Rendering unto Cæsar the Things that are Cæsar's: Complex Trust Models and Human Understanding

Stephen Marsh[1], Anirban Basu[2], and Natasha Dwyer[3]

[1] Communications Research Centre,
3701 Carling Avenue, P.O. Box 11490, Stn. H. Ottawa, ON K2H 8S2, Canada
steve.marsh@crc.gc.ca
[2] Tokai University, 2-3-23 Takanawa, Minato-ku, Tokyo 108-8619, Japan; and
School of Informatics, University of Sussex, Brighton BN1 9QJ, UK
abasu@cs.dm.u-tokai.ac.jp, a.basu@sussex.ac.uk
[3] Victoria University, Footscray Park Campus, Ballarat Road, Footscray, 3011, Australia
natasha.dwyer@vu.edu.au

Abstract. In this position paper we examine some of the aspects of trust models, deployment, use and 'misuse,' and present a manifesto for the application of computational trust in sociotechnical systems. Computational Trust formalizes the trust processes in humans in order to allow artificial systems to better make decisions or give better advice. This is because trust is flexible, readily understood, and relatively robust. Since its introduction in the early '90s, it has gained in popularity because of these characteristics. However, what it has oftentimes lost is understandability. We argue that one of the original purposes of computational trust reasoning was the human element – the involvement of humans in the process of decision making for tools, importantly at the basic level of understanding why the tools made the decisions they did. The proliferation of ever more complex models may serve to increase the robustness of trust management in the face of attack, but does little to help mere humans either understand or, if necessary, intervene when the trust models fail or cannot arrive at a sensible decision.

1 Introduction

Computational Trust is the study, formalization and implementation of human trust in computational settings [1]. It was intended to be used by autonomous systems of all types in their missions to perform their duties with humans in mind. This qualification is both interesting and important – there is an implicit, at least, expectation that the systems in question are working for humans, perhaps where humans co-exist with the systems themselves. Thus, we can imagine autonomous agents or devices in eCommerce scenarios [2,3], Smart Cities [4,5], Information Systems [6] and ubiquitous computing in general [7,8]. We have seen trust explored for example in user interfaces [9,10] and 'real world' marketplaces [11]. A recent paper [12] tellingly did not build any trust model, noting that current trust models are inappropriate. Trust, of course, is everywhere people are, and if the people are using technology, then it makes sense, at some level, to allow trust to be used there too.

Trust has unquestionable utility – it has been used by humans for millennia to manage risk. Of course, humans have issues estimating and handling risk [13], a subject beyond the scope of this paper, but we could argue that this is another reason for trust being so important to people. People make trusting decisions, for better or worse, almost in the blink of an eye, and adapt these decisions over the lifetime of a relationship [14]. We should mention that, of course, they make mistakes all the time, but one of the strengths of trust is the implicit acknowledgment of the potential for mistakes in the face of incomplete information, and the fact that it is used

anyway and still manages to work both at an individual and a societal level [15]. Perhaps one triumph of irrationality over rationality.

We are faced with something of a dilemma in the user of trust in technology, however. There is a noticeable trend to more and more complex models, using deeper mathematical techniques and constructs. Complex models (involving trust or related reasoning) have been applied in places as diverse as eCommerce [2,16] and Mobile Ad-Hoc Networking [17], through to financial systems (for instance in algorithmic trading). The use of such complex models in systems where time is of the essence, or even where attacks are both prevalent and can have far-reaching effects, is *potentially* both sensible and timely, but the failures can be interesting, to say the least – witness the Flash Crash of May 6th, 2010 for instance.[4] The use of 'trust' models in these instances is troubling to us because they explicitly preclude the human element, and it is the human element that makes trust trust.

This paper examines the purposes of computational trust, its human element, and the fact that, without the human element, there is a lost link. It goes on to argue that with increasing complexity the human element is ever-more difficult to enlist, even in circumstances where, as almost certainly must happen, the trust model cannot accomplish its task. We argue that the use of the term 'trust' in such systems is misleading, and as such potentially dangerous, and that the increasing complexity of the models does not advance the field or its understanding or applicability in systems where it can be at its best.

2 The Need for Models

Trust has long been seen as a tool for the reduction of complexity [18]. It allows people to exist in complex societies, where there are far too many considerations of what *might* go wrong, by taking certain things 'on trust' and not considering them. We can extend this to computational infrastructures in various interesting ways. Developers and creators of ICT systems are encouraged to engage trust, and solve the trust 'problem,' quickly because trust, in this mode reduces uncertainty. This in turn can speed up transactions and 'grease the wheels' of e-commerce [19]. The complexity-reduction nature of trust is important here because, if trust is not resolved, a user can stay in a cycle of exploring possibilities [20]. Since trust results in the foreclosure of some future possibilities [21], this can remove the need to consider them, and results in a more satisfying, or at least more quickly engaging, experience. This is why trust is often viewed as a type of confidence, even when the confidence is not well placed. Distrust is at least as important in this respect also. Often seen as a negative attribute, distrust can in fact holds benefits for similar reasons. Distrust can also resolve a complex scenario, closing down possible paths for the individual to balance [22]. For this reason, an understanding of both trust, distrust, and everything inbetween is needed [23,24]. Perhaps ironically, whilst trust models can help create systems that more accurately follow humans, ostensibly the topic of this paper, they can also, properly executed, help in developing a greater understanding of the phenomenon *in* humans [1].

That said, we arrive at a juxtaposition where models in the technical setting have become increasingly complex, and as has in fact generally been the case, overly contextual and applicable in narrow domains. We see these as problems. The latter problem, perhaps, feeds the former, and is in a cycle fed by it. One reason that they are problems is that they are incomprehensible to the very humans that they, in *all* cases, must inevitably serve [25], but another is that they are not, in the final analysis, approximations of trust at all.

Models of any type, be it trust or anything else, helps us simplify our understanding of a phenomenon in natural sciences or social sciences. Empirical models are expected to be built on

[4] See, e.g. http://www.iosco.org/library/pubdocs/pdf/IOSCOPD354.pdf

actual observations of large samples of data. We argue that models of a 'real' phenomenon (in this case, trust) should be, as close as possible, a fit to that real phenomenon. This is, perhaps obviously, because we would wish to use those models so as to confidently explain (and in some instances predict with some accuracy) future occurrences of the same phenomenon. We can happily accept that models of complex phenomena must in themselves be inherently complex, but we do not accept that models should be so complex as to be beyond the understanding of their creators, or, most tellingly in the case of models used to interact with humans in sensible ways, beyond the understanding of the very people that are intended to use them.

3 Where the Models Work, and Where They Don't

Trust models perform best when a general and abstract understanding of a context and associated trustworthiness of others is required, rather than detailed advice about a specific situation. For instance, the reputation system of Ebay can provide an adequate picture of whom the most trustworthy parties are – who deliver goods on time, who do not overcharge, whose goods are as described and of consistent expected quality, for instance. However, the model cannot say whether a user should definitely trust a particularly seller in a specific situation, particularly if the parties in question do not meet the measurement standards of the system. Consider for example a buyer who has found a rare item. The buyer may greatly desire the item, but the seller has a low rating. In this instance, the buyer can override caution and 'trust despite the low reputation' – Ebay gives reputation values that the buyer can use to ascertain their own trust levels. Since here the trade off for the reward is great, the trust calculation is weighed by the rarity of the item. Thus, Ebay can only guide to a certain point, and we have to be aware that trustworthiness (or reputation as calculated by Ebay) is no more than a guide in this respect.

Models also work when the users concern is a predictable one shared by the majority of others and does not stray far from the objectives others have. Models are not good at handling nuance. For instance, the movie recommender system Netflix can only work to a certain extent. For instance, over time and with experience, a user's taste changes [26]. Models encourage conformity, which in turn, creates conditions for the model to work, creating a self supporting system. Ahn and Esarey [27] explain that systems that provide some sort of information about what is expected as trustworthy behaviour (for instance, Wikipedia), actually foster trustworthy behaviour because users can learn from the system.

We find ourselves here in danger of a cardinal sin for trust and reputation systems – a conflation of trust and reputation – and this is an ideal example for us to approach and consider from the point of view of this paper – that an adequate understanding of what the model is trying to do is absolutely necessary for its successful use. Bearing in mind that trust decisions are based on two questions – how much do you have, and how much do you need (cf [1]) – the next steps should be clear in any evidence-gathering approach. Consider embedded reputation systems. In all instances of such systems the system is informing its user of the *reputation* of the person or thing they are considering. Reputation is a societal judgment of the past behaviour (hopefully in context) of the thing being considered. It is not an indication of the amount that the potential trustee can trust the other. For example, building from Allman et al. and others [28,29], [30] suggested the formulation of a reputation framework for network clients based on their behavioural histories in order to inform service providers to make decisions about future service provision. Although, on a continuous numeric scale $[-1 \quad 1]$ for the sake of comparability, the contextual reputations developed therein did not equate to or imply *trust* at the level of human connotation. Rather, the reputation was only an indicator. What service providers would do with such reputation would be policy specific, and outside the remits of the reputation framework.

In this context, we should also look at inherent challenges with numerical comparability. While during most part of our daily lives, we deal with numbers, we also do so with agreed upon (within some "community") standard units. These units are nothing but enforcements of connotations of numerical values. For example, a 500ml. bottle of water anywhere in the world will contain very close to 500ml. of water (excluding the minor measurement errors) because milliliters is an already agreed upon standard measurement unit. This is good because that helps our society to function. But, assigning numbers to anything without units takes away the very essence of comparability for which we resort to using numbers in the first place. Even if we suggest that such numerical values are only indicators that people should use to inform their trust decisions, the indications carry little meaning in the absence of comparability. Since, numbers are really hard to assign and units are even harder to define in such cases, perhaps, we should call on the use of more qualitative comparators, e.g. partial order. For example, the reputation of A is higher than B and the reputation of C is also higher than B gives us a reasonable (at least to human mind) qualitative indicator and stops us from making the mistake that reputation of A is equal to C. Humans will, inevitably, tend to make that mistake of comparison if they are presented with concrete numbers. Then the point about trust with humans is that we can safely ignore the internal workings of the human mind when it comes to making a trust based decision, and infer those workings based only on observation. On top of that, we also envisage that if we were to ask about such trust decisions from artificial agents, we should be able to get understandable explanations.

The reason for considering this is the following: models work best when they are best understood. Part of the responsibility for that understanding is the responsibility of the person using the system, that is clear. However, part of it must rest with the system itself, and how it is represented to the person. If reputation is shown as a 'trust' value, it is of little surprise when the sinful conflation occurs and mistakes are subsequently made. On the other hand, it is equally unsurprising to find that overly complex models for trust-reasoning technologies result in diminished, or zero, understanding by their target users, with, we suggest, a corresponding lack of engagement. In the next section we explore this problem further, and look at the concept of 'foreground trust' [31,25], which aims to bring trust right into the user interface by encouraging users to make their own decisions based on the evidence.

4 Putting People First

Trust and reputation models are there for people in much the same way as technologies are there for people – they make certain things more straightforward, or more easy. The implicit assumption that we are making is that all technologies are deployed in order to help, in some way, at least a subset of people. Thus, there is always a human in the loop [32] of technological advancement and deployment. It is indeed difficult to find instances of technology where this is not in some sense, the case – and anything that might be found begs a question, we conjecture, as to its worth. That said, we acknowledge the fact that some systems may have people so far from the immediate consideration that they are effectively ignored. That said, we propose that any system that uses trust or reputation must explicitly acknowledge and make room for the human element. If this is the case, it follows that the models must have not only predictive power, but explanatory power also.

4.1 The User Interface

User interfaces are important. They convey system states, valuable information, environmental states, history and predictions to users. In the case of trust (and reputation) systems, they

convey recommendations, perspectives, decisions (or, better, suggestions) evidence and requirements. They might also ask for information in order to better do their jobs. They should also ask for help when necessary. Their design is vitally important. The terminology they use, as discussed in the previous section, has power to enlighten or confuse. Most especially, if complex models are used in the background for whatever reason, the user interface has a role to make the model understandable without losing any of its predictive power. Thus, if we consider models of the weather, an interface showing movement of air masses, and associated weather patterns, is infinitely more accessible to anyone than a mass of numbers on a screen, or the output of some set of algorithms. The same must be true for trust models.

4.2 When Systems Fail, People Pick up the Pieces

Trust models and their operationalized reputation and trust management systems are ideal targets for attacks and misunderstandings. Whilst the attacks possible on reputation systems are relatively well understood, we would argue that spotting these attacks autonomously is difficult. There are nuances to attacks (hence their efficacy) that computerized systems are not good at spotting. Humans, however, are much better at spotting anomalies in trust reasoning and reputation. And it is to people that we must always turn when at the edges of trust – which include boundaries close to thresholds, cold-start problems, and possible ongoing attacks as well as less stress-laden situations (see e.g. [33]).

Designing trust systems that hide the way in which they work must result in misunderstandings, misinterpretations, and mistakes by the very people that the systems are serving. This is guaranteed to result in dissatisfaction and disengagement.

5 Is it Trust?

There are countless ways of defining trust, but most definitions engage with the notion of confidence and risk and there is an acknowledgement that there is a sense of the unknown embedded in the concept of trust. Möllering [34] goes as far to argue that there is something magical about trust, otherwise a concept is not trust. Thus models that claim to have removed uncertainty as much as possible are removing trust. If a situation is clear and the future outcome known, then trust is no longer the issue and the relevant ways of understanding a scenario is better described with notions such as control and security.

Making a trust model more complex does not solve the problem when a model interfaces with human reality. A trust model could dictate how much information a user is safe to reveal to others. In reality, humans are swayed by other factors. For instance, most people are aware that a birth date is information required to verify ones identity and most people are aware of identity theft. Even though there is this awareness, a large number of people put their birth date on their social networking profile presumably so that others can wish them a happy birthday. Vanity wins over sensible behaviour. And who can say that what people do in reality is wrong? To do so is to make a value judgment that prejudices rationality. So any trust model that tries to dismiss how humans actually work is not engaging with human reality. A trust model that can function without humans needs to be able to compute ambiguity; unpredictable irrationality, a continual state of uncertainty and lack of clarity, without any claims on what is a correct position. Ambiguity is different to complexity. Complexity is when a model can process a large amount of variables and conditions.

There are some human phenomena that cannot be short-cut [35], and we argue that trust is one these concepts. In the context of on-line dating, Stainer et al. [12] demonstrate how a machine can calculate a trust interaction, but when it comes to two people interacting, that

calculation can mean very little. While not claiming that humans have a monopoly on trust, we argue that machines are not able to solely process trust without the input of humans. This is because trust is both a rational and irrational phenomena. Trust is a grey concept rather than a black and white, binary position. Computers are renowned for dealing with calculations quickly. Humans are schooled at dealing with ambiguity. Trust models that can incorporate the best of both human and machine work will excel.

6 Render unto Cæsar: a Manifesto

Trust models are approximations of human trust, and as such should be used sensibly and designed so that, whenever possible, humans can be involved. Our basic premise throughout this paper is that any system deployed in the 'real world' in fact influences, is influenced by, and/or works on behalf of humans (whether they like it or not!). When we consider this, we can examine a set of requirements for any trust model that may be designed and deployed in this world. We do not expect this is a complete list, and would expect it to evolve as our understanding of computational trust evolves.

(1) The model is for people.
(2) The model should be understandable, not just by mathematics professors, but by the people who are expected to use and make decisions with or from it.
(3) Allow for monitoring and intervention. Understand that a human's conception of trust and risk is difficult to conceptualise. Many mathematical and economic models of trust assume (or hope for) a 'rational man' who makes judgments based on self-interest. However, in reality, humans weigh trust and risk in ways that cannot be fully predicted. A human needs to be able to make the judgment.
(4) The model should not fail silently, but should prompt for and expect input on 'failure' or uncertainty.
(5) The model should allow for a deep level of configuration. Trust models should not assume what is 'best' for the user. Often design tends to guide users towards what the owner or developer of the site thinks what people should be doing [36]. However, only the user can make that call.
(6) The model should allow for querying: a user may want to know more about a system or a context. A trust interface working in the interest of the user should gather and present data the user regards as relevant. Some of the questions will be difficult for a system to predict and a developer to pre-prepare, so a level of dynamic information exchange is necessary.
(7) The model should cater for different time priorities. In some cases, a trust decision does need to be made quickly. But in other cases, a speedy response is not necessary, and it is possible to take advantage of new information as it comes to hand. A trust model working for humans needs to be able to respond to different timelines and not always seek a short-cut.
(8) The model should allow for incompleteness. Many models aim to provide a definitive answer. Human life is rarely like that. A more appropriate approach is to keep the case open; allowing for new developments, users to change their minds, and for situations to be re-visited.

7 Conclusions

Trust models are becoming more prevalent, applied to many places in many different contexts. Models can be applied in narrow contexts or be much more generic and descriptive (see for example [37]), but we suggest in this paper that models of trust, pure or applied, are, as part of sociotechnical systems deployed in the world at a concrete level, human-oriented. If this is

the case, computational trust models are only useful in situations where the people who interact with them *understand* them in a reasonable way. It is unsatisfactory to have eCommerce models, for instance, that use mathematical tools that are indecipherable to people who use eCommerce systems (and agents) because the ultimate endpoint is then a shift of trust from the other person to the model (I will do what it says because I can't figure out why, and it's smarter than I am) – this is not what computational trust in its origin was intended to achieve. Without human understanding and focus, trust models are not trust, but a mere statistical, probabilistic or other mathematical approach to uncertainty. We have provided a small manifesto for computational trust models that we hope can be of some service to the community – as a discussion point, as the start of a set of requirements above and beyond 'does it work in this instance, or against this attack?' and as a reminder that, (to paraphrase Einstein) in all of our considerations, humans are the root.

References

1. Marsh, S.: Formalising trust as a computational concept. PhD thesis, University of Stirling, Department of Computing Science and Mathematics (1994)
2. Ping, W., Jing, Q.: A mathematical trust model in e-commerce. In: 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07). (2007)
3. Noorian, Z., Marsh, S., Fleming, M.: Multi-layer cognitive filtering by behavioural modeling. In Tumer, Yolum, S., Stone, eds.: Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011). (2011)
4. Mahizhnan, A.: Smart cities:: The singapore case. Cities **16**(1) (1999) 13–18
5. Shapiro, J.: Smart cities: Quality of life, productivity, and the growth effects of human capital. Technical report, National Bureau of Economic Research (2005)
6. Li, X., Valacich, J.S., Hess, T.J.: Predicting user trust in information systems: A comparison of competing trust models. In: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 8 - Volume 8, Washington, DC, USA, IEEE Computer Society (2004)
7. Shankar, N., Arbaugh, W.A.: On trust for ubiquitous computing. In: Workshop on Security in Ubiquitous Computing, UBICOMP 2002, IEEE Computer Society (2002) 44–54
8. Sillence, E., Briggs, P.: Ubiquitous computing: Trust issues for a "healthy" society. Social Science Computer Review **26**(1) (February 2008) 6–12
9. Riegelsberger, J., Sasse, A., McCarthy, J.D.: Shiny happy people building trust?: photos on e-commerce websites and consumer trust. In: Proceedings of the 2003 conference on Human factors in computing systems. (2003) 121–128
10. Riegelsberger, J.: Trust in Mediated Interactions. PhD thesis, University College London (2005)
11. Wakeman, I., Light, A., Robinson, J., Chalmers, D., Basu, A.: Bringing the virtual to the farmers' market: Designing for trust in pervasive computing systems. In Nishigaki, M., Josang, A., Murayama, Y., Marsh, S., eds.: Proceedings IFIP Conference on Trust Management (IFIPTM 2010), Morioka, Japan. (2010)
12. Stanier, J., Naicken, S., Basu, A., Li, J., Wakeman, I.: Can We Use Trust in Online Dating? In: Proceedings of the International Workshop on Trusted Communications in Decentralised Computing (workshop in IFIPTM 2010), Morioka, Japan. (2010)
13. Zeckhauser, R.J., Viscusi, W.K.: Risk within reason. Science **248** (May 4th 1990) 559–564
14. Boon, S.D., Holmes, J.G.: The dynamics of interpersonal trust: resolving uncertainty in the face of risk. In Hinde, R.A., Groebel, J., eds.: Cooperation and Prosocial Behaviour. Cambridge University Press (1991) 190–211
15. Bok, S.: Lying: Moral Choice in Public and Private Life. Pantheon Books, New York (1978)
16. Zhang, Z., Zhou, M., Wang, P.: An improved trust model in agent-mediated ecommerce. Int. J. Intell. Syst. Technol. Appl. **4**(3.4) (February 2998) 271–284

17. Sun, K., Xu, R., Deng, J., Haynes, L., Li, J.H., Gruenwald, L., Sanchez, C., Weber, G., Mayhew, M.J.: Securing manet databases using metadata and context information. In: Proceeedings of MILCOM 2008: Military Comunications Conference. (2008) 1–6
18. Luhmann, N.: Trust and Power. Wiley, Chichester (1979)
19. Fukuyama, F.: Trust: The social virtues and the creation of prosperity. Free Press, New York, USA (1995)
20. Cofta, P.: Trust, Complexity and Control. Volume 829528313. John Wiley and Sons, New Jersey, USA (2007)
21. Goffman, E.: Frame analysis. Harvard University Press, Cambridge, USA (1974)
22. Cofta, P.: Distrust. In: Proceedings of Eight International Conference on Electronic Commerce. (2006)
23. Marsh, S., Dibben, M.R.: Trust, untrust, distrust and mistrust — an exploration of the dark(er) side. In Herrmann, P., Issarny, V., Shiu, S., eds.: Trust Management: Proceedings of iTrust 2005, Springer Verlag, Lecture Notes in Computer Science, LNCS 3477 (2005)
24. McKnight, D.H., Chervany, N.L.: Trust and distrust definitions: One bite at a time. In Falcone, R., Singh, M., Tan, Y.H., eds.: Trust in Cyber-Societies. Volume 2246 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, Heidelberg (2001)
25. Marsh, S., Noël, S., Storer, T., Wang, Y., Briggs, P., Robart, L., Stewart, J., Esfandiari, B., El-Khatib, K., Bicakci, M.V., Dao, M.C., Cohen, M., Silva, D.D.: Non-standards for trust: Foreground trust and second thoughts for mobile security. In: Proceedings STM 2011, Springer (2012)
26. Lathia, N.: Evaluating collaborative filtering over time. PhD thesis, University College London (2010)
27. Ahn, T., Esarey, J.: A dynamic model of generalized social trust. Journal of Theoretical Politics **20**(2) (2008) 151–180
28. Allman, M., Blanton, E., Paxson, V.: An Architecture for Developing Behavioral History. In: Proceedings of the Workshop on Steps to Reducing Unwanted Traffic on the Internet. (2005)
29. Wei, S., Mirkovic, J.: Building Reputations for Internet Clients. Electronic Notes Theoretical Computer Science **179** (2007) 17–30
30. Basu, A.: A Reputation Framework for Behavioural History. PhD thesis, University of Sussex, UK (January 2010)
31. Dwyer, N.: Traces of Digital Trust: An Interactive Design Perspective. PhD thesis, School of Communication and the Arts, Faculty of Arts, Education and Human Development, Victoria University (2011)
32. Dautenhahn, K., Alan H, B., Canamero, L., Edmonds, B., eds.: Socially Intelligent Agents: Creating Relationships with Computers and Robots. Kluwer Academic Publishers (2002)
33. Kaur, P., Ruohomaa, S., Kutvonen, L.: User interface for trust decision making in inter-enterprise collaborations. In: ACHI 2012 : The Fifth International Conference on Advances in Computer-Human Interactions. (2012)
34. Möllering, G.: Trust, institutions, agency: towards a neoinstitutional theory of trust. Handbook of trust research (2006) 223–233
35. Donath, J.: Signals in social supernets. Journal of Computer-Mediated Communication **13**(1) (2007) 231–51
36. Thaler, R., Sunstein, C.: Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven, USA (2008)
37. Castelfranchi, C., Falcone, R.: Trust Theory: A Socio-Cognitive and Computational Model. Wiley (2011)