Investigating Anthropomorphic Algorithms

Tom Wallis

Computing science has developed algorithmic formalisms of human-like traits, such as trust and comfort, for a little over two decades. These allow a computational agent to interact with other agents in its environment in more subtle, flexible 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"). However, these algorithms are largely studied in isolation. This proposal describes research to investigate the behaviour of combinations of these algorithms within populations of computational agents. It is hypothesised that equipping computational agents with a richer set of anthropomorphic traits lessen their vulnerability to attacks such as reward hacking[1].

Anthropomorphic Algorithms are algorithms which simulate a social human trait in an artificial agent. Examples of anthropomorphic characteristics that have received computational treatments include trust[2], reputation, comfort[3], and more recently responsibility[4]. These algorithms have largely been studied within the artificial intelligence and agent based research communities.

These anthropomorphic algorithms are undergoing continual improvement[5], but some problems remain unexplored. Exploration of the techniques within the scope of AI development might yield solutions to problems of corrigibility[6] and reward hacking[7] in AI safety research. The technique implies that human behaviours can apply to non-biological "minds", useful to philosophy of mind research[8].

There are numerous examples of formalisation of human traits within computational models, both in research and practice. An example would be reputation formalisms developed by companies such as eBay, which users interact with as seller ratings. Variants of these formalisms are employed as the basis of more complex models, such as Eigentrust[9]. Eigentrust's reputation model — based on eBay's — is then used as a basis for a more complex trust model. While this model has little basis in behavioural science, there exists strong empirical evidence that use of such formalisms can lead to much stronger peer-to-peer networks with "trustworthy" interactions between peers.

However, practical use of these formalisms can be challenging, for a number of reasons. Testing these formalisms and their implementations reliably can be a particular difficulty, as they tend to have no

[1] 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

[2] Stephen Paul Marsh. Formalising Trust as a Computational Concept. *Computing*, Doctor of(April):184, 1994. doi: 10.2165/00128413-199409230-00010; and Cristiano Castelfranchi and Rino Falcone. Social Trust: A Cognitive Approach. 2001

[3] Stephen Marsh, Pamela Briggs, Khalil El-Khatib, Babak Esfandiari, and John A. Stewart. Defining and Investigating Device Comfort. *Journal of Information Processing*, 19(7):231–252, 2011. ISSN 1882-6652. doi: 10.2197/ipsjjip.19.231. URL http://ci.nii.ac.jp/naid/110008508036/en/

[4] Tom Wallis. Investigating computational responsibility. MSci thesis, currently in progress., 2017; and Ian Sommerville. Models for responsibility assignment. In *Responsibility and Dependable Systems*, chapter 10

[5] 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

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

[7] 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

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

overarching framework for their implementation or testing[10]. While this is problematic for reliably carrying out research on anthropomorphic algorithms, the lack of established engineering guidelines for developing, testing and deploying these formalisms means that industrial use is also elusive.

Developing these guidelines is a complicated matter. In part, this is due to the nuanced nature of behavioural science. It is further complicated, however, by the different natures of the formalisms themselves. Some formalisms are developed so as to be rigidly logical[11], while others are primarily focused on behavioural traits and only then analysed algorithmically[12]. Developing architectures and guidelines for these models, therefore, requires an understanding of their nuance.

WITHIN THE REALM of computing science research, there is potential for anthropomorphic algorithms to be developed, that capture other human traits such as responsibility and loyalty. Other potential research could leverage other behaviours, such as stereotype adoption. Other research in this area is addressing the communication of computational traits via human interfaces[13], helping fields such as human-computer interaction to create more authentically human-feeling interfaces. Anthropomorphic Algorithms are an important area of research which can be expected to flourish in both computer science and interdisciplinary research.

This section of new work ends here

ONE PREREQUISITE TO THIS RESEARCH is to develop a software architecture which allows for the combination of multiple traits. Current methods for designing multiple traits would involve developing a single formalism which modelled multiple traits. However, an architecture which combined these traits would permit combining sociological and psychological theories which have already been formalised, without the additional work of creating an overarching formalism for every trait combination.

COMBINING SEVERAL TRAITS has a great deal of practical utility. To illustrate, one can imagine designing an interface to a mobile phone which takes into account a device's "feeling" of trust and comfort — two traits which have already been formalised into anthropomorphic algorithms.[14] A mobile phone might have a degree of trust in the person it identifies as using it; less trustworthy users might be prohibited from accessing more sensitive information, such as medical information stored in systems like Apple's HealthKit database.

The device might also have a degree of comfort which is dimin-

[10] Partheeban Chandrasekaran and Babak Esfandiari. A model for a testbed for evaluating reputation systems. *Proc. 10th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications, TrustCom 2011, 8th IEEE Int. Conf. on Embedded Software and Systems, ICESS 2011, 6th Int. Conf. on FCST 2011*, pages 296–303, 2011. DOI: 10.1109/TrustCom.2011.40

[11] Cristiano Castelfranchi and Rino Falcone. Social Trust: A Cognitive Approach. 2001

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

[13] FIND THIS FIND THIS. Find this find this. FIND THIS FIND THIS

[14] Stephen Marsh, Pamela Briggs, Khalil El-Khatib, Babak Esfandiari, and John A. Stewart. Defining and Investigating Device Comfort. *Journal of Information Processing*, 19(7):231–252, 2011. ISSN 1882-6652. DOI: 10.2197/ipsjjip.19.231. URL `http://ci.nii.ac.jp/naid/110008508036/en/`; and Stephen Paul Marsh. Formalising Trust as a Computational Concept. *Computing*, Doctor of(April):184, 1994. DOI: 10.2165/00128413-199409230-00010

ished when the user it identifies as using it begins acting erratically. One would expect the degree of trust it had in a previously trustworthy user to decrease as a result — perhaps its initial assessment of the user's trustworthiness was mistaken, or perhaps its identification of its user is incorrect. If erratic behaviour in this case decreased a feeling of trust, then trust and comfort are in some way linked.

One can imagine similar situations where trust might affect the device's degree of comfort, where switching from a trusted to an untrusted user might result in a sharp decrease in comfort. It is important to design effective and simple ways to model these and other situations, then, so as to make the engineering of these useful systems as simple as possible. Research — in a range of fields anthropomorphic algorithms touch, including and extending beyond computing science — would be affected too, as the construction and testing of anthropomorphic systems would be simplified if their creation can be simplified.

THE NEED for this system, which involves the integration of several traits, should be done via an agreed-upon software architecture and a formalised method for creating these formalisms. An anthropomorphic architecture would permit the combination of several systems, without the need to develop a new formalism for each combination.

A good architecture would also explore the possibilities for a general format of these formalisms. For a formalism to work in the architecture, it might need to adhere to certain requirements. This lends an opportunity to develop a suitable format for the formalisms themselves, providing guidelines for new formalisms to be constructed against and adding coherency to the growing inventory of formalisms being studied.

APPLICATIONIN INDUSTRY is one of the key advantages of a general architecture by which to combine these formalisms. With the weight of this additional engineering lifted, combined formalisms can begin to have impact in real-world products. The deployment of these formalisms would be safer, also, due to their testability within a properly engineered framework.

Tim: I've added this since we spoke 25/01.

The development of this architecture would incur a generalised architecture for single formalisms, too. An architecture would have to house an arbitrary number of formalisms which adhere to the same implementation details; a single formalism should be able to operate within this architecture without being combined. As a result, research and real-world deployment of single formalisms can be strengthened also. This would have a similar effect to testbeds already developed for reputation systems[15] — what differentiates this

[15] Partheeban Chandrasekaran and Babak Esfandiari. A model for a testbed for evaluating reputation systems. *Proc. 10th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications, TrustCom 2011, 8th IEEE Int. Conf. on Embedded Software and Systems, ICESS 2011, 6th Int. Conf. on FCST 2011*, pages 296–303, 2011. DOI: 10.1109/TrustCom.2011.40

work from the already developed work is that of implementation detail. Rather than developing workflow and classification systems, this architecture would have a software-engineering focus for maximal impact.

FURTHER ADVANTAGES include easing the early stages of developing a system that uses anthropomorphic algorithms, outside of implementing the algorithm itself. Choosing a formalism and testing it becomes simpler when that formalism is surrounded by a well-documented architecture, as replacing that formalism and testing different formalisms can be expected to entail less change to the code base and its structure. Developing with a series of integrated formalisms may become complex, but architectures which reduce coupling between the formalisms and still permit them to influence each other would permit more nimble alterations to the product's anthropomorphic components.

New work ends here.

I PROPOSE THAT this work is a suitably sized and impacting topic for PhD level research. I also believe that the project holds much value, due to the pressing need for the anthropomorphic architecture to be developed. I am also inspired, however, due to the fact that my own masters level research involves developing new anthropomorphic algorithms for responsibility, and that the existence of this architecture would permit exciting new research opportunities is an observation born of my own enthusiasm for the topic.

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[16] 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.[17] Sociotechnical modelling tackles similar problems to anthropomorphic algorithms, and my familiarity with computational models of human traits shows my familiarity with the field.

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 formalism will play a vibrant role in the creation of mixed-trait models.

[16] 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.

[17] 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

My familiarity with the state-of-the-art, and my proven ability to push the state of the art in anthropomorphic algorithms research, is a testament to my suitability in carrying out this important next development in the field.

MY CONFIDENCE IN MY FIELD is evidenced by my activity in the area. I have in-progress articles on the impact anthropomorphic algorithms can have in the philosophical arena, and will be talking on anthropomorphic algorithms' interdisciplinary potential at the "Let's Talk About [X]" undergraduate research conference in February 2017. My willingness to contribute to the field even in my spare time, and to speak publicly on the research as it stands, shows my dedication to research in anthropomorphic algorithms as a field. My confidence and experience in software engineering for sociotechnical systems, and my familiarity with anthropomorphic algorithms as a field, primes me to continue this research in a unique way.

Thank you for your consideration.