CompSci_colour.pdf

# Investigating Computational Responsibility

William Wallis (2025138)

April 9, 2017

## ABSTRACT

*Currently, models are produced for responsibility modelling which have their roots in logic. These models, while sophisticated, suffer from a lack of pragmatism: for guiding agent behaviour in sociotechnical simulations, logical models are not always ideal. In the similar field of trust modelling, algorithmic models which emulate social behaviour produce useful results while being easier to understand, implement, and reason about. In this paper , a proof-of-concept responsibility modelling platform adopting the algorithmic formalism style employed by trust modelling is produced, and its utility evaluated.*

*paper? report? project?*

## 1. INTRODUCTION

A growing area of research lies in the formalism of human traits into computational representations. These algorithms make computers more human-like; for that reason, they are referred to here as “Anthropomorphic Algorithms” . A similar term, “human-like computing”, has also risen in popularity lately. Human-like computing does not strictly focus on the implementation of formalisms of human traits, however, which is the area of interest for this report.

improve the introduction of the Anthropomorphic Algorithms term

This implementation interest, realised in the study of anthropomorphic algorithms, presents an interesting sociotechnical problem. They present an opportunity to alter the behaviour of actors in a sociotechnical system, and to do so in a way that is easy to reason about. This alternation of behaviour is done by the algorithmic implementation of a *formalism*. Formalisms present a concrete definition — by process, mathematical definition or semantic description — which can be used to construct an anthropomorphic algorithm. These formalisms tend to attempt to model in one of two ways:

1. Modelling the trait as a useful metaphor
   These models tend to be inaccurate with regards the social science surrounding the trait that they model. However, they make a trade-off between this accuracy and the model’s utility. For example, the notion of trust as a metaphor for a type of behaviour might be useful in information security research, but what matters in the formalisms implemented for this research is the formalism’s utility in information security — *not* whether the formalism accurately represents human trust.

2. Modelling social science directly
   These models attempt to accurately model the traits they concern. This can be useful for fields such as sociotechnical modelling, as well as social sciences research. There are also interesting applications for these models in interaction study: making interfaces interact with users in a human-like way, and representing the states of these traits to the users, are valuable research areas which are more applicable to these type-2 formalisms than to type-1 formalisms.

In reality, most formalisms and their implementations lie somewhere on the spectrum that these two types define.

## 2. STATEMENT OF PROBLEM

Computational formalisms of human traits are a growing field of research, with applications in lots of different areas. A problem with these anthropomorphic algorithms is that there is limited breadth to the scope of existing research in the field (as is demonstrated during the background survey in section 3). The metaphor of the human trait in these algorithms remains largely unexplored.

Breadth in the application of the metaphor is important, however. The importance stems from the utility in the human metaphor when designing systems:

- Human-Computer Interaction can make use of behavioural metaphors to relay complicated internal states to a user. Storer et al. [12] demonstrated methods by which a mobile device might dissuade certain user actions by expressing its “discomfort” or lack of “trust” in its interaction design.

- Information Security can make use of behavioural metaphors in order to increase difficulty of access when negative system states are encountered. A system might allow access on a graded scale, dependant on internal states of trust, comfort, and confidence.

- Theoretical advancements in smart city technology[15] might increase a city’s resilience by integrating notions of responsibility into public services and the environment on a community scale.

While similar results can often be achieved using regular techniques, the human metaphor allows for a better communication between a human user and complicated system states. All of the above examples center around this notion; however, the applications extend beyond Human-Computer Interaction research.

The lack of application of the human metaphor is a complicated mosaic of related factors. For example, research into anthropomorphic algorithms holds particular challenges, as a result of its strongly interdisciplinary nature: it requires a research team to understand the nuances of sciences as well as social sciences, and sometimes even humanities. Not only does the research team require the ability to understand these nuances, but the research must take into account their different natures. This often causes divergence in the philosophies of sociotechnical research. Some researchers view sociotechnical systems from the perspective of largely human-based systems with abstract, social behaviour. Others see sociotechnical systems as a combination of dynamic, mathematical processes which produce more technical emergent phenomena. This hints at a third complicating factor, (the second being a lack of convergence in research focus): a lack of a consistent modelling paradigm. Some research focuses largely on actor interaction-style modelling techniques [1], while others rely on purely graphical modelling [8], or on mathematical modelling techniques [13].

These issues together pose an issue for research in anthropomorphic algorithms: a formalism of a human-like trait is only useful to certain researchers, for certain types of models, with certain sociotechnical philosophies. Their lack of broad application is therefore unsurprising; these factors compound to produce yet another, which is that the breadth of traits formalised and researched is very small. The largest degree of research is easily conducted in the field of Trust; other traits, such as Comfort, have recently been attempted also [6].

Recently, some interest has been shown in research pertaining to modelling and formalising *responsibility*. Logical models of social trust exist which could be turned into a proof-of-concept formalism of responsibility with only a small addition: adding an obliging term in a similar way to the Deontic Logic's system for obliging[14], allowing an agent to effectively delegate a task to another which is deemed responsible in discharging responsibilities — the agent selected being known to be trusted already, by C&F's already established work. The scheduling of tasks based on trust is a simple extension of existing models, or an application of existing models.

A model of responsibility might be more than simple task allocation, however. Some logical models of responsibility attempt to model ethically responsible decisions [2]. Deontic logic's obliging term was in itself an attempt to create a logic which was suitable for the calculation of whether an agent was obliged to ensure certain goals, or perform certain tasks. Another angle might be to perceive a model of responsibility as something which might allow a responsible sociotechnical agent to choose responsibilities to discharge, rather than blindly executing tasks they are provided with in a trust-oriented model which simply delegates tasks.

The latter has a number of potential applications. One such possibility would be to implement agent awareness of remote task execution via RPC. Should an agent on a network be given a procedure to execute which is perceived through a responsibility formalism to be unusual, the procedure may not be executed, or may be rejected upon receipt by the responsible agent. Similar applications have proven effective in trust literature, particularly the Eigentrust algorithm [**?**] , where information security is enhanced by inferring agent trustworthiness.

In a realistic sociotechnical system, an agent's behaviour is often informed via a feedback loop. It is important, therefore, to allow an agent to learn better "responsible" behaviour over time, through an analysis of its sociotechnical environment and other factors. In acknowledging that anthropic traits are most useful when they account for both introspection — so as to direct an agent's own behaviour based on the formalised trait — and extrospection — so as to judge and learn from that trait in other agents.

To achieve these goals, two research questions were formulated:

1. How can a computational formalism of responsibility direct the decisions made by an intelligent agent?

2. How can an intelligent agent assume the consequences of actions it makes, the decisions other agents make, and its general environment, so as to direct its interpretation of responsibility?

Answering these questions requires the construction of a proof-of-concept formalism of responsibility which is suitable for directing agent behaviour in an algorithmic manner, as opposed to existing logical methods.

## 3. RELATED WORK

A broad range of literature must be reviewed to properly understand the research at hand, due to the problem's broad nature. Particularly, this paper will focus on three areas: popular algorithmic trust models; broader sociotechnical systems research; and relevant philosophical literature.

paper? report? project?

### 3.1 Trust Modelling

Ordinarily, when constructing a new anthropomorphic algorithm, one would draw on literature regarding other anthropomorphic algorithms which formalise the trait being modelled. However, no such formalism exists for responsibility modelling; therefore, insights from early trust modelling may provide useful information on how to proceed, given that responsibility formalism now is in a similar stage to early trust formalism. Trust is particularly appropriate as a comparative trait to responsibility, as the two traits have many similarities (as investigated earlier).

#### *3.1.1 Marsh*

The seminal anthropomorphic algorithm for trust is found in Marsh's 1994 formalism[7]. In this paper, a formalism is described which has a number of useful qualities: it is modelled on a per-agent basis as opposed to calculating trust across a group of agents, has foundations in social sciences, and is largely algebraic, staying clear of modelling via logic.

Of particular interest is Marsh's separation of three different degrees of granularity in trust judgement. Marsh identifies that human agents have a basic degree or weighting which applies to their trust — it would not be uncommon to assert that "person X is very *trusting*". This is, however, distinctly different to an assertion that "person X is trusting, but *doesn't trust person Y*" — that is, a person's basic level of trust is distinctly different from a person's more directed degree of trust toward another agent in particular. Marsh calls this "General Trust", differentiating it from the earlier "Basic Trust". A final distinction, "Specific Trust", identifies an agent's degree of trust in another with regards a specific task (which could be considered "person X's degree of trust in person Y *in doing task* $\alpha$).

These separations are useful in a few key ways. One is that it identifies some of the key parameters in the model of trust: at the very least, Marsh's formalism of trust requires an agent, two agents, or two agents and a task or action in order to calculate a degree of trust. This is useful when generating a model of responsibility, as trust and responsibility share some common features: like trust, responsibility usually concerns two agents, and a task that one agent is responsible for. Unlike trust, responsibility modelling contains hierarchies: one agent might be considered the authoritative figure, which *delegates* a task to another. For the purposes of this paper

paper? report? dissertation? project?

, the latter agent will be referred to as a "delegee". The similarities in the parameters of the formalism affirm the notion that responsibility and trust are similar in concept, and also serve as a starting point for imagining what an algorithmic formalism of responsibility might be like.

Another useful insight from these separations would be that, when considering responsibility, judging how responsible an agent is can be done from the same three vantage points. Basic responsibility would colloquially be "benefit of the doubt", general responsibility would be how responsible an agent in all modelled capacities, and specific responsibility would be another agent's calculated responsibility with regards some task, resource, or other modelled subject of responsibility). Not only does this also affirm trust and responsibility's similarities, but it helps in providing further structure to the new formalism.

A final important property of Marsh's formalism is its graded nature. This model does not produce a boolean indicator of whether to trust or not; rather, it opts to calculate a number between 0 and 1 of the *degree* of trust one agent ought to have in another. This feature allows for as granular a model of trust as any given application requires, and allows for more nuanced comparison between agents.

Write more of this? Maybe less on the separations, and more on its social science background or lack of logical foundation?

### 3.1.2 Castelfranchi & Falcone

Another useful formalism of trust to consider comes from Castelfranchi & Falcone [3] (often referred to as "C&F theory"). While this formalism has logical foundations, its large popularity makes it an interesting comparison to Marsh's formalism.

In contrast to Marsh's formalism, C&F is graded only in more complex forms. At its root, C&F is a logical formalism built on boolean calculations of goal and belief state satisfaction. However, C&F also segregate different aspects of trust in their formalism: different elements of the basic formalism represent competence, disposition, and dependence. They define competence as one agent's belief and will that another agent can successfully achieve a goal by a certain action, disposition as the belief and will that the other agent is willing to perform an action to achieve a goal, and dependence as the effective delegation of a task for the purpose of achieving a goal (expressed via logical statements).

Curiously, the C&F model of trust contains all of the same parameters as Marsh's, plus a fourth: the goal to be achieved by an action. One may take the position that a goal is also important in modelling responsibility — but it is not immediately apparent that it is necessary to include it. Therefore, C&F uncovers the need when building a formalism of a trait to make choices as to the formalism's philosophy regarding the trait.

In particular, two philosophical differences between Marsh's model and that of C&F are immediately apparent: where Marsh separates his calculation of trust into a basic/general/specific granularity, C&F separate trust's different constituent parts, each of which must be satisfied for trust to be present. Another philosophical difference between the two is the assertion as to whether the goal of an action factors into one's trust. These philosophical decisions can make concrete, important differences to a formalism's constitution — which is plain to see when considering that even the parameters of the model differ, something fundamental to the formalism's definition of trust. Therefore, these choices as to the approach to responsibility are important to note in the formalism produced during this report

report? paper? Dissertation?

.

### 3.1.3 Eigentrust

Eigentrust[**?**] is a formalism which takes a notably different approach to Marsh and C&F's formalisms: rather than attempting to model human trust accurately, it models trust as a metaphor for the truly desired behaviour.

Calculations of trust in Eigentrust have their foundations in eBay's model for reputation: star ratings based on a summation of satisfaction scores. In particular:

$$s(i, j) = sat(i, j) - unsat(i, j)$$

...where *sat* is the number of satisfactory interactions between two agents and *unsat* the unsatisfactory ones, represents a "local trust value" that an agent $i$ has in another agent $j$. Through some linear algebra, these local trust values are accumulated and gradually turned into a global score of responsibility, accounting for all agents' opinions of each other. This global score could also be considered to be an agent's *reputation* — in this way, Eigentrust generalises trust as a certain application of a reputation formalism, and bootstraps its own formalism on another trait.

While Eigentrust has proven particularly effective in its intended domain, it is a formalism designed expressly for the purposes of network and information security. Eigentrust achieves this by a number of philosophical differences to Marsh and C&F's respective formalisms:

- Eigentrust operates in a distributed way. Unlike the local scoring system employed by Marsh and C&F's formalisms, Eigentrust has all agents report their local trust scores, so as to create a distributed ledger of more general trust scores.

- Eigentrust does not model trust directly, nor does it claim to model it accurately — unlike C&F's formalism, which is strictly intended as a model of human behaviour, and Marsh's, which simulates it, Eigentrust uses "trust" as a description of an agent's behaviour.

- Eigentrust is not concerned with the cause of satisfactory or unsatisfactory interactions; it focuses entirely on accumulated positive/negative scores. In this way, Eigentrust somewhat models Marsh's "general trust", but makes no assertions as to trust's composition or how to reason about it.

In these respects, Eigentrust represents an interesting alternative end of the spectrum between anthropomorphic algorithms which treat their trait as a metaphor, or as a social behaviour to accurately simulate. While Eigentrust does not represent a very useful foundation for our responsibility formalism, it does highlight two things:

1. The responsibility formalism required to test the research questions should be more similar to Marsh and C&F's formalisms than Eigentrust: there is no specific use case to design for, so designing with a trait as a metaphor would not be appropriate.

2. Should it be necessary, a trait's formalism could be bootstrapped using a pre-existing formalism of another trait. Whether this is a suitable way to answer the research questions above is harder to address; the possibility should therefore be considered.

### 3.1.4 FIRE

FIRE[4] is another trust modelling system which provides a focus on being able to judge trust using information from many different sources. For example, it treats direct experience information in a different way to information collected by third parties. It is also a model which considers multiple different traits: it incorporates reputation information into its judgement of trust. In part, FIRE is able to do this because it segregates different information sources into different measurements, which are tabulated into a score after their measurement.

FIRE's inclusion of information from multiple sources paints other trust formalisms in a slightly different light. However, this feature is not necessary for all trust scenarios: the decision to trust (or not trust) is made with different amounts of information for agents in different simulations. One can imagine a two-agent simulation, where information about each agent's interactions would be assessed by exactly one source — the other agent, which judges the former's reliability. It is plain to see that FIRE is designed to be a formalism treating traits as a metaphor, similarly to Eigentrust.

This philosophical decision makes FIRE an unlikely candidate as a foundation of a responsibility formalism, as its specific application area — sociotechnical models with multiple types of information to consider — is more complex than is necessary for a proof-of-concept responsibility formalism, and would be more complicated than the research questions posed require. However, the philosophical choice it raises regarding different types of information, and the nature of the information which is being reviewed when calculating the responsibility score, is an important one.

## 3.2 Sociotechnical Systems

While these anthropomorphic algorithms are useful to consider in isolation, their application within the realm of sociotechnical systems is important to their design. Moreover, the nature of responsibilities is touched within the broader sociotechnical systems area of responsibility modelling — research on the delegation and discharge of responsibilities, and how to reason about them.

### 3.2.1 Ian Sommerville

Ian Sommerville was a prolific writer in the field of sociotechnical systems, who was responsible for much of the current literature on responsibility modelling[5, 10, 9].

What other references can go here?

Sommerville's responsibility modelling systems often happened to be graphical[5], a paradigm for computational responsibility which may prove hard to convert to an anthropomorphic algorithm. This is because graphical representations of system states do not naturally present themselves as a numerically analysable format for information — rather, graphical presentations are useful for exposing sociotechnical system state. This feature of sociotechnical modelling, and particularly responsibility modelling, is useful for risk and impact analysis[8]

Sommerville's writing, however, presents a wealth of interesting insights which may be useful in understanding the context of the anthropomorphic algorithm, and understanding some possible choices as to the formalism's philosophy.

### 3.2.2 Tim Storer and Russell Lock

While describing a graphical responsibility modelling format for the InDeED project — notably lead by Ian Sommerville — Storer and Lock provide some useful foundations for other models of responsibility in a technological report on modelling responsibility[11].

Most interestingly, Storer and Lock provide a useful definition of a responsibility:

> A duty, held by some agent, to achieve, maintain or avoid some given state, subject to conformance with organisational, social and cultural norms.[11]
>
> ...
>
> Responsibilities are the duties to be discharged by agents as described...

Numerous things in this definition are useful. Similarly to the work done by Mash and C&F, the definition provides an insight into some possible fundamental parameters of a responsibility:

**A duty:** responsibilities can be considered as some action an agent is *obliged* to conduct.

**achieve... some given state:** the obligation which a duty is defined by can be described as a change of *state*. Storer and Lock note that this state change can have different modes: it can be achieved, but also maintained or avoided. It is worth noting that this very general description of responsibility indicates that the graphical formalism designed for the InDeED project was somewhat socially accurate.

**comformance with... norms:** agents can be delegated responsibilities in most useful formalisms of responsibility — Storer and Lock note that these responsibilities must not conflict with norms that an agent holds.

Unlike Marsh or C&F's more mathematical definitions of trust, this semantic definition lends itself nicely to conversion to some responsibility model; partly because it already summarises the responsibility trait, but also because it describes an intuitive social definition of responsibility while remaining very simple. Therefore, a responsibility formalism which somehow extended this model, while retaining the power of the anthropomorphic algorithms described earlier would provide a very useful starting point for the desired responsibility formalism.

Another useful definition from Storer and Lock's report presents itself when discussing the structure of a responsibility in their model:

> Responsibilities may be composed of other responsibilities.

This composibility is an interesting property, which may also imply that some responsibilities can be broken down into a form of sub-responsibility. This property would be useful to remember when modelling responsibilities in the desired formalism. It also implies that there may be a sort of "atomic" responsibility, that more complicated responsibilities are composed of.

## 3.3 Philosophical Literature

### *3.3.1 P.F. Strawson*

### *3.3.2 Thomas Scanlon*

### *3.3.3 Sloman*

# 4. A FORMALISM OF RESPONSIBILITY

## 4.1 Design Decisions and Philosophy

### *4.1.1 Constraints*

### *4.1.2 Obligations*

### *4.1.3 Responsibilities*

### *4.1.4 Agents*

# 5. EVALUATING THE FORMALISM

## 5.1 RQ 1: Acting on Responsibilities

## 5.2 RQ 2: Judgement of Responsibility

# 6. FUTURE WORK

## 6.1 Advancing the formalism

# 7. DISCUSSION

# 8. REFERENCES

[1] G. Baxter and I. Sommerville. Socio-technical systems: From design methods to systems engineering. *Interacting with computers*, 23(1):4–17, 2011.

[2] 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.

[3] C. Castelfranchi and R. Falcone. Social Trust: A Cognitive Approach. 2001.

[4] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. Fire: An integrated trust and reputation model for open multi-agent systems. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pages 23–27. IOS Press, 2004.

[5] R. Lock, T. Storer, I. Sommerville, and G. Baxter. Responsibility modelling for risk analysis. 2010.

[6] S. Marsh, P. Briggs, K. El-Khatib, B. Esfandiari, and J. A. Stewart. Defining and investigating device comfort. *Information and Media Technologies*, 6(3):914–935, 2011.

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

[8] F. C. Paul Wallis. *The OBASHI Methodology*, volume 1. The Stationery Office, 1 edition, 2010. The offical manual for the OBASHI Methodology.

[9] I. Sommerville. Causal responsibility models. In *Responsibility and Dependable Systems*, pages 187–207. Springer, 2007.

[10] I. Sommerville. Models for responsibility assignment. In *Responsibility and dependable systems*, pages 165–186. Springer, 2007.

[11] T. Storer and R. Lock. Modelling responsibility. Technical report, Project Working Paper 7, InDeED Project (April 2008), 2008.

[12] V. T. Tim Storer, Karen Renauld. Find this find this. 2020.

[13] A. Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39, 2012.

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

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