Engineering Standards for Anthropomorphic Algorithms

Tom Wallis

1 Proposed Approach

1.1 Abstract

The field of human-like computing — and the study of algorithms which mimic human behaviours especially[1] — is one of increasing importance in academic and industrial circles. Academic circles are increasingly looking to use the metaphor of anthropomorphic behaviour in information security, human-computer interaction, and fields tangentially related to computing science such as urban planning and smart city development.

Developing these anthropomorphic algorithms can be complicated, however. They require rigorous study in social sciences such as psychology, anthropology, and sociology, as well as the understanding of artistic studies such as philosophy. This added complication means that pursuit of human-like computing requires interdisciplinary study to effectively research and implement these systems. This complexity often results in simple models of one human behavioural trait, rather than more involved multiple-trait models.

This limitation is problematic: the complexity of the field, and its obscurity relative to other fields such as pervasive computation or quantum computing paradigms, mean that many researchers are not drawn to the field as a possible opportunity.

In this proposal, a research opportunity is described which can solve both of these issues by producing currently absent tooling and methodologies for the field. Once successfully completed, the tools and methodologies produced would reduce the interdisciplinary complexity of the field, and create jargon and tools which reduce the friction involved in undertaking this research from multiple angles. These tools and methodologies would, in turn, permit currently difficult-to-pursue research which engineers multiple-trait anthropomorphic algorithms. These models would strengthen the industrial utility of this field, as the utility of the models is compounded with more traits — garnering more interest in the field, and putting existing research to better use, though applications in smart cities, voice assistants, and more.

1.2 Problem Outline

Human-like computing, as a research field, has grown significantly in recent years — particularly with regards literature on trust formalisms. However, the field is unusual, in that the development of an anthropomorphic algorithm requires an understanding of not only computing science, but also social sciences: a formalism's accuracy depends on its psychological and sociological perspective. More anthropomorphic formalisms require an understanding of other fields, such as philosophy and ethnography also, as their modelling of human traits — and affect on our culture — is critical to understand during the model's creation. This interdisciplinary nature is one of the field's greatest strengths and most curious aspects.

It is also one of its greatest weaknesses. The requirement for a formalism to have a well-defined psychological and sociological model, as well as potential ethnographic and philosophical perspectives, so as to be implemented and evaluated by a computing science researcher, means that only particularly polymathematical researchers can undertake the research — assuming that it arouses their interest in the first place. The alternative, an interdisciplinary team who can perform the research with a shared understanding of different components of the formalism, has its own complications, communication can be hampered by the differences in different parties' jargons. Moreover, the aims of researchers with different backgrounds can differ: some fields, such as the social sciences, have an interest in modelling human activity accurately, but computing scientists and philosophers can find the models useful as a metaphor in their studies — as a thought experiment for philosophers, and in human-computer interaction for computing scientists.

Therefore, no suitable system currently exists for undertaking this research. Either a researcher adept in both computing science and social science is required, or a team with unusually good communication skills, each of the members of which should understand the (complicated) jargons of the others.

Moreover, this interdisciplinary disparity can create further tensions, as no guidelines exist on how these models should be implemented and tested. A team of differing backgrounds will naturally diverge on how a formalism should be evaluated, and its creation can be complicated by the lack of clear guidelines on its implementation and evaluation. A model of multiple human traits can quickly couple the many traits together, and should one trait need to be altered, this can ripple through a particularly sophisticated model to cause major setbacks. Given the difficulty communicating between team members, and the complexity of the project for a single researcher, these major setbacks should be expected at present.

[1] Henceforth referred to as "Anthropomorphic Algorithms".

Solving this problem has its own complexities. This particularly can be seen when analysing the intricate nature of psychological and sociological research on a single topic.[2]. The problem can be tackled however — as will be seen — by separating the engineering from the theory, and creating guidelines and tooling which simplify the model's creation and structure.

1.3 Approaching the Problem

To approach the problem of simplifying the engineering, one must first analyse what trends exist currently — a standard should fit the direction that the field is currently taking. Once undertaken, however, this analysis will lend itself to the creation of:

1. A methodology for formalism creation and evaluation.
   This should help the engineering effort to get out of the way of the research, being the definition and evaluation of a formalism of a trait.

2. Tooling to support this methodology.
   This will help drive adoption of the methodology, as well as ensuring that evaluable, well-engineered formalisms are easier to create — strengthening the case for industrial applications.

Methodology/Guideline Component

To create the methodology appropriate for solving the problem of the creation and engineering of these anthropomorphic algorithms, it will be important to analyse multiple aspects of the existing literature.

For example, it is critical that the methodologies and guidelines produced are in line with existing models. Moreover, it is vital that these methodologies and guidelines are suitable at a number of levels: they must support not only the engineering of a model, but the engineering of models with psychological, sociological, ethnographic or philosophical perspectives on their respective traits.

The methodologies and guidelines created would note only support the creation of a model, but would support the creation of several models, ideally of different traits. This would increase the degree to which different models can be compared, and can be used as the basis of other models.

Once these methodologies and guidelines exist, jargons around this framework can be made concrete, providing one jargon that all members of a research team investigating anthropomorphic algorithms and formalisms of human traits can learn. This would simplify the existing research, and solve some of the issues in interdisciplinary communication.

[2] As an example, consider the differences between Luhmann's approaches to trust [10] compared to Deutsch's [6] Deutsch believed that trust was inherently a perspective of an individual regarding the world, whereas Luhmann's perspective centred around the broader-scale sociological impact that trust has.

Tooling Component

Once the methodology and guideline component of work is complete, tooling for the system can be constructed which supports this as a specification. This tooling would take the form of engineering techniques, such as appropriate design patterns, as well as libraries which encourage the construction and design of a formalism according to the guidelines.

Ideally, these guidelines would simplify model construction so as to greatly reduce the technological barrier to entry; perhaps enabling social sciences researchers to implement the models themselves without a dedicated software engineer/computing science researcher. The feasibility of this goal is dependant on the results of an in-depth background survey, which would identify the complexity of the methodologies and guidelines, as well as whether one set of methodologies and guidelines can sufficiently cover all anticipated models.

An aspect of this tooling component would be demonstration that the tools would work; this can be done in two stages.

1. Re-implementation/redesign of existing models according to the methodologies and guidelines, built using the tools created.

2. An industrial proof-of-concept, using the tools to create models with the simplicity, reliability, and power to have commercial application.

1.4 Use Cases

One possible application of this work would be simpler methods for using anthropomorphic algorithms in passive authentication. This would build on already published work applying anthropomorphic properties to device security[4], strengthening the system and allowing the formalisms used to be evaluated more reliably. Evaluation becomes more reliable as the engineering of the anthropomorphic algorithm becomes more reliable, and can be tested against other similar anthropomorphic algorithms engineered using the same techniques.

Another potential application would be smart city development. Currently, many programmes exist for the development of smarter cities[12], yet anthropomorphic algorithms see little application in the space of urban planning. Many applications of anthropomorphic algorithms can be envisaged[18], resilient city development and anthropomorphically responsible emergency response being two examples.

Voice assistants in commercial electronics are becoming more popular, as can be seen from the rise in popularity of technologies such as Apple's "Siri", Amazon's "Alexa", and the Google Now assistant. From a human-computer interaction perspective, these technologies rely in part on a human-like interaction mechanism through the use of a human-like voice, and natural language parsing.

Designers can further the anthropomorphic metaphor used in many of these devices through the application of anthropomorphic algorithms — however, no well-founded engineering standard exists for this technology, preventing its broader commercial application. A potential use case would be the integration of anthropomorphic algorithms to these consumer electronics products.

2 Background

Related literature on anthropomorphic algorithms varies widely, due to the multitude of different traits which are formalised. However, the trait with the greatest degree of pre-existing literature is trust; this background survey will therefore cover the nuances of trust research particularly, as an example of the different natures of anthropomorphic algorithms which the methodologies and tools would need to support.

2.1 Trust

Many different approaches have been taken to trust modelling. For example, Marsh's seminal model of trust[11] is largely founded in psychology and sociology research; Eigentrust[8] instead uses trust modelling as a metaphor on which to base network security algorithms. This subsection will explore the myriad ways in which trust modelling can vary in its academic philosophy.

Marsh

Marsh's model of trust is the first specifically computational perspective on a human behavioural trait[3]. It draws on psychological and sociological work to formalise a general theory of trust, and creates a mathematical representation of this theory. The theory is then tested by evaluating it in application to reinforcement learning agents.

One particularly interesting notion Marsh provides is that Trust can be considered from three degrees of detail:

- Basic Trust
  This is an agent's general inclination to trust; from a human perspective, one might consider it the "trusting-ness" of the agent.

- General Trust
  This is an agent's inclination to trust another, specific agent. An example of general trust might be a student's degree of trust toward their thesis advisor — the student might trust the advisor completely, not at all, or somewhere in-between.

- Specific Trust
  This is an agent's inclination to trust another, specific agent to enact a task or complete some goal. An example of Specific Trust might be that a student might trust their supervisor completely to write a reference, but not to fly a Boeing 777.

In creating these degrees of detail, Marsh attempts to replicate the way that human beings trust when considering different aspects of a scenario. One can imagine asserting, "He's rather trusting", "She trusts her advisor", or "they don't trust each other to write a grant proposal" — each having its own related, yet distinct meaning. In integrating this notion into his model of trust, Marsh creates an algorithm which satisfies the literature he cites for both psychological and sociological perspectives on trust.

Worth noting is that Marsh's model permits graded degrees of trust for all of these types. That is to say, Marsh's model can associate numbers to its measurement of trust, and so weigh up different degrees of trust to make decisions.

C&F

Castelfranchi and Falcone[3] — abbreviated to C&F in popular literature — created their own model of trust, based on a logical formalism of social trust. Their formalism defines logical predicates which define the nature of trust in terms of confidence, dependence, and disposition, which are also defined by them in logical terms.

The degree of detail Castelfranchi and Falcone provide creates a simple method for calculating and assessing whether one agent trusts another, in a boolean fashion, in a way which scales to complicated multi-agent systems (or "MAS") effectively. This is more difficult in Marsh's model of trust, which is designed to measure degrees of trust from multiple agents' perspectives, and involves more complex calculation as a result.

C&F, however, has its own problems. For example, the boolean logical approach that C&F provide makes gradations of trust difficult to calculate. While they provide one possible approach at the end of their original paper, the elegance and simplicity of the non-graded approach is lost to the additional detail.

C&F in its non-graded form therefore has middling applicability in the real world; however, it has served as the basis of much further work[7], including more powerful modal logics.[9]

It is clear to see that, between even these two early models of only one trait, a significant degree of difference exists in the implementation detail of the two formalisms. Moreover, their philosophies with regards social sciences differ; Marsh taking an approach between sociology and psychology, C&F leaning more to the sociological aspects. However, both models are created with the intention of directing the behaviour of intelligent agents. Therefore, at least some similarities exist which might be used to construct methodologies and guidelines around.

[3] Birkhoff's early work on the mathematics of Aesthetics[2] being a possible exception, though Birkhoff's intention was not to create a computational model — "Aesthetic Measure" pre-dates the Church-Turing thesis by about three years.

Eigentrust

In comparison to the social sciences-oriented approaches taken by Marsh and C&F, Eigentrust[8] takes a very different approach. Eigentrust leverages a reputation-centric approach, basing its algorithms on star ratings often used by e-commerce platforms such as eBay and Amazon for rating their traders. The core of the algorithm rests on a very simple calculation of overall satisfaction:

$$s_{i,j} = sat(i,j) - unsat(i,j)$$

. . . where sat is a function representing the number of interactions agent i has had with agent j which they deem "satisfactory", and $unsat$ the function representing the number of interactions deemed "unsatisfactory". From this simple formula, Eigentrust calculates agent j's reputation from the perspective of agent i, and through a series of more complex transformations using linear algebra, arrives at a distributed ledger of reputation scores which takes into account the input from all agents.

Eigentrust is therefore different to Marsh and C&F's respective formalisms in important ways. Examples include:

- Networking agents together
  Agents under Eigentrust discuss each others' assessments of reputation with each other; to facilitate this, protocols for communicating reputation scores between these agents must be established. Unlike models created by Marsh and C&F, Eigentrust permits an agent to trust or distrust another agent by taking into account the interactions the other agent has had with third parties.

- Evaluation
  The intended application of Eigentrust is directing the behaviour of an intelligent agent, but unlike Marsh or C&F's models, Eigentrust is designed to be directly applied in areas such as network security. While C&F and Marsh can evaluate their models based on whether the actors behave in more "trusting" ways, for some metric of trust, Eigentrust is evaluated via its efficacy in protecting nodes on a network from unsatisfactory interactions (such as downloading viruses or receiving bad packets).

- Trust as a useful metaphor, rather than creating a realistic anthropomorphic model
  One can see from this difference in evaluation that Eigentrust is fundamentally an application of trust as a useful metaphor in security engineering. Eigentrust's goal is not a realistic representation of trust, but a more limited application of trust to fulfil a specific purpose.

These differences would, at first glance, indicate that differences between formalisms can be wide-reaching enough that one set of methodologies or guidelines for their research would not cover models for even one trait; however, while their implementation and application *do* vary considerably, Eigentrust's direction of agent behaviour indicates that some middle ground can indeed be reached. However, due to the broad spectrum of possible formalisms a set of methodologies and guidelines would need to cover, one must be particularly familiar with a range of formalisms to successfully cover this range. The work to be undertaken, therefore, is far from trivial, and would require a great degree of time to cover properly.

2.2 Reputation

Rather than using reputation to bootstrap a model of trust, some formalisms simply model reputation. REGRET[13] from Sabater & Sierra is one such formalism.

The perspective of REGRET is that reputation is the "opinion or view of one (agent) about something"[13]. This definition of reputation has no specific roots in psychology or sociology; however, REGRET makes use of social sciences research indicating that social structure and recency of events are important factors in an anthropomorphic agent's assessment of reputation. The authors' philosophy is that reputation is fundamentally an opinion; therefore, it is unsurprising that a main focus of this model is the successful modelling of an agent's opinion. Other factors assessed include "social reputation", being an agent's inclination to inherit opinions which other agents hold (where some social connection between the two exists).

REGRET lies in the middle of the spectrum carved by socially anthropomorphic models, such as Marsh's, and metaphorically anthropomorphic models, such as Eigentrust. Due to its practical nature, it is easy to evaluate and to implement in real-world use cases; however, it draws on social sciences research heavily enough to have a particular lean toward a socially centred model. This balance makes REGRET a good example of a middle ground which can be carved between models.

REGRET also presents an important use case of the guidelines and methodologies produced as a part of the proposed research. A model such as Eigentrust, in its original detail, is tightly coupled to its implicit model of reputation; future work might combine the reputation modelling provided by REGRET into Eigentrust's calculation of graded trust, so as to explore how Eigentrust fares when modelling responsibility differently. Current approaches to the problem of anthropomorphic algorithm design and engineering do not present elegant methods for combining the different aspects of Eigentrust in a modular, uncoupled way. However, with appropriate guidelines, methodologies and tooling, this limitation of the current status quo need not create more difficult future research; however, these methods produced must apply not only to Trust modelling, but to Reputation, a largely unrelated behavioural trait, also.

2.3 Responsibility

Similarly to Trust modelling, some varied work has been done in the pursuit of responsibility formalisms.

2.3.1 Deontic Logic

An early mathematical framework of responsibility can be found in the popular deontic logic[17]. Deontic Logic is a modal logic, similar to Kramdi's modal logic for trust.[9] It is worth noting that Kramdi's logic for trust is substantially more advanced than deontic logic is for responsibility — therefore, while deontic logic is elegant in its simplicity, it is of limited practical use.

Some attempts are made to overcome this. DeLima et al., for example, create more complex responsibility models which are able to model the allocation and discharge of tasks with greater detail, even in complex MAS.[5] However, further research is yet required to develop DeLima et al. 's model into a fully featured anthropomorphic algorithm for responsibility.

Wallis

Lately, some attention has been paid to responsibility modelling via anthropomorphic algorithms.[19] While previously logical models of responsibility has been developed [1], responsibility formalism work currently underway (by the researchers involved in this grant proposal) provide a socially inspired responsibility formalism in the same vein as Marsh's trust modelling formalism.

This work presents a model of responsibility inspired heavily by sociotechnical systems research — particularly that of Ian Sommerville[15]. In sociotechnical literature, responsibilities are actions which can be discharged; the history of discharged responsibilities discussed as "consequential" responsibilities, and obligation to act in the future identified as "causal" responsibilities. Similar demarcations are made in philosophical literature by Scanlon[14] and by Strawson[16].

As indicated earlier — and as a result of its philosophical inspiration — Wallis' model of responsibility lies on the Marsh and C&F side of the aforementioned spectrum. This might be to be expected: where C&F might consider trust as the belief that another agent will act, Wallis' model of responsibility treats a responsibility as an obligation to act in a certain way. The two traits can, by certain philosophies, be seen as very similar: Wallis' model of responsibility might work particularly well with Marsh's model of trust, for example. By contrast, there is very little similarity in Eigentrust's model of trust and Wallis' model of responsibility. Therefore, the approach one takes when initially creating a model is important; should the software engineer in a research team be unable to grasp the nuances of the social science aspect of a formalism, it may fail to fulfil its requirements simply by being insufficiently similar to the perspectives of other formalisms it is designed to work alongside.

2.4 Discussion

As can be seen, there exists a great breadth in the nature and detail of different anthropomorphic algorithms.

Some models use their respective traits as a metaphor by which they can represent some human-like behaviour; other models use research in philosophy and social sciences for an anthropomorphic realism. Even within the latter camp, differences in perspective with regards social science research can greatly affect the nature of the resulting model.

The creation of a set of guidelines by which this broad research can be carried out is clearly a complex task, which requires a great deal of interdisciplinary knowledge, and a particularly communicative and open-minded set of researchers to enact. However, the research is also urgent: in only 23 years since Marsh's seminal computational model of trust, the field has exploded into myriad forms, which vary wildly. To combat further divergence of the field, which would require even more work to generalise into methodologies and guidelines, this research should be undertaken as soon as possible.

3 Methodology

4 Risks

5 Impact

National Importance

Academic Impact

References

[1] F. Berreby, G. Bourgne, and J.-G. Ganascia. Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for Programming, Artificial Intelligence, and Reasoning*, pages 532–548. Springer, 2015.

[2] G. Birkhoff. *Aesthetic Measure* . 1933.

[3] C. Castelfranchi and R. Falcone. Social Trust : A Cognitive Approach. *Trust and deception in virtual societies*, pages 55–90, 2001.

[4] H. Crawford, K. Renaud, and T. Storer. A framework for continuous, transparent mobile device authentication. *Computers & Security*, 39, Part B:127 – 136, 2013.

[5] T. De Lima, L. E. Royakkers, and F. Dignum. A Logic for Reasoning about Responsibility. 2008.

[6] M. Deutsch. Cooperation and trust: Some theoretical notes. 1962.

[7] A. Herzig, E. Lorini, J. F. Hübner, and L. Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 2009.

[8] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The Eigentrust algorithm for reputation management in P2P networks. *12th International Conference on World Wide Web (WWW)*, page 640, 2003.

[9] S. Kramdi. A modal approach to model computational trust. *PhD Thesis*, 2015.

[10] N. Luhmann. Familiarity, confidence, trust: Problems and alternatives. 2000.

[11] S. P. Marsh. Formalising Trust as a Computational Concept. *Computing*, Doctor of(April):184, 1994.

[12] T. Nam and T. A. Pardo. Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times*, pages 282–291. ACM, 2011.

[13] J. Sabater and C. Sierra. REGRET: Reputation in gregarious societies. *System*.

[14] T. M. Scanlon. Justice, responsibility, and the demands of equality. 2006.

[15] I. Sommerville. Causal Responsibility Models. In *Responsibility and Dependable Systems*, pages 187–207. Springer London, London, 2007.

[16] P. F. Strawson. Freedom and resentment. *Proceedings of the British Academy*, 48:1–25, 1962.

[17] G. H. von Wright. Deontic logic. *Mind*, 60(237):1–15, 1951.

[18] T. Wallis. Anthropomorphic algorithms. Let's Talk About [X] — https://youtu.be/RGUeYQzRsOQ, 2017.

[19] T. Wallis. Investigating computational responsibility. MSci thesis, currently in progress., 2017.