposal/blob/master/ai_safety/cs/proposal.pdf., 2016

Future work

Computational Responsibility is an important formalism for the development of behaviour-centric AI Safety algorithms. Responsibility as a human behaviour has a deep importance for the way we interact with other people: the understanding, for example, that we have been delegated a task with a certain degree of importance. However, responsibility as a concept can describe other things too: for example, a social responsibility not to litter is imposed partly by law, and partly by social convention. Different human agents experience this responsibility to different degrees, depending on cultural background, among other factors.

This is to demonstrate that responsibility formalisms, while very useful, are underdeveloped as a research field. Necessary research presents itself in a number of ways, including:

- More thorough and nuanced formalisms of responsibility
- Formalisms for sociological perspectives of responsibility
- Development of practical applications so as to showcase the potential impact of responsibility formalisms in particular to the wider research community
- Research into containers and other engineering techniques for anthropomorphic algorithms — such as responsibility — to be more easily implemented and integrated into intelligence projects

Note that most of these pertain specifically to responsibility formalisms. Anthropomorphic algorithms in general present a great, untapped research opportunity with far-reaching consequences.

Specific Work pursuing a DPhil

I propose that a body of work suitable for a DPhil at Oxford University can be found in researching nuances in responsibility formalism, combined with the analysis of anthropomorphic algorithms as a wider field of research in the form of engineering techniques for anthropomorphic algorithms. Particularly, research on engineering techniques which combine anthropomorphic algorithms' effects on

an intelligent agent would be developed. These would be tested in two ways:

1. Combining the results of responsibility, trust, and comfort — formalisms which currently exist, and all centre around the allocation and discharge of goals
2. Combining different types of responsibility formalism, which model different types of nuance in the concept of responsibility. In this way, a more human-like formalism of responsibility might be borne from the covering of a wider subspace of the space of possible responsible situations.

To complete the latter research, further responsibility formalisms would be developed. Upon completing this work, advances in the study of anthropomorphic algorithms would allow for serious engineering and research in combining human-like traits, as well as more complete literature on responsibility formalisms and the development of related interdisciplinary study. Anthropomorphic Algorithms themselves would have the potential to grow from a modelling and research field, rather than being considered a curiosity of sociotechnical research. It would also have impact on AI Safety research, and philosophical fields such as theory of mind.

My suitability

Having developed the only existing anthropomorphic algorithm for responsibility formalism, as well as having experience in sociotechnical systems research, I am very well suited to pursue this particular research project. My honours year dissertation[6] developed new sociotechnical modelling techniques for introducing human-like variance to workflow modelling with minimal overhead, so as to create human-readable workflow models which dynamically introduced human error during runtime. The models, and the dissertation project, were pure Python code. A paper on the work is currently in progress.[7]

Following this, my MSci research project is to create a responsibility formalism — all references here to responsibility formalisms are my own work. I am excited to continue the work, and know that once properly developed as a research field, responsibility and its associated anthropomorphic algorithms present a deep and prolific research opportunity.

Given my familiarity with the software engineering involved,

[6] My honours year dissertation received an A2, and the award for the best software engineering of my year — despite it being a research project specifically instead of the engineering projects it competed with.

[7] Tom Wallis and Tim Storer. Simulating variance in socio-technical behaviours using executable workflow fuzzing. Paper on the dynamic human-like variance of a workflow model, complete with a suitable modelling framework in Python, currently in progress. Recent draft can be found at http://github.com/probablytom/fuzzi-moss-paper., 2016

sociotechnical systems in general, and being the first to specialise in the responsibility formalisms the proposed work concerns — as well as having a deep understanding of the philosophical work involved[8] — I am certain that I am exceptionally qualified to undertake this research. I am excited to begin.

Thank you for your consideration.

Cordially,
Tom Wallis.

[8] Tom Wallis. On anthropomorphic algorithms. Paper on the philosophical utility and implication of anthropomorphic algorithms on theory of mind and AI Safety, currently in progress. Recent draft can be found at https://github.com/probablytom/PhD-Project-Proposal/blob/master/ai_safety/cs/proposal.pdf., 2016