Better Engineering for Anthropomorphic Algorithms

Tom Wallis

1 Introduction

Anthropomorphic algorithms — sometimes referred to as human-like computing — is a growing field which designs computational formalisms of human traits, such as responsibility, trust, or comfort. The field is unusual, in that while it is firmly rooted as a study in computer science, it often relies heavily on research in social sciences such as sociology and psychology: anthropomorphic algorithms are therefore inherently multidisciplinary, and require both mathematical and social reasoning to successfully research.

The most well-researched trait in anthropomorphic algorithms is trust. Many models have been presented, with a range of perspectives concerning trust as a trait to be modelled — and therefore, a range of approaches to its formalism. It is important, because of this range of approaches, to note that some formalisms often treat the trait they model as a metaphor, useful to describe the intended functionality of an algorithm. Eigentrust[5] is one such example. Other models, such as Marsh's model of trust[8] takes a starkly opposite approach, attempting to model trust as it occurs in human behaviour foremostly.

Given the ways that these approaches differ, it can be difficult to identify where building a new model might begin. Anthropomorphic algorithms has a wealth of potential applications as a field, particularly as the variety of traits modelled begins to grow. This is marred by the inherent difficulties of researching and engineering a subject which eludes definition through its potential for variety.

It is important, therefore, to explore the various ways these formalisms are similar, so as to understand what potential avenues can be explored when embarking on research in the field. Particularly, reviewing the existing methods may elucidate whether there is any commonality between different methods: similarities might be used to create guidelines or software tooling to ease the process of engineering these formalisms.

2 Aesthetics

A very early human factor formalism[1] can be found in the work done on "Aesthetic Measure" by Birkhoff[1]. Birkhoff's study of aesthetics was an attempt to find an elegant but accurate mathematical model of something ordinarily thought to be human — in a similar vein to modern anthropomorphic algorithms research.

Birkhoff's work brought him to a representation of aesthetic measure as a ratio of a subject's order to its complexity:

$$M = \frac{O}{C}$$

Birkhoff's work was instrumental in demonstrating that human factors could be represented by mathematical (and therefore computational) processes. One reason for his success in demonstrating this was that Birkhoff was particularly focused on backing his research up with theory from social sciences: having a psychological foundation for his mathematical abstraction dissuades critics from arguing that his work lacked practical foundation.

The practical foundation of Birkhoff's theory strengthens the argument that a human factor can be represented by an equation or process, and later trust modelling sometimes followed a similar approach [2]. Particularly, recurring themes which originate in the work on aesthetic measure include:

- A quantifiable foundation in logic or mathematics
- Concurrent inspiration from humanities, such as philosophy, and social sciences

It is important to recognise the importance of this combination in Birkhoff's early work in this field. As this is effectively one of the earliest anthropomorphic algorithms (though it was not developed with algorithmics in mind), it produces a spectrum of different perspectives to take in anthropomorphic algorithms research: on one side, rigorous science, and on the other, emulating the reasoning and findings of the social sciences.

This spectrum makes creating a set of commonalities between different models difficult, because a model entirely composed of modal logic will be at odds with a model built as a formalism of social features, without mathematical rigour being considered as the formalism is designed. Other commonalities, perhaps pertaining to the nature of their applications, should therefore be considered; as will be seen, this observation will form the motif of this review.

3 Trust

Computational trust literature is varied, and can be influenced by a variety of fields and perspectives.

[1] For the sake of clarification, we define a "human factor" as an element of a social or sociotechnical system which arises from human behaviour, such as Trust. A "formalism" is here a well-defined representation of the thing being formalised, such as the human factor of trust.

[2] as will be seen in in Marsh[8] and Castelfranchi & Falcone[2]'s models

Eigentrust[5] treats trust as a metaphor, used as a tool [3]. Other models, such as the model from Castelfranchi & Falcone, treats trust as a social phenomenon useful to simulate accurately. We will here explore the merits and results of the various methods at these extremes, and where the two approaches meet.

3.1 Marsh's seminal model

Marsh's early model of trust[8] takes a socially-inspired approach to its trust modelling. Here, Marsh takes social notions of trust which he is able to quantify, and exploits this quantification to build his trust model. Specifically, Marsh's model is built on a subdivision of trust into three degrees of detail:

1. Basic Trust
   This is the form of trust a trusting agent has arbitrary other agents, by default. In human terms, one might consider basic trust to be how "trusting" a person is, or how predisposed a person is to trust.

2. General Trust
   This is the form of trust a trusting agent has in another. In human terms, general trust might be considered the disposition one has to trust another agent: "I trust my sister, but not my brother" is an expression of general trust, and an example of its differences as compared to basic trust.

3. Specific Trust
   Specific trust is the form of trust a trusting agent has in another agent to successfully achieve a goal or perform an action. "I trust my sister to drive me to the airport, but not to fly the plane" is an example of how specific trust again differs from general trust in its degree of detail.

Marsh's separation of concerns between these different levels of detail indicate that different things factor in to our discussion of trust when viewed from a social perspective. Competence matters greatly when discussing Specific Trust, for example. Discussing on the level of Basic Trust, by comparison, lends no subject of trust for which to judge competence. The social nature of Marsh's model makes it very easy for humans to reason about, making it easy to interpret and explore. However, Marsh's model becomes more complex as a result.

In constructing this model, Marsh also relies on a wide variety of literature, making its construction as a model more complicated due to the interdisciplinary nature of the research. Marsh juxtaposes psychological literature from Deutsch[4] with sociological literature from Luhmann[7], two well-respected trust researchers in social sciences, so as to achieve a model which functions both at the level of the individual, and of their societies.

Marsh's model is fundamentally mathematical in nature, however: this allows it to be a *computational* model[4]. While it is similar in philosophy to Birkhoff's work[5], the similarities — particularly their mathematical nature — point to a possible path for future work on guidelines and tooling for Anthropomorphic Algorithms.

3.2 Logical Models

Rather than developing a model which is purely quantitative, some researchers construct models in logic. These models have the benefit of being simple to produce and test, and often simple to understand.

Because of the prevalence of the logical model as a format by which anthropomorphic algorithms have been constructed, it is important to analyse these so as to see what similarities with social-based models can be found. If such similarities exist, future work in tooling and guidelines for engineering human-like computing is feasible.

3.2.1 Castelfranchi & Falcone

Castelfranchi and Falcone's logical model of social trust[2][6] provides an alternative method for modelling social traits. Details of trust from social science research is presented as a series of logical predicates, eventually defining such terms as competence, dependence, and disposition in logically rigorous ways.

This degree of detail provides an elegant computational framework for assessing whether an agent ought to trust another. It also provides a simple method for assessing trust in a complicated multi-agent system (or "MAS") — this is made more difficult in Marsh's model of trust, which is designed for the assessment of trust from various agents' perspectives.

C&F is a popular model, and has inspired a series of models which have built on its foundation (such as Kramdi's modal logic[6], discussed below). C&F also lends an alternative mathematical discipline to the conversation on trust, allowing for an assessment of trust from a logical perspective. In this way, C&F contributes greatly to the literature and positions itself as a formidable model.

These logical models, however, have their limitations. Because a logical model typically outputs "true" or "false" for a decision on trust, rather than a numerical interpretation of trust, comparisons of trust become complicated. Models such as Marsh's show that not all forms of trust are equal, even if we might consider them trust. To combat this, the authors provide modifications to their model toward the end of the introductory paper which provide degrees and quantifications of trust; the model is made far more complex as a result, and some of the benefits of its former elegance are lost.

[3] for network security in the original example

[4] For the purposes of this review, a computational model is a formalism of a trait, the design of which permits implementation in software.

[5] Birkhoff and Marsh both rely heavily on research in the social sciences, but formalise their findings mathematically for a computational formalism.

[6] Often referred to as "C&F" or "C&F theory".

3.2.2 Modal Logic

Kramdi's recent modal logic for trust [6] provides another potential avenue to explore the impact of trust modelling in logic. This work has a focus in security applications, but does not strictly model trust in its own right: instead, similarly to deontic logic [12], Kramdi's trust logic is a platform by which trust scenarios can be modelled and computed.

Kramdi's model has similarities to Marsh's, in that it is founded in a concern for the actions an actor makes. It also has similarities to C&F's model, in that it concerns itself with beliefs of agents. Similarly to the earlier models, it also benefits from being backed by research in the social sciences; Kramdi's model is a suitable alternative model to C&F for some purposes, such as the security scenarios for which it was produced.

The more abstract mathematical nature of Kramdi's model, making use of more abstract forms of logic such as modal logic and dynamic epistemic logic, mean that the ease of understanding that a socially-based model such as Marsh's is lost. It benefits, however, from a heightened practical applicability.

3.2.3 Brief Discussion of Logical Models

Logical models, similarly to the mathematical models of Marsh and Birkhoff, ultimately direct the behaviour of some agent; they are also both used in learning contexts for a gradually better "understanding" of the modelled trait on the part of the agent. While the use cases for the models can differ significantly, practical development using the formalisms would ultimately use the model as an influencing factor in guiding behaviour: a possible format that all of the formalisms presented so far might fall into could be an intelligent agent which uses the output of a formalism at any time "t" as a parameter in its utility function. Implementation such as this would permit any of the models presented so far to fulfil their intended purpose using the same basic architecture.

3.3 Reputation-based models

The models shown so far have all modelled trust directly — that is, as models of trust they assess the nature of trust, and produce a formalism of trust and its intricacies.

Reputation-based models operate slightly differently. An alternative approach to modelling trust is to bootstrap the formalism over a formalism of reputation, which is frequently much simpler. Conceptually, the equivalence these models sometimes make between trust and reputation makes sense: one would expect an agent to trust another if the other agent was known to be reputable, and for a lack of reputability to indicate that an agent perhaps should not be trusted.

From this string of reasoning, rather than a deep reading into psychology and sociology literature, trust models based in reputation tend to mimic trust rather than more accurately emulating it. Particularly, one can see that these models are built for specific purposes: Eigentrust[5], for example, uses reputation to assess the reliability of servers on a network, using this as its metric for trust. This method is effective, but one should bear in mind that it treats the trait it models as a conceptual metaphor in the construction of a *tool*.

This difference between the two approaches is very important, because it implies that the philosophy of different state-of-the-art formalisms varies much too wildly to base a common software architecture or research guidelines on. Therefore, specific implementation details should be targeted instead. However, those implementation details differ slightly in the models we have already seen; whether any similarities exist should therefore be explored in the specific literature on the subject.

3.3.1 Eigentrust

Eigentrust[5] is a reputation-based trust formalism designed originally for peer-to-peer communication. To this end, it relies on a very simple aggregation of satisfaction an agent i has in previous interactions with a peer j:

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

...where $sat()$ is a function indicating the total number of satisfactory interactions between two peers, and $unsat()$ the number of unsatisfactory interactions.

The model then makes use of some linear algebra to create a distributed matrix of trust scores, derived from this satisfaction (indicating reputability), to arrive at a model which allows an agent on a network to assess the reputability — and the trustworthiness by extension — of every other peer on the network, taking into account the interactions that *all* peers have had with that agent. The evaluation of the algorithm shows that it withstands certain attacks as a result of the mathematical formulation of the distributed matrix, including resilience to cliques of agents which inflate each other's satisfaction scores.

Eigentrust's implementation is not terribly complex, but allows for a model of trust which works particularly well for trust modelling in MAS. Moreover, the model is built for practical applications, meaning that unlike the earlier models — with sometimes inelegant solutions to modelling trust, and philosophy that computer scientists often cannot relate to — Eigentrust can be implemented and tested fairly well using standard tools and methodologies.

Eigentrust represents a difficulty for a general architecture or tooling for anthropomorphic algorithms generally, however. Eigentrust's distributed matrix means that implementing it as a parameter of an intelligent agent's utility function, or similar architectures which call upon the model to generate a score at time "t", doesn't scale very well if the distributed matrix isn't kept up-to-date. The formalism still directs an agent's behaviour, however; the limiting factor here is the peer-to-peer nature of this model's application. Notes on alternative architectures are supplied in this review's discussion.

4 Non-Trust Anthropomorphic Algorithms

While Trust is easily the most broadly-studied anthropomorphic algorithm, other traits have also been formalised, to varying degrees of depth. As one imagines that research interest in these formalisms will only grow over time, it is important to analyse this research also when developing a common architecture or research guidelines for developing anthropomorphic algorithms.

4.1 Reputation

Rather than formalising trust through reputation, some anthropomorphic algorithms model reputation directly. REGRET[9] is one such model.

The REGRET formalism is a simple mathematical construct which models reputation as the "opinion or view of one (agent) about something"[9]. While its definition of reputation is not directly rooted in psychological or sociological study, their model uses social sciences research to influence the way that it uses certain human-like concepts such as social structure and recency[7]. Other human concepts it utilises include "social reputation", being the fact that, socially, an agent is often perceived to inherit the reputation of a group it happens to join.

REGRET's simplicity and somewhat social philosophy puts it in the middle of the spectrum drawn between Eigentrust's very mathematical formalism and Marsh's social sciences-inspired model. Evaluation of the model shows that aspects such as social reputation lend it a utility practically, while being easy to implement and rooted in psychological research. The balance that it strikes appeals to the goal of building models which accurately represent their trait, while also being pragmatic to engineer. This balance should be supported and encouraged by architectures and guidelines for research in anthropomorphic algorithms.

An important note on REGRET's success is that, should a useful reputation model exist, it can be used to bootstrap traits like trust using mechanisms seen in models such as Eigentrust[5]. However, coupling the engineering of multiple traits together is bad engineering practice. Especially, when undertaking complicated interdisciplinary research in a field such as anthropomorphic algorithms, one wants to avoid the added difficulty of re-engineering the implementation of an experiment only for the purpose of testing an alternative model. A better approach would be to be able to easily swap out one trait for another, even in multi-trait systems such as the hypothetical trust-built-on-REGRET model. Such multi-trait models would be useful to create, yet the engineering of REGRET and other reputation formalisms is entirely unexplored.

4.2 Responsibility

While not as extensive as the literature on trust, responsibility as a field has a wealth of research to draw on. Much of this, unfortunately, does not come from computationally useful backgrounds, but *does* act as a suitable candidate for designing a responsibility formalism around. Examples include the sociotechnical work on responsibility, particularly in dependant systems[11] and in sociotechnical modelling[10].

Examples of computational responsibility work are mostly logical. The most simple and elegant example of this is deontic logic[12], a modal logic for obligations[8]. Deontic logic retains the simplicity of trust models such as C&F, but unfortunately, is too simple to do sophisticated modelling in. Additionally, deontic logic is developed as a tool to consider obligations mathematically, but not for practical emulation of responsibility within an intelligent agent.

These shortcomings are addressed in more recent work on logical formalisms directly addressing responsibility[3], which provide more sophisticated logical tools to analyse and reason about logical methods. The contribution of DeLima et al. is largely a construction of methods for reasoning about task allocation, and is well-suited for task allocation, even in MAS. However, further work needs to be done in the logical system provided to construct a full formalism of responsibility — and, being an entirely logical body of work, DeLima et al. provide a model which is more theoretical than pragmatic. Still, the model shows that one can anticipate a more complete model of responsibility soon, increasing the need for multi-trait engineering techniques as described earlier — computational responsibility models are currently in development[13].

5 Discussion

The formalisms and traits analysed over the course of this review show that anthropomorphic algorithms, as a research field, is very varied and complex. Particularly, the breadth of research interests is large: traits can be accurately emulated, or used as a metaphor, and the degree of mathematical, logical, and social rigour varies dramatically.

Curiously, no guidelines or common implementation details present themselves. However, these traits are used always to better the decisions made by intelligent agents. Therefore, allowing for an easy implementation alongside artificial intelligence and machine learning projects is vital: as these algorithms become more complex and multi-trait models begin to appear more frequently, the development and testing of these models will

[7] The authors note that, in psychology research, it is often the case that how recent an event is can affect its weighting in human assessments of traits — their citations refer to persuasion and opinion. It is worth noting that opinion factors greatly into this model due to the nature of the researchers' philosophy on reputation i.e. that it is an opinion, for practical purposes.

[8] In responsibility literature, obligation and responsibility are often referred to interchangeably, due to the similarities in the topics.

come to rely on small implementation details.

The varied nature of existing formalisms, however, makes a standard architecture difficult to envision. Further research on the topic is clearly required to complete such a standard, and to ensure that it is suitable for the myriad improvements the field will see in the future.

To speculate, a common architecture might include a middleware or microservice-oriented approach which allows for simple functions or entire separate processes to return information in a standard manner. This approach would work especially well for MAS-oriented formalisms, such as Eigentrust, or possible future formalisms where information about peers in the MAS is kept in a centralised storage which requires its own process to operate. To verify this approach, however, and refine details to produce guidelines on implementation details, it is imperative that future work on the subject is carried out.

6 Conclusion

Variation in the approaches of different formalisms, and the fact that different traits are modelled in different ways — sometimes bootstrapping one from another — means that a general architecture is hard to envision. An example of one approach is provided in this review. Without some further work in research guidelines for anthropomorphic algorithms, however, the increasing complexity of the field will make advancements in areas such as multi-trait models difficult to pursue.

The fractured state of the art is also one of the more interesting and useful features of anthropomorphic algorithms as a field, however. As the field sees application from network security to human-computer interaction, variety in the nature and purpose of these models is necessary for its use in a wide range of applications.

Demonstrably, a common architecture (or at least a set of research and engineering guidelines) is vital for future research, and is not provided by the field in its current state. The field's youth, however, implies that this fracturing should be expected only to grow more problematic over time: its current degree of variety has developed in a relatively short space of time. Therefore, it is not only imperative that the proposed work is undertaken, but is undertaken soon, so as to limit future fracturing in the field. Anthropomorphic Algorithms' youth, however, is also a good thing: it limits the degree of variance already available in the field, and provides a small set of differences to accommodate in a common architecture. Therefore, this research should not only be undertaken soon, but should be undertaken *now*, to utilise the limited scope of the problem.

References

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

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

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

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

[5] 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.

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

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

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

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

[10] R. Simpson and T. Storer. Formalising responsibility modelling for automatic analysis. In *Lecture Notes in Business Information Processing*, 2015.

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

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

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