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[9]
- Stricter formal definitions on Trust, from a cognitive standpoint[2]
- Some responsibility modelling, from a logical formalisation[12]

Finish reading this!

- Some work on reputation [3]
- Models of computational comfort models[8].

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[15]. 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 § 3.5.

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

Write something concrete here regarding fields it might be applicable in, like decision theory or AI safety or network security or sociotechnical modelling

1.1 An early rebuttal of some common criticisms

One easy criticism to make of these anthropomorphic formalisms is the argument that, say, a trust formalism doesn't represent "true" trust. To address this point early, a responsibility formalism such as the one proposed need not be an entirely human-like representation of responsibility for every definition. Rather, there is a utility in an agent giving the *appearance* of responsibility. (If one follows the deterministic school of thought, there is also an argument that there is no difference[5].) Whether one considers it "true" responsibility should arguably be secondary to whether responsibility-like traits are useful to have computational frameworks for; we will see that these traits are indeed useful, and so that the criticism is moot.

1.2 Terminology

Fill with terminology, in a style similar to honours dissertation

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, simply building "responsibility" into a system without any understanding of responsibly-made decisions, or ability to learn obligations and duties in a concrete way, means we lose a great opportunity. Moreover, 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 that consider topics like obligation and duty. A concern arises: do artificially intelligent agents, which we put at the helm of concerns like network security and road safety, miss out as a result of their failure to consider duty and obligation? 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? And if so, which responsibility is more important?

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. Clearly, this is a problem for decision theory; but one where an understanding of responsibility may be of great help. Unfortunately, this example may be only expository as of yet: a practical model of a self-driving car with this degree of responsible awareness would be rather complex, and a model this advanced is beyond the scope of this project.

The second: two intelligent agents raise or lower the price of a book they manage according to percieved changes in the market. This is common practice on large sites where prices of unusual books can fluctuate according to sudden rises in demand, as seen in § 2.

Confirm figure referencing style

In § 2, 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.

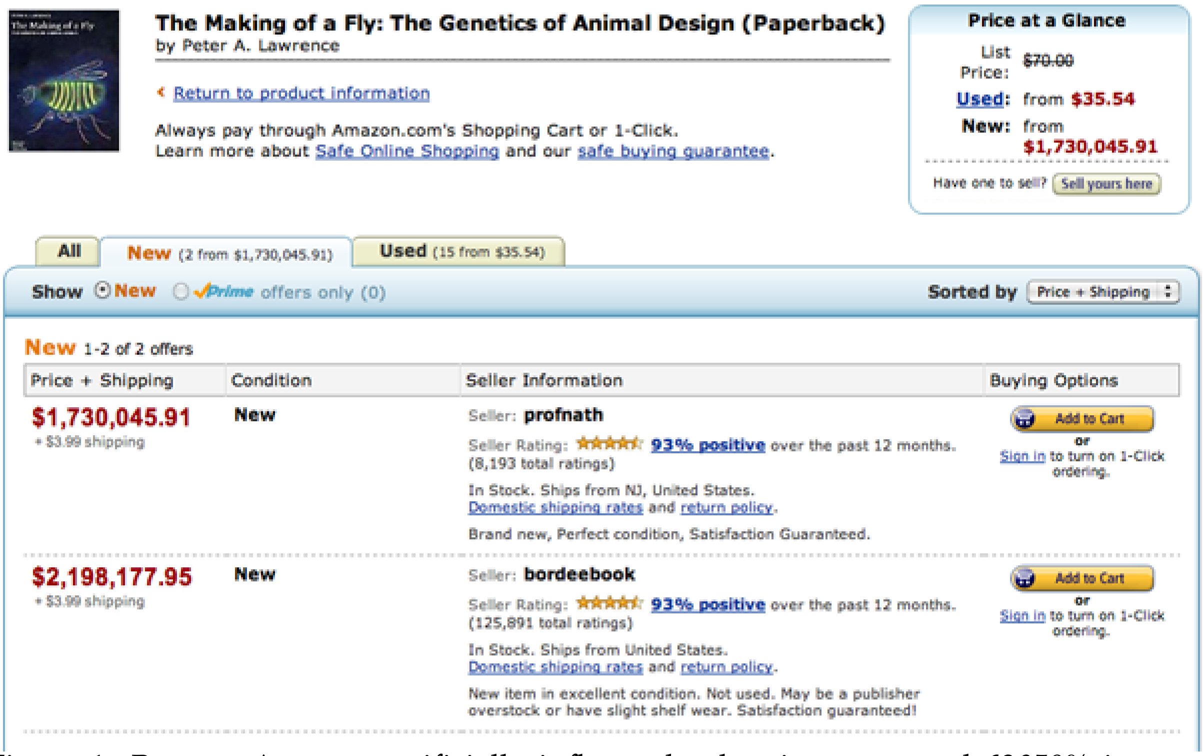

Figure 1: Bots on Amazon artificially inflate a book price to around *62850%* its used price

Kevin Slavin discusses the idea that we have begun to design a world *"for algorithms, with nothing but a big red button, labelled "stop" "*[13]. 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 from the existence of models for concepts such as Trust which can solve similar HCI problems [11], mimicking human traits computationally has its benefits. Moreover, we can be certain that, just as with trust modelling, useful and thorough responsibility models can produce work in machine ethics[10], sociology[7], and clearly, computing science. We can therefore expect that a computational model of responsibility will yeild similar results — perhaps breaking new ground in other fields, which traits like Trust or Comfort have less relevance to.

2.1 Reflective Agents

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[14], 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.

Therefore, when developing the proposed model of responsibility, one has to wonder what the components of the machine mind would be, such that it could house some useful definition of "responsibility". This useful definition need not be accurate; however, it will require the emulation of fundamental human attributes in order to successfully simulate.

For example: C&F define a "cognitive" agent as the lower limit of an agent's requirements for human traits for trust. 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*.

Therefore, we might define our own high-level requirement of responsible computational agents:

> Only a reflective agent can be "responsible" for its actions; only an agent which can *reflect on its obligations when choosing an action*.

This definition of a "responsible agent" as a "reflective" agent is important, because when considering its obligations, a responsible agent should be able to gauge whether to act in a certain way, weighted by their responsibility for a given obligation's fulfilment. As the model of responsibility developed begins to take shape, necessary components of those obligations — the responsibility equivalent of trust's goals and beliefs — will come to light.

2.2 Interpretive Agents

Unfortunately, simply reflecting on one's responsibility is not the only high-level requirement we can forsee needing for a responsible computational agent. Humans do not simply reflect on their obligations and duties before deciding on what their next actions might be. Human agents also see those obligations and duties through their own lens; they interpret their responsibilities according to certain factors which may influence their "feeling" of obligation.

One can see this, for example, in people's respect for law or social convention. Sime citizens of a community might feel that it is imperative not to ride a bicycle on the pavement in Britain, as it is technically illegal. Others may well avoid the road traffic by making use of pedestrian areas if there aren't many pedestrians allowed, regardless of the law. Another example might be crossing the road; if a small child is present, parents of the child may well be teaching it to cross the road safely. To cross at a "red man" then, regardless of the presence of cars, somewhat derails the parent's lesson. It may even give the child an example of why they should be allowed to cross the road when they want. To cross the road at a "red man" does not respect one's influence over the situation at hand; in other words, the *responsible* thing to do is to wait for the lights to change.

Better examples? Examples of both subjective interpretation and comparing two subjective interpretations? (Some agents might perceive one responsibility with the same score as more important than another responsibility, another agent might get it the other way around)

However, it is clear that not everybody thinks this way; many cyclists ride on the pavements, and stopping at a red light to aid a parent in teaching their child might be considered extreme by some. The subjective nature of responsibility belies its interpretive nature: human actors interpret their obligations according to their beliefs, knowledge, and preferences, amongst other things. Therefore:

> Involved in a "reflective agent"ś judgement of their responsibilities is a subjective component: an interpretive function which converts information about an obligation or duty into a subjective score of responsibility.

This way, the human-like subjectivity of responsibility can be simulated.

2.3 Reactive Agents

There is one last definition important to our understanding of what a basic responsibility formalism might look like. If we want to represent a human-like sensation of trust, it's important that we acknowledge that human agents become more or less trustworthy over time. Children, for example, can be trusted much more once they mature and become adults. Moreover, that adult agent will probably be slightly less likely to trust than it was as a naive child. Therefore, to simulate human-like responsibility, agents' interpretation of responsibility should change over time.

As we'll see, responsible agents whose sense of responsibility changes as a reaction to its environment is a notion which is backed not only by observation, but by the moral philosophy which underpins the proposed work at the end of this review. For now, we can define these "reactive agents" as agents whose sense of responsibility shifts as their environment is seen to change:

> Only a "reactive agent" has a changing subjective outlook on the world; it *changes its reflection on its own and other agents' responsibilities depending on its environment*.

Clunky? Reword?

It is clear to see that for a reactive agent to change its reflection on responsibility, it must be an interpretive agent; in other words, the set of reactive agents is a subset of the set of interpretive agents. It should also be clear to see that a useful responsible agent should be both *reactive* and *reflective*: the set of reactive, reflective agents have a perception of the world which is subjective, can change according to its environment, and uses its sense of responsibility when making decisions.

2.4 Using the definitions

These definitions are useful in that they allow us to begin to see what might compose a responsible agent. We might begin to build a computational model of responsibility which embodies these traits by investigating the following research questions:

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?

These are the research questions I seek to answer in the course of this project.

3 Background Survey

Unlike Computational Responsibility, Computational Trust is a topic which has a surprising degree of pre-existing literature. Marsh [9] 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[9] 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[9], 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[9]. 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[9]:

   > 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[9]:

   > ...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.

3.3.1 Strawson

One example of research with utility in a computational way is that of Peter F Strawson, particularly in his seminal essay, Freedom and Resentment [16]. Strawson's topic actually revolves around whether determinism has any impact on "free will" — a discussion clearly outside the scope of this project — but in forming his argument creates some key conepts that we can use to consider the applicability of a responsibility formalism to a computational system, as well as touching on what that formalism would look like.

Strawson's fundamental argument can be construed as being that determinism doesn't affect what human factors — like Responsibility and Trust — mean, because these concepts are founded on the relationships between human actors, rather than being inherant to the human actors themselves. As computer scientists, we can extrapolate this argument out to a sociotechnical environment. That is: Strawson's argument applies to both socila *and* technological agents within a sociotechnical world. Using Strawson's reasoning, then, we can firmly conclude that it *does* make sense to create "responsible" computers, because responsibility makes sense as a trait for a computer to have in the same way it might a human actor.

Strawson's insights don't stop there, though. He also produces an interestingly rigorous analysis of how ordinarily fuzzy human factors can be formalised appropriately:

> Indignation, disapprobation, like resentment, tend to inhibit or at least to limit our goodwill towards the object of these attitudes, tend to promote an at least partial and temporary withdrawal of goodwill; they do so in proportion as they are strong; and their strength is in general proportioned to what is felt to be the magnitude of the injury and to the degree to which the agents will is identified with, or indifferent to, it.

[16]

Strawson captures an essential component of Marsh's computational trust model: an actor's actions influence the percieved trustworthiness of an agent in proportion of the magnitude of the "trustworthiness" of their actions. This concept, Strawson's elucidation shows, is one which can also extend to responsibility; it influences all of the human factors Strawson seeks to address in his essay (i.e. all human factors).

This gives us an insight into how a realistic computational trust model might function: its perception of responsibility should alter depending on external factors. We are aware, however, that people's actions are not the only things which can affect our feeling of responsibility: all sociotechnical factors might. For example, the "Bystander Effect"[4] occurs when one feels less responsible for acting in an emergency situation because of the number of people present who could help; in the end, nobody does, because every person's sense of responsibility is weakened. However, it is not weakened by any person's actions: it's influenced by a set of sociotechnical factors, of which the feelings that Strawson discusses are a subset.

A suitable conclusion one might draw, then, would be that a valid computational model of responsibility *must* take into account an analysis of the sociotechnical environment of the responsible agent. It would also be valid to draw the conclusion — as a result of the first of Strawson's points discussed — that it makes sense to talk about "trusting" and “responsible” computational agents as if they were humans. Another important conclusion to draw from this part of Stawson’s discussion is that this is one way in which agents’ interaction is meaningful: therefore, there is a significant body of work to be undertaken in the combination of trust and responsibility formalisms to the field of HCI. Some of this work is already underway[8, 7].

3.3.2 Aside: Philosophy’s impact on Computing Science

It’s worth noting that Philosophy and Computing Science are unusual bedfellows. However, this need not be the case. As we’ve seen already, and will continue to observe, there is a definite need for computing science to make use of and contribute to research involving human factors — whether those factors be sociological, psychological, ethnographic, or even philosophical. What’s more, philosophical research may well have an impact on the ethics and cultural implications that certain advances in Computing Science might entail. When discussing matters of potentially existential levels of risk, it is important to have the foresight to construct solid, informed arguments about our role in the grander scope of things.

To this end, I believe collaboration between Philosophy and Computing Science should be a more routine affair. Some interdisciplinary work of this form is already underway: Nick Bostrom’s Future of Humanity Institute at Oxford University and the Machine Intelligence Research Institute in the USA are two excellent examples of Computing Science and Philosophy coming together to produce vitally important work, which greatly influences the progression of the field of AI Safety.

3.4 Comparing Trust and Responsibility

3.5 What work is missing?

sectionProposed Approach

Develop arguments that the responsibility formalism might actually be put to good use, as per § 1

3.6 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 [6] 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.

4 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] Peter Fischer, Tobias Greitemeyer, Fabian Pollozek, and Dieter Frey. The unresponsive bystander: are bystanders more responsive in dangerous emergencies? *European Journal of Social Psychology*, 36(2):267–278, 2006.

[5] Ted Honderich. Free will, determinism and moral responsibility – the whole thing in brief.

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

[7] Michael W. Macy and Robert Willer. From Factors to Factors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28(1):143–166, 2002.

[8] 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.

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

[10] James H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.

[11] Andrew Patrick, Stephen Marsh, and P Briggs. Designing systems that people will trust. *Security and Usability: Designing Secure Systems That People Can Use*, pages 75–100, 2005.

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

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

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

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

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