Proposing a Model of Computational Responsibility

William T. Wallis

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Masters project proposal

Contents

1 Introduction

Computational Responsibility is a field with little to no existing literature. Rather than a focus on *responsibility*, researchers have so far tackled a variety of other social topics through computational formalisation:

- Marsh's seminal work on Trust[6]
- Stricter formal definitions on Trust, from a cognitive standpoint[2]
- Some responsibility modelling, from a logical formalisation[7]
- Some work on reputation [3]
- Models of computational comfort models[5].

Finish reading this!

While there is no direct literature on responsibility formalisms, then, we can see that there exists a wealth of literature for a responsibility formalism to be inspired by.

A responsibility formalism is useful in the same ways that formalisms of human traits such as reputation and trust might be; however, a responsibility formalism has the potential to have impacts in areas trust and reputation might not. For example, imbuing an intelligent agent with a sense of responsibility might provide it a greater degree of corrigibility[10]. An agent overseeing network security which understands its responsibilities within a much larger security system might better prioritise its duties when confronted with an unusual situation. Computational responsibility frameworks might help better model the emergent phenomena in sociotechnical systems, combine with traits like trust and comfort to make a more anthropomorphic device for better HCI, or perhaps help predict human actions in large computational models of human actors. We will explore some of these practical applications in Section 4§ 4.

However, it is certain that a uses for these formalisms present themselves at every turn.

2 Statement of Problem

Computational responsibility is a complex area with lots of incidentally related work, but no specific relevant literature. Instead of focusing on the responsibilities of artificial agents, their responsibilities are implied by the construction of the agent itself. It might employ algorithms for driving without human guidance, or classify network traffic in an attempt to flag attempts at a system's security. In these instances, lots of somewhat-related work has been done on computational *trust*: can one artificial agent trust another?

However, this approach is short-sighted. While trust and responsibility are intrinsically linked social concepts, no work has been done to migrate the models of trust to new models of responsibility. A concern arises: do artificially intelligent agents, which we put at the helm of concerns like network security and road safety, actually communicate its understanding of its assigned duty with other agents it collaborates with? Two examples present themselves.

The first: a car might drive along a residential street and identify a squirrel running across the road in front of it. It calculates a high probability that, unless it swerves out of the way of the squirrel, it may kill it. It simultaneously identifies that, in the country it is driving in, the law states that it should swerve to avoid killing animals if possible. Computational responsibility introduces itself into the problem in that the car should also have a social understanding: will the swerve endanger humans? How strongly should it weight that probability into the action it chooses? Is it also responsible for, say, conserving fuel for environmental reasons? The key here is that the car has many goals to ascertain; while some are more immediate than others, it should have the capacity to weigh *multiple, arbitrary responsibilities* up to surmise what its next action is.

The second: an artificially intelligent agent watches the price of a collection of books in an online store. This is common practice on large sites where prices of unusual books can fluctuate wildly.

Here one artificial agent is known to have artificially inflated the price of a book; another agent has *also* inflated the price according to the seeming market trend. The first agent, seeing that the book is rising in value and now underpriced, inflates the price of its own copy, and the cycle continues until a human intervenes.

Kevin Slavin discusses the idea that we have begun to design a world *'for algorithms[8], with nothing but a big red button, labelled ś́top'*. The precession of this design trend marches on, relentless — but algorithms, rather than their interfaces, can be built with humans in mind. A mutual understanding of responsibility would allow one algorithm in this cycle to delegate the price inflation of its book to the other, breaking the cycle, so long as the concept of responsibility for a task is mutually understood. This is where the second, real-world example of computational responsibility lies.

As can be seen in the model proposed by Castelfranchi & Falcone in their formulation of

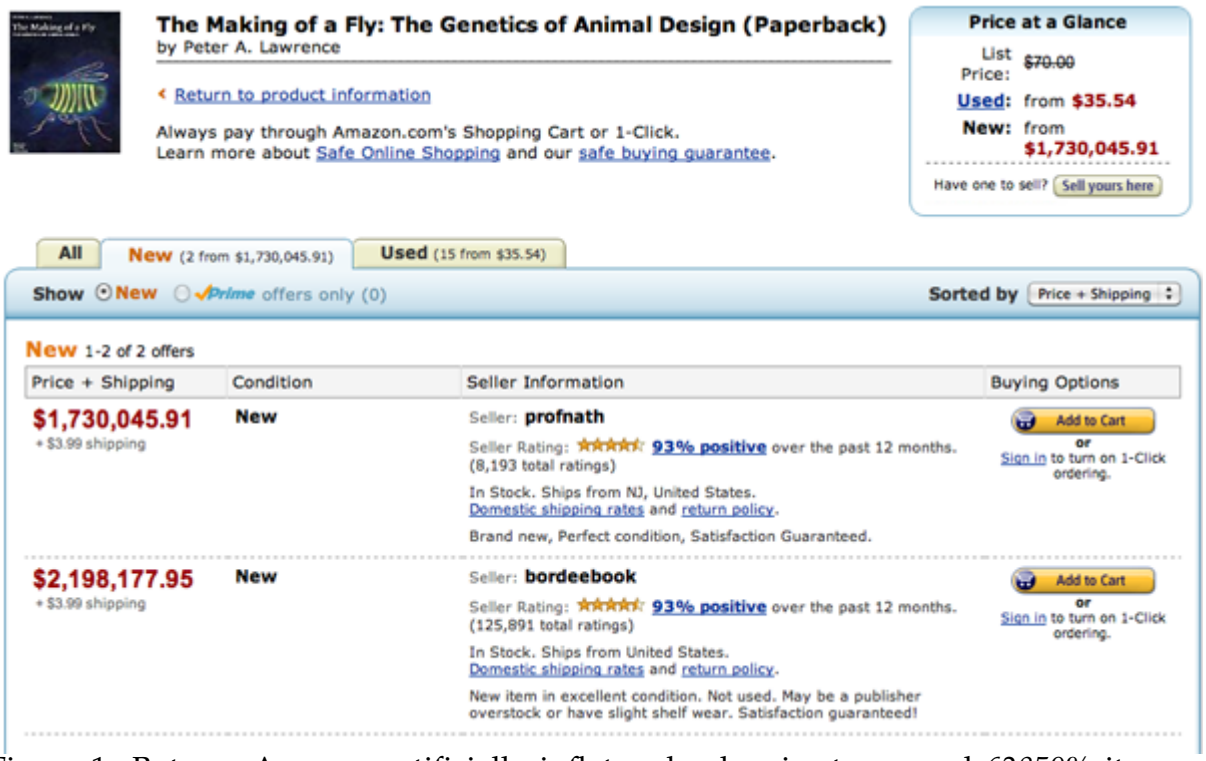

Figure 1: Bots on Amazon artificially inflate a book price to around *62850%* its used price

cognitive trust[2] (usually referred to as *C&F Theory*), a formulation of trust surrounding one or more actors, subjectively assessing tasks and goals, which takes into account social and technical factors in its modelling, is already present. Fortunately, this model is very well accepted by the computational trust community! Therefore, some work presents itself: does an adaptation of C&F theory suit a practical model and implementation of computational responsibility? Secondly, one is also led to wonder: how well would such a model solve the example applications of computational responsibility explained earlier?

2.1 A need for a cognitive computational responsibility

As we move into a world increasingly dominated by algorithms and shaped by their decisions, there is a clear requirement for responsible systems. One problem arises: how can we be certain that an algorithm's internal conception of responsibility is 'human-like'? Early work by Sloman describes the notion of a 'space' of minds[9], and this concept is useful here. An artificial mind need not be human-like, or even biological-like; it can occupy an entirely different area of the space of minds altogether.

This is problematic when imposing human-like concepts onto machines. The effort of imposing a social, human-like construct onto a 'mind' that may not easily host the concept means that one runs the risk of imitating the trait in a useful way, but not in an accurate way. If the model of the human trait doesn't stem from the same basic concepts

as the human model does, it cannot be relied on to behave in a human-like way all of the time.

Therefore, while we have already demonstrated a need for computational responsibility, there is another requirement that must be satisfied: *cognitive* computational responsibility.

C&F define a cognitive agent as:

> "Only a cognitive agent can "trust" another agent; only an agent *endowed with goals and beliefs*"

This definition doesn't quite fit our purposes — as will be seen, our definition also requires the concept of *obligation*. However, it can be seen that this definition is deliberately high-level in order to simulate the important components of a human trusting agent. A cognitive agent can be seen as an agent which, for the task it is set out to do, is modelled in a *high-level, human like way*.

Thus, a cognitive model of responsibility is necessary; an ordinary model might suffice for theoretical or research purposes, and may be useful in analysing scenarios bound tighter by, say, law, than they are by the normal human's cognition.

3 Background Survey

Unlike Responsibility, Trust is a topic which has a surprising degree of pre-existing literature. Marsh [**?**] draws inspiration from as early as David Birkhoff's 1930s work in creating an 'Aesthetic Measure', where the famous mathematician created a quantification of Aesthetics. While some dispute that such subjective topics can be boiled down to a single number (or array thereof), much work to the contrary has now been completed. Like Marsh, we should start from the beginning.

3.1 Social Sciences and Mathematics

3.1.1 Birkhoff's Aesthetic Measure

One of the earlier formalisms of a human factor[1][2] was Birkhoff's definition of Aesthetic Measure[1]. In it, Birkhoff defines the notion of Aesthetic Measure as a ratio of Order to Complexity:

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

Birkhoff's work inadvertently gave rise to the notion that human factors can be represented by mathematical equations and systems. Birkhoff's formalism of aesthetics became popular for a few reasons, but one of particular interest to later Trust modelling work was that Birkhoff put a great degree of effort into backing his work up with psychological theory. In this way, Birkhoff's formalism could be said to be a *psychological* formalism.

Later trust modelling work followed in Birkhoff's footsteps here. Indeed, Birkhoff gives a solid foundation for the model-creating method later employed by Marsh[6] and Castelfranchi & Falcone, as it is:

- Founded on mathematical or logical principles which are *quantifiable*
- Heavily inspired and directed by related work in psychology, sociology, and philosophy

[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.

[2]Also for the sake of clarifying a sociotechnical system, a sociotechnical system is a system composed of human tendencies and behaviours, such as Trust, alongside technical activity, such as a computer or a steam engine. An example might be a coffee shop:

- Humans take orders and manage the running of the shop
- Technology is responsible for complex activities such as taking payments and forcing steam through coffee at high pressure

so there are both social and technical actors and behaviours in the "system" of a day-to-day coffee shop.

The marriage of social studies with mathematical rigour will be a recurring theme of the work related to Computational Trust.

3.1.2 Deutsch

Following the quantifiable, mathematical work done by Birkhoff, logical and arithmetic formalisms of human factors followed. One of the earlier and more widely adopted models for Trust came from Deutsch in 1962. Deutsch is a psychologist who did swathes of work in the topic of cooperation, touching on Trust during the 60s.

Deutsch's formalism of trust wasn't immediately quantifiable, but presented one of the earliest well-defined definitions of trust. To paraphrase Deutsch's formalism in "Cooperation and Trust: Some Theoretical Notes", 1962:

CITE THIS

- An actor is presented with a choice between two paths.

  A: No change

  B: The actor takes some action, of ambiguous outcome. A possible gain is associated, *P*, and some possible risk is associated, *R*.

- The actor assesses that the outcome of choice *B* relies on the behaviour of another actor.

- The actor assesses the action they may take and resolves that the strength of *R*, likelihood of *R* as an outcome, or both are higher than the respective *P* measurements.

- The actor is said to be *trusting* they take path *B*.

This formalism introduces some interesting notions. For example, it is unclear as to whether the outcome of choice *B* can rely on the same actor making the decision; can one trust oneself by Deutsch's definition? Another interesting analysis of the implications of Deutsch's model is that it does not rely on the *accurate* measurement of risk and utility, but just its perception — trust is subjective, and based on the trusting actor's perspective on the world.

Rather than characterising trust by the parties involved, Deutsch's formalism is characterised by *risk and utility*. A simple quantification of Deutsch's formalism could be devised, therefore, where risk and utility are quantified by simple assessments using utility functions and a form of risk analysis. Even so, the outcome of this quantified system is a single bit: trusting or not trusting. This does quantify trust, but only technically speaking, and this quantification is weak in its expressiveness. It gives no remit to suggest that one might trust one person over another, for example, as there are no orderable degrees of trust.

Deutsch offers many different ideas as to why and how trust or trust-like behaviour can come about, however. This list is taken from Marsh 1994[6], where explanations of all nine can be found:

1. Trust as Despair
2. Trust as Social Confirmity
3. Trust as Innocence
4. Trust as Impulsiveness
5. Trust as Virtue
6. Trust as Masochism
7. Trust as Faith
8. Risk-taking or Gambling
9. Trust as Confidence

Deutsch's given model above specifically targets formalisation of trust as confidence.

3.1.3 Luhmann

Luhmann, a sociologist who also worked in Trust and related fields, had his own take on formalisms of Trust: that trust was a social tool for reducing the complexity of a social system. Specifically, Lohmann sees trust as being a method whereby agents in a social system can reduce their exposure of *risk* to each other. According to Luhmann, "Trust... pressuposes a situation of risk."

CITE THIS

Luhmann's work is therefore difficult to form quantitative formalisms from, as his thesis stems from a risk analysis perspective, which can be particularly difficult in a sociotechnical system. However, Luhmann's work remains interesting; a formalism of a human factor like trust would be incomplete without considering the properties of individual human actors as well as these properties' emergent effects in the larger sociotechnical space. For small systems, these social-level properties may not present themselves very strongly; however, most human factors are present regardless of the scale of the system being modelled. Therefore, a formalism of a human factor which fails to consider both psychological and sociological aspects cannot be complete.

3.2 Modern [Computational] Trust methods

3.2.1 Marsh's formalism

The earliest quantifiable formalism of trust which provides computability, flexibility, and an inspiration from the sociological and psychological work above is that of Stephen Marsh in 1994[6]. Marsh's work breaks trust up into three core quantifications, where each variable takes some value in the range $[-1, 1)$:

1. Basic Trust
   This is the general degree of "trustingness" about an agent, or that agent's ordinary inclination to trust.

2. General Trust
   General trust is trust in the context of the agent being trusted. Marsh's original description begins[6]:

   > Given two agents, $x, y \in \mathcal{A}$, to notate 'x trusts y' we use: $T_x(y)$. ...The value represents the amount of trust x has in y here.

   So, General Trust can be seen to be the trust that an agent x has in y.

3. Situational Trust
   Trust doesn't exist in a vaccum, and the only variable isn't the subject of x's trust; y may have varying degrees of competency in performing an action. Therefore, Situational Trust can be seen to be the trust x holds that y can actually perform some task, α. Marsh helpfully gives the example[6]:

   > ...whilst I may trust my brother to drive me to the airport, I certainly would not trust him to fly the plane!

3.3 Philosophy of Moral Responsibility

Philosophy regarding moral responsibility is an area whose literature is both wide and deep. That said, not all moral repsonsibility literature is relevant to a computational repsonsibility project; lots of it is designed from a social analysis perspective which would be difficult to implement in any useful way. Other areas, however, present more promise for studies regarding formalisms.

One example of research with utility in a computational way is that of Peter F Strawson, particularly in his seminal essay, Freedom and Resentment [11]

3.4 Ben Colburn

3.5 Comparing Trust and Responsibility

3.6 What work is missing?

4 Proposed Approach

4.1 C&F Close, but no cigar

As it turns out, cognitive computational trust models that already exist are almost but not quite appropriate for modelling responsibility, too. The C&F trust model requires only four main ingredients to formulate a cognitive trust model:

1. x, a truster
2. y, a subject of trust
3. g, a goal of x
4. α, an action of y

This model gets us close to where we need to be to model responsibility; like responsibility modelling often does, it assumes two agents. There also exists some goal which can be met, which — to use C&F terminology — is *delegated* by x to y. Y can achieve this goal through some action, α. So far, all of this forms the beginning of a foundation for cognitive responsibility; what turns delegation of a task into the consignment of responsibility is that of obligation, and the understanding of obligation.

It is evident that trust and responsibility models are, even in the human-like cognitive approach, very similar. However, crucial differences mean that we cannot directly apply C&F theory to the idea of computational responsibility.

Therefore, I propose that research must be carried out to ascertain whether C&F can, as a model, be adapted simply to account for an agent's responsibility. In addition, research must be carried out to implement this model in a BDI logic agent, enabling the evaluation of the new model's success.

WE SHOULD BE ABLE TO ADAPT [4] TO IMPLEMENT OUR NEW COGNITIVE RESPONSIBILITY MODEL IN A BELIEF, DESIRE, INTENTION AGENT MODEL. THEY DO A DIRECT APPLICATION OF C&F TO BDI, WE SHOULD TOO.

5 Work Plan

Panic, write the report in a 36 hour caffeine-induced fever dream

Todo Notes

References

[1] G D Birkhoff. AESTHETIC MEASURE.

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

[3] 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.

[4] Jomi F Hübner, Emiliano Lorini, Laurent Vercouter, and Andreas Herzig. From cognitive trust theories to computational trust.

[5] 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.

[6] Stephen Paul Marsh. Formalising Trust as a Computational Concept. *Computing*, Doctor of(April):184, 1994.

[7] Robbie Simpson and Tim Storer. Formalising responsibility modelling for automatic analysis. In *Lecture Notes in Business Information Processing*, 2015.

[8] Kevin Slavin. HOW ALGORITHMS SHAPE OUR WORLD.

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

[10] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. *AAAI Workshop on AI and Ethics*, (2014):74–82, 2015.

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