Proposing a model of computational repsonsibility

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[3]
- Stricter formal definitions on Trust, from a cognitive standpoint[1]
- Some responsibility modelling, from a logical formalisation[4] *I'm still reading this, Tim!*
- Some work on reputation [2]

something something there's a gap here where nobody's applied machine learning to the problem for teaching artificial agents about the social concept of responsibility

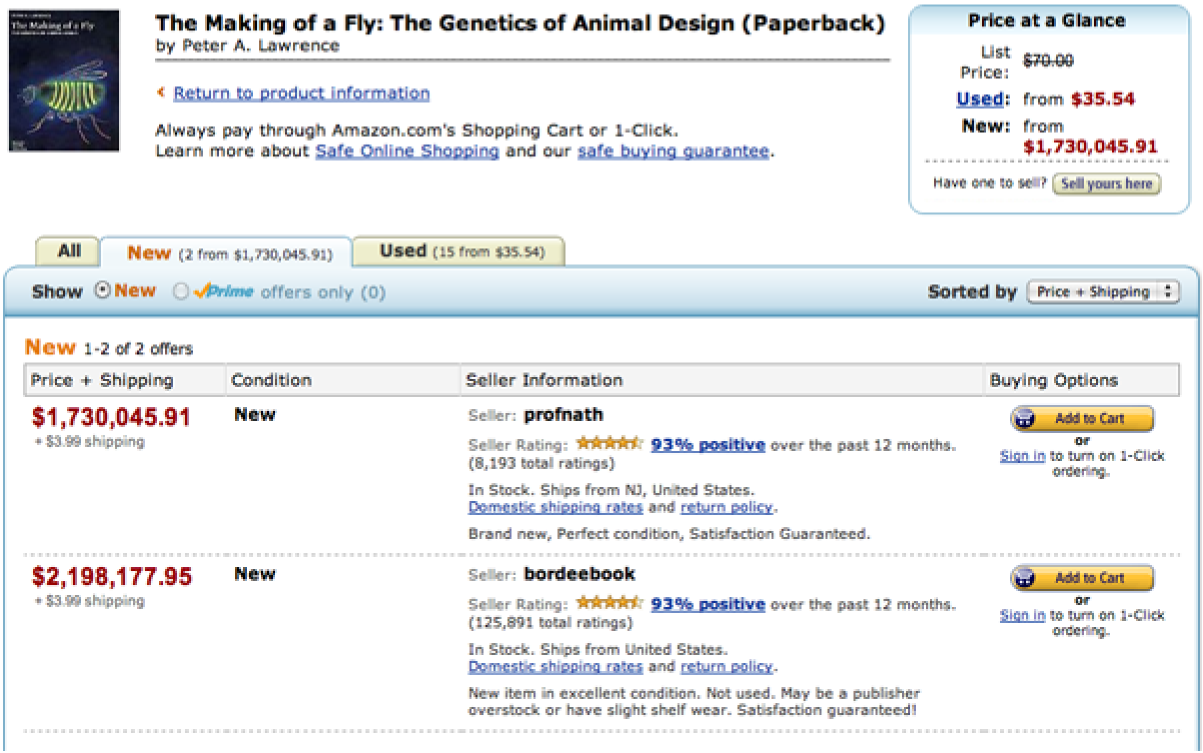
2 Statement of Problem

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.



The Making of a Fly: The Genetics of Animal Design (Paperback)
by Peter A. Lawrence

‹ Return to product information

Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

Price at a Glance
List Price: ~~$70.00~~
Used: from **$35.54**
New: from **$1,730,045.91**
Have one to sell? Sell yours here

All | **New** (2 from $1,730,045.91) | **Used** (15 from $35.54)

Show ● **New** ○ ✓Prime offers only (0) **Sorted by** Price + Shipping

New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information	Buying Options
$1,730,045.91 + $3.99 shipping	**New**	Seller: **profnath** Seller Rating: ★★★★½ **93% positive** over the past 12 months. (8,193 total ratings) In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.	Add to Cart or Sign in to turn on 1-Click ordering.
$2,198,177.95 + $3.99 shipping	**New**	Seller: **bordeebook** Seller Rating: ★★★★½ **93% positive** over the past 12 months. (125,891 total ratings) In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher overstock or have slight shelf wear. Satisfaction guaranteed!	Add to Cart or Sign in to turn on 1-Click ordering.

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

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[5], with nothing but a big red button, labelled 'stop'"*. 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 cognitive trust[1] (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[6], 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

What work's been done already?

Follow up question: what work hasn't been done yet?

4 Proposed Approach

What's my method for tackling the problem?

5 Work Plan

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

References

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

[2] 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.

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

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

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

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