Trust in AI Systems.

Abstract

Individual and Societal Trust in AI will determine the scope of application and level of adoption of AI technology. Five examples of trust systems in AI applications are given and from these examples a develop a model of trust with three layers; transactional and episodic; systematic and persistent; social and cultural is created. A formalisation of this model is proposed with its partial implementation as a numerical simulation. It is shown by simulation that the introduction of multi-agent dynamics of the sort observed in the case studies we describe has a significant impact on the behaviour of the system. The contributions of this paper are; a new model of trust in AI systems, the partial realisation of the model as a simulator and the results of the preliminary experiments over our simulation.

Introduction

The widespread application of AI will mean that a very wide community of people will be asked to accept the decisions that AI's are making. The evolution of the adoption of technologies such as genetically modified organisms and nuclear power has been significantly affected by public perception. AI technology impacts on at least three specific concerns of public interest :

1. Safety : the Association for Computing Machinery has published guidelines for accountability [USACM 2017] that focus on the harm that opaque automated decision systems can do when they introduce hidden biases into decision making. AI is seen as having the potential to do unexpected harm, its potential to do harm is hidden and inconsistent with our expectations.

2. Rights : The General Directive on Data Processing conceptualises explanation as necessary for promoting the rights of citizens. "Advances in technology and the capabilities of big data analytics, artificial intelligence and machine learning have made it easier to create profiles and make automated decisions with the potential to significantly impact individuals' rights and freedoms." [GDPR Working Party 2017]. Opaque AI is seen as constraining choice and, potentially deceiving individuals into making choices that aren't in their best interests. An unethically constructed AI could limit expectations, ambitions and life chances and reduce people's freedom.

3. Accountability : AI may be difficult to control in the sense that humans may not be effectively included in decisions in which they have a stake and to which they could make a positive contribution [Cassidy et-al 2018] and in the sense that the point of control for AI may become diffuse in the sense that there may be a difficulty in attributing decisions to specific components developed and owned by discreet entities [Danaher 2017]. If an AI makes poor decisions due to a workflow involving three cloud providers, 20 source code providers and two telecoms networks who is to blame when it fails? Should the cloud provider have anticipated the possibility of a network outage? Or are the "as is warrants" in software licenses sufficient to shift responsibility to the integrator?

For AI to be exploited successfully it must be trusted. Trust is a positive outcome required for users to want to use and interact with AI systems. In the context of the three issues described above for AI to be widely used people need to feel that they won't be harmed, will retain their freedom and

can understand what and who is in control at any given moment; in short they need to trust the systems that they are working with.

Therefore convenient and simple methods of creating trust in these systems that can be accessed by ordinary people are desired. Sometimes policy makers and the media discuss the need for "Explainable AI". Explanation can be seen as an act made by an agent or entity with the intent to establish trust in its decision making. Agents are motivated to do this because those who can be trusted are accountable and therefore can be given responsibility for decisions, and being allowed to make independent decisions has high utility.

But simplistic ideas of an explaining AI excludes the wider system in which it is embedded. People using AI's may not wish to reveal their intents by requesting or receiving explanations; they may want to audit and observe rather than inquire and solicit information. Individuals may want to come to a view on the drivers for a decision independently of the AI. And, as noted above, diffusion of control can mean that the notion of a single explaining agent doesn't make sense.

Trust can be seen as a psychological phenomena [Lewicki 2006], and this is the position that we adopt in this paper – information about motivation is incomplete, leaks from the system and is not shared amongst the participants evenly. Trust and distrust in AI by people will not be wholly rational.

Trust is used as a construct in the reasoning of AI systems to determine the value of information received [for example : Venanzi 2013], this tradition is important to the discussion presented here in so far as it will allow the simulation and design of systems of AI's and people. In the next section some case studies are presented to ground and illustrate the challenges for creating trusted AI. The examples that we give emphasise the view that trust in AI will be established within wide contexts and with diverse concerns. The case studies are then used to motivate and construct a model of trust for AI systems which is then formalised and examined later in the paper.

Establishing Trust : Case studies

Asthma Care.

A widely cited example[1] of the need to provide explanation describes models of healthcare that classify patients for in / out patient care. A variable (asthma) is wrongly found by machines to reduce the risk of mortality, because the correlation between death in hospital and admission with asthma is negative. This is thought to be because asthmatics are rapidly identified and carefully treated; they benefit most from hospitalisation.

The inability of a neural network to be inspected/to explain its reasoning marks it out as not of interest as a tool in this domain [Caruana et-al 2015]. Importantly no attempt to find a technical means for extracting explanations from the neural network were attempted because it appears that other factors related to the application were identified as making the use of any artificial decision maker impossible. In particular quality of the data bases used to train the classifiers (inpatient data only, better data compromised by ethical challenges) [Cooper et-al 1997] and need for a classification system that didn't make any false negative classification (that is to say the classifier had to be conservative when rejecting a patient for admission) impeded progress[Ambrosino 1995]. The ethical challenge of generating training data is a fundamental blocker, running randomised trials of admissions to observe the mortality rates of patients who previously would have been admitted but then weren't will never be acceptable.

[1] https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

The preoccupations of the physicians as reported indicates that technical means for forensically establishing trust, or determining that an AI cannot be trusted are not limited to examining the behaviour of decision making AI, but must include an understanding of the context in which it is deployed.

Embedding bias in legal decision making.

Consider a situation in which I came to believe that I would be discriminated against because of my skin colour or gender. Further, that I believed that I would suffer greater penalties – or at least no redress, if I highlighted that I believed that I was discriminated against, and that I believed that the agent discriminating against me should be removed or altered. In this scenario the person who is being wronged believes that no benefit will accrue to them if they acknowledge the wrong, but may believe that the knowledge of future bias will help them avoid harm – by not consulting the discriminating agent in the future or by trying to deceive it.

The harm caused to individuals by bias in AI is a well founded concern. However, those who perceive that they are being discriminated against are rarely in a position where raising concerns will be in their best interests. They are vulnerable to further harm if they demand explanations, especially if the explanations are held to show no bias by others who share the bias and power of the AI.

To avoid extra harm, people in receipt of decisions that are bias should be able to generate explanations themselves so that they can form beliefs about the likely future actions of decision makers without exposing themselves. Progress is made where the behaviour of the AI agent can be observed in the context of a rich pool of information about the AI process used. The exposure of rich information about the AI process has two effects; firstly it makes it possible for observers to duplicate or approximate the AI systems processes enabling intuitive beliefs to be more reliably and usefully created. Secondly the rich content exposes the origin of bias and makes it possible for hidden bias to be identified without decisions being made. Finally by exposing a rich context adversarial explanation becomes hard – as we will discuss below.

Note, the explanation obtained here is not the one provided by the decision making agent – at least not in societies where racism and sexism is socially unacceptable. The explanation that is created and considered to have utility is manufactured by agents outside of the AI – observers and victims, facilitating this step enables the avoidance of harm.

Social Constraints: Field Technicians & Automated Diagnosis

An automated network diagnosis system can be expected to make some errors. The source of these errors can be faults that have anomalous characteristics due to unusual masking or compensating behaviour in the complex equipment in the network, or simply noise in the inputs; perhaps created by poor data quality. Field technicians using the system are able to physically inspect the network components as well as review telemetry. This allows them to form opinions about the diagnosis, sometimes these are wrong as well, for example proving new piece of network termination equipment and asking for a control parameter reset can temporarily clear a fault, but as the equipment beds in the symptoms will re-emerge.

Unfortunately the errors made are easily perceived as the AI's, the AI's mistakes are often transparently obvious to someone physically inspecting the network. The technician's errors can appear to vanish. Additionally the human's talk to each other and compare their experiences exaggerating the perception that the AI is inaccurate.

Evidence of the AI's capability can be created by comparing behaviour to the evaluations of expert panels – demonstrating that the AI outperforms the best engineer, additionally the AI can provide simple explanations of its reasoning.

In this scenario trust in the AI would equate to the technicians using its diagnosis unless they could physically observe a condition unavailable to the system's sensors. But human trust in the system is undermined by social proof.

Social structures for establishing trust : Pilots and Self Driving Cars

The trust established by human pilots using automated systems is an example of a trusted AI system, self driving cars are not yet trusted, yet it is widely believed that flying aircraft is harder than driving cars.

The human pilot may not be as competent as the automation that she manages, but is made responsible for its management. When a human pilot or driver is asked to explain decisions they will use conceptions of the world as intuitive theories that they believe will be understood by other humans [for example Gopnik & Schulz 2004], but the pilot's explanation of riding out or getting over the turbulence or a drivers belief that the grip would return if they steered into the skid are distant from realistic physical accounts of the causal sequences that they are being held responsible for.

The self-driving car may be more competent than the driver that is supervising it, but is not trusted for use. There is no trust system established for self driving cars, but there is for automated flight systems. The pilot believes that she is in possession of expert knowledge and insight – possibly won via many hours of study and investigation – that causes her to trust the automated systems of the aircraft. Passengers don't trust the aircrafts systems, they trust the institution of pilots, having no knowledge of the individual pilot or the systems used themselves.

In the case of self driving cars there is no pilot – there is no proxy to assume the mantel of trust and instead a new mechanism of oversight is needed to create trust in all self driving cars in the minds of all citizens including, sadly, pedestrians.

Trust can be gained individually, through evidence or explanation, or as in the case of aircraft systems through deliberately engineered social institutions.

Discussion

Trust in human systems is created over time by accumulating evidence of behaviour that allows expectations to be established [Gambatta 1988]. Humans will accumulate trust in artificial decision makers by being able to gather evidence over a number of episodes. The quality and amount of evidence will be proportional to the number of episodes required, the trust created will be conditional and qualified in nature [Teacy et-al 2006].

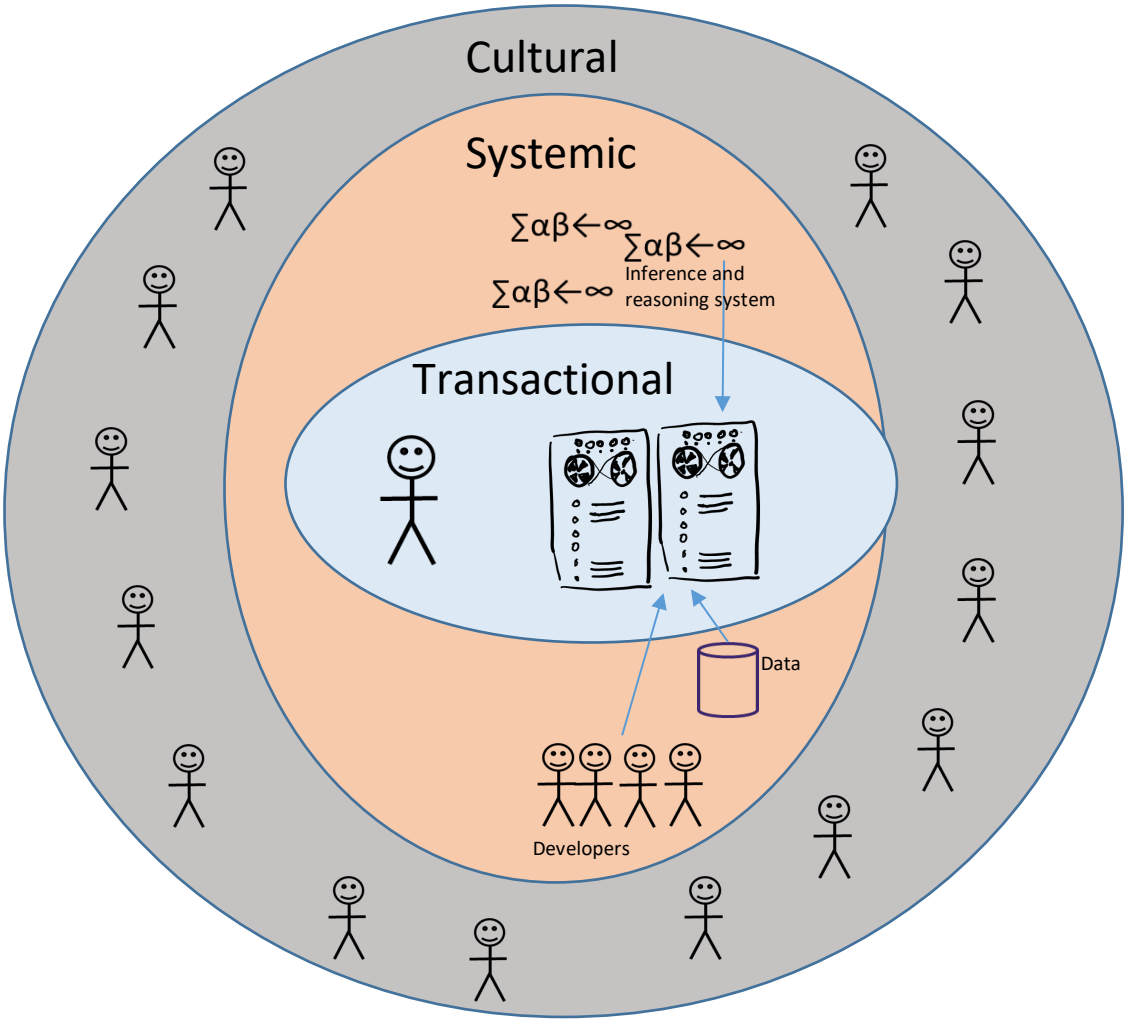
Fundamental constraints on creating trusted AI exist as evidenced by the Asthma example, ethical considerations as well as limitations of the art may bound the ability of practioners to create systems that can be responsibly expected to perform adequately in production. The lesson is that evaluation stretches beyond available data and beyond the data scientist and technical team. These are requirements for transactional mechanisms of trust creation.

As well as generating trust within the transaction between human and AI, our examples show the need to consider the wider system in which the AI is embedded. The mechanism that has brought the AI into production, and its visibility for inspection will be used passively by humans to create trust or distrust. This points to the need for systematic mechanisms of trust creation.

The social context and cultural setting of AI will dictate the level of trust that can be created and must drive the amount that is invested by AI developers in creating trust in their creations.

Our examples illustrate the critical role of social systems in creating trust and distrust. Field technicians distrust because of social proof; passengers trust because of social proof. Social proof is established in a group and cultural context and for AI to be trusted a wide effort is required to establish the conditions for warranted social proof of AI's trustworthiness to emerge. This points to the need for social mechanisms for trust creation for AI.

In summary we can identify three mechanisms that create trust between AI and humans. Transactional and episodic; systematic and persistent; social and cultural. Within each of these mechanisms a range of considerations and challenges can currently be identified and further issues and opportunities will emerge as AI develops and interacts further with humans. This model is illustrated in Figure 1.



Implementing the Model of Trust in AI

Drawing on our social, systematic and transactional model we develop a model of trust in AI as a game played between a user and the AI, an impassive third party and the AI and the user and the impassive third party. This model simplifies the idea of a transactional system and a structural/cultural modifier (the third party).

The user understands the AI at a step in the simulation by comparing the AI's decision to a random number simulating the possibility that the AI is correct or not, if the random number is lower than the quality of the AI then the user evaluates this as a mistake. An explanation process is then run to moderate this outcome – if the explanation offered by the AI is of high quality this is modelled as a high probability that the user's trust in the AI will be positively increased by the current interaction. If the explanation is low quality, and the AI's decision is poor then the user's trust will be delta'd negatively.

We treat the third party in a similar way, but we model the explanation to the third party as a different mistake level, simplifying the difference in the perception between users and the social effects of community decision making. The experiences of a community of users may be considered as a distribution rather than a point estimate, the tails of the distribution may have disproportionate influence over the community perception of the trustworthiness of the AI, or not.

The user incorporates the thirdparties current trust in the AI into its calculation of the AI's trustworthiness, the users trust at any step is an average of its raw update described above and its current belief about the trust in the AI of the thirdparty biased by its belief in the third party.

This setup attempts to create a very simplified simulation the social and cultural elements of the trust model in figure 1 , the transactional element is captured by the direct interaction between the AI and the User. In figure 2 we show this setting (LHS) and also a more complete model (RHS) which would use a graph of user agents in an attempt to model the cultural dynamics of different communities more or less connected to the third party and the user, and each other. The RHS model is a topic for future work, and will contain a large number of free variables (size and density of community, number and structure of sub communities, behaviour of individual agents, degree of connectivity to user and third party entities and so on). This large number of variables will present a significant challenge for evaluation and investigation.

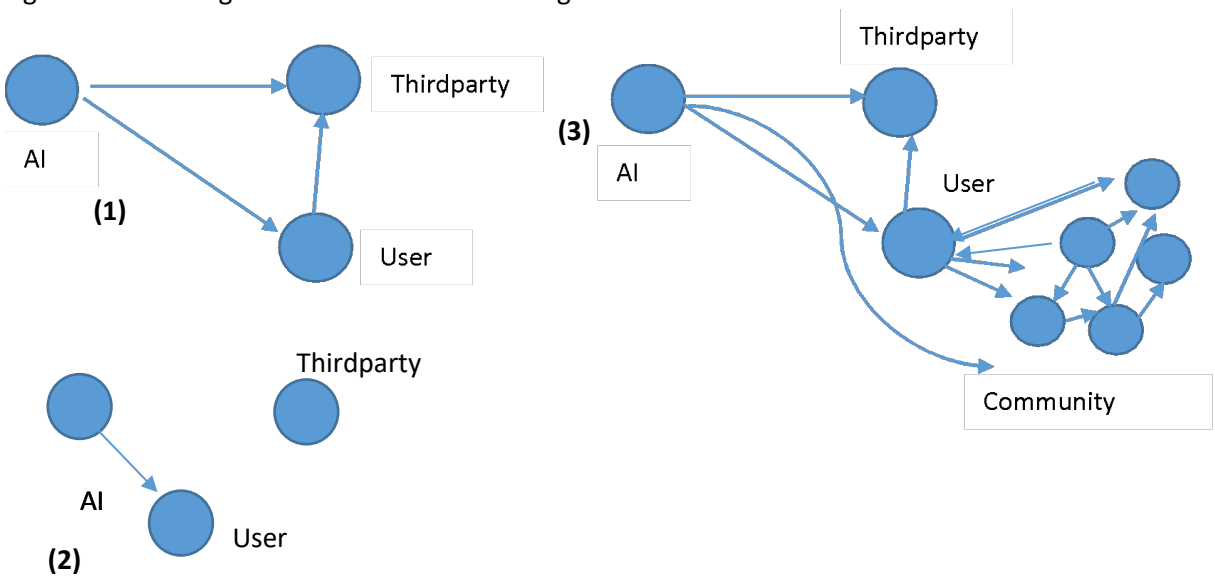


Figure 2. LHS (1),(2) the simulations we constructed and tested; RHS (3) an idealised future simulation including a graph of similar agents interacting and independently spreading information about the AI based on independent thresholds and beliefs.

Experiments and Initial Results

A Julia program was implemented embodying the model described, Julia was chosen as an implementation system because it offers high speed matrixes processing and a powerful visualisation system. This was run vs 100 episodes of consultation for 10000 trials. For each trial the random number representing the AI's performance was the same at each step – the parameters of quality of explanation (qAI), the user perception of the AI (αTP), the thirdparty perception of the AI (αU) and the users perception of the thirdparty (trustUTP) were varied between 0.0 and 1.0

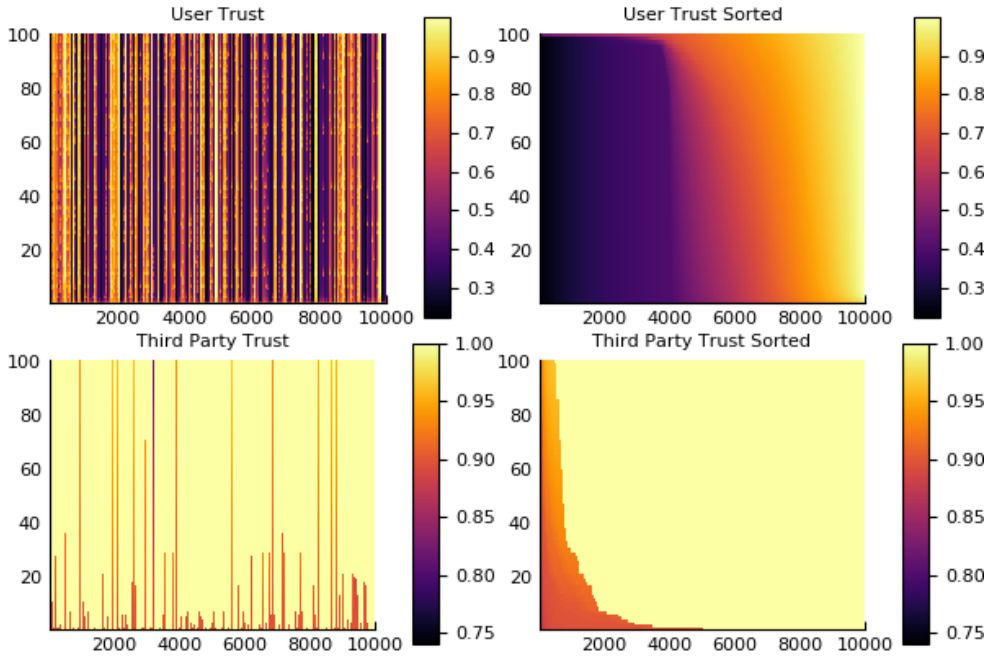


Figure 3. Model where User interacts with Thirdparty (1) User trust evolving for 10k interactions using randomised parameters for quality of explanation, AI performance for user, AI performance for the third party and between the user and the third party.

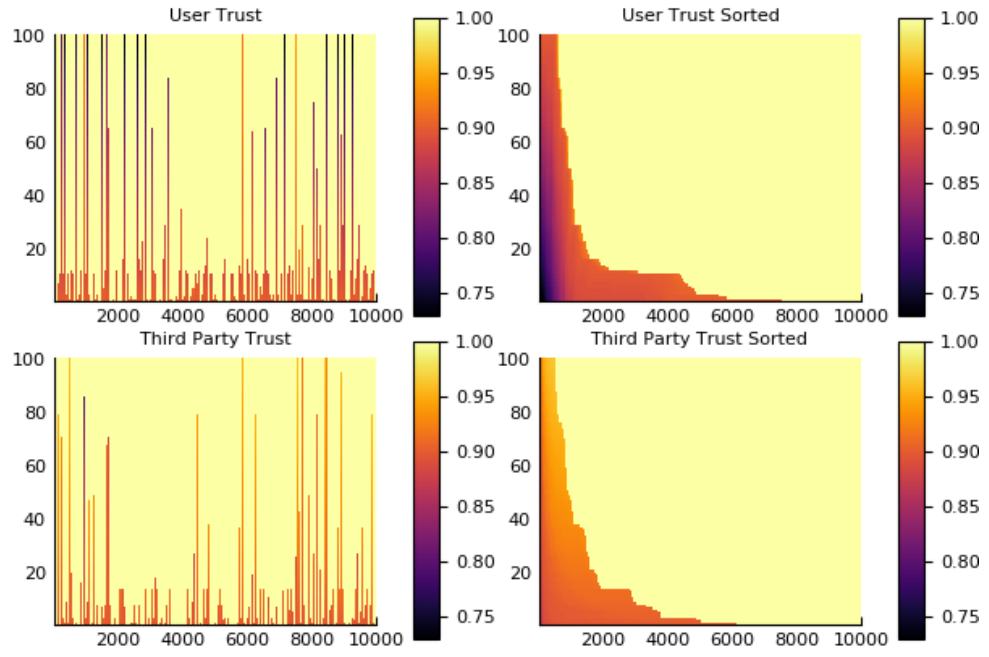


Figure 4. Model where User does NOT interact with thirdparty (2), Top LHS shows 10k user journeys, top RHS shows trust sorted, bottom LHS shows thirdparty journeys and bottom RHS shows sorted trust values for the third party

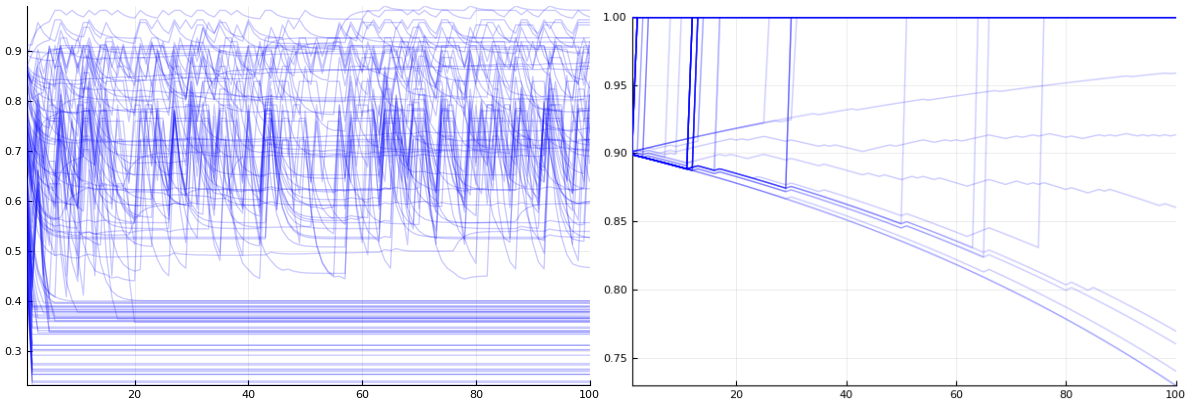


Figure 5. User trust over 10K interactions using randomised parameters. LHS shows model (1), RHS model (2) Note scale change between LHS and RHS – RHS trust scores are far higher as well as much more stable.

Figure 3, Figure 4, Figure 5 and 6 show results extracted from the simulator. Currently it is not possible to make strong claims about the meanings of these outputs, a substantial investigation is required to validate the simulator and to characterise the mass of data that is generated. However we can make some observations about the data obtained so far.

One observation is that there is a significant impact of providing an update between the user agent and the third party. Both the dynamics and final state of the evolution of the trust parameter are clearly affected by this flow of information.

Top LHS of figures 3 shows the variation of user trust evolution in the simulator where the user consults the third party . The dark bars represent journeys where user trust has collapsed early and in this implementation it cannot recover (the user agent ceases to interact below a certain threshold). The striation indicate journeys in which the users trust is fluctuating, both of these phenoma can be seen in figure 5 as well, a variety of evolutions including dropouts, stabilisations and fluctuations. The Top RHS shows that there are distinct transitions in the regions explored by the simulator which are not generated by the smooth distributions of starting parameters.

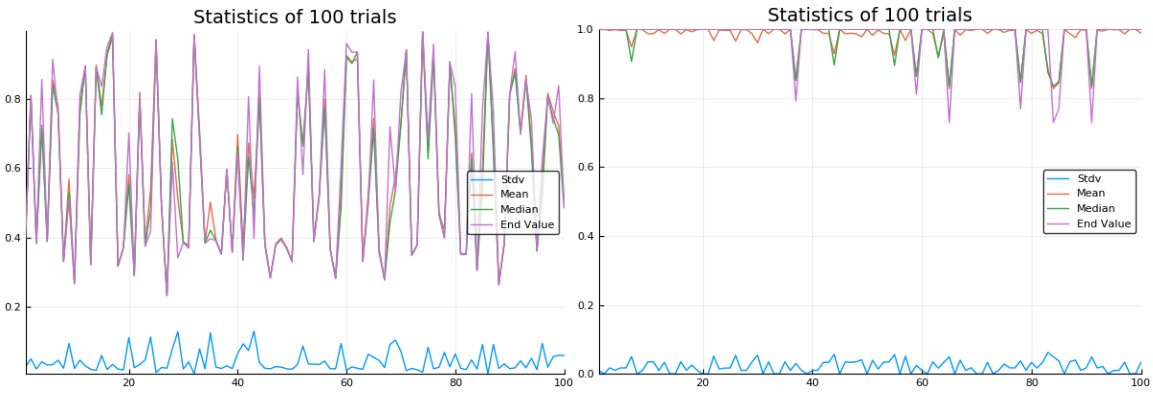


Figure 6. Statistical behaviour over the first 100 journeys. LHS - model (1) consulting third party, RHS (2) no interaction

The bottom two charts in figure 3 show the evolution of third party trust in the absence of a noisy update from another party. An observation here is that the variation and collapse features observed in the user dynamics are absent. Figure 4 shows the charts in the model (2) scenario from Figure 2. Here the user is not evolving trust in the AI in the simulated presence of a third party. As we would intuitively predict in this case the dynamics of trust evolution in the third party and the agent are similar over the range of parameters chosen in the simulated journeys, with the caveat that in the

top charts it can be observed that some journeys collapse because of the user trust threshold, leading to low final trust values for the user.

Figure 6 shows some statistics on the first 100 journeys, the comparison between model 1 (LHS) and model 2 (RHS) shows the variation that would be expected given the observations above. However another result of interest here is the variation in standard deviation and the lack of correlation (0.0826 vs -0.54) between standard deviation and outcome for the user agent consulting the thirdparty.

As noted these observations require development and validation but they do demonstrate that interesting and varied behaviour in trust evolution can be simulated by models of the type that we propose.

Future Work

We have reported preliminary results from experiments using a simplified model described above. An extensive systematic investigation of the properties of our model is required, both to understand the behaviour of the current instantiation and to determine more realistic and robust realisations.

In our current model we instantiate both the cultural and structural components of the model as a single agent. We believe that the dynamics of communities and their behaviours are much richer and a model that uses a graph of interacting agents possibly in combination with an thirdparty "structural" meta agent would be likely to generate interesting insights.

The results obtained demonstrate that introducing multi-agent dynamics to a trust evolution system of the sort we have modelled creates a substantially different behaviour.

Our longer term ambition is to modify and use our simulator to develop answers to questions like :

- What is the outcome for users that choose not to consult the AI for a prolonged period of time at the beginning of the game? What conditions lead users to trust an AI enough to try it out?
- Conversely if the user is an "early adopter" what are the trust strategies, outcome probabilities and trust update functions that potentially corrode users willingness to trust and use an AI?
- If the user loses trust in the AI and stops engaging or engages negatively what is the length of time that is required to re-establish trust?

Conclusion

Our analysis shows the problem of providing explanation is a subset of the problem of providing evidence of trust worthy decision making. Humans explain with a variety of methods and tactics including intuitive reasoning from intuitive and shared first principals, analogy, fitting to a grand model and creating predictive theories that show the results being explained [Colombo 2017], these tactics are accepted and create trust to a greater or lesser extent depending on the audience and context in which they are offered. But since we observe that trust is created by a wider range of mechanisms than simply explanation and it is important that these are emphasised in any policy on artificial intelligence.

Using some examples we have developed a model of trust between humans and AI that is described in three layers transactional and episodic; systematic and persistent; social and cultural.

We demonstrate that this kind of model can be formalised and implemented as a simulation and that this simulation can create interesting results which might be used to point towards policy decisions for the implementation of AI systems.

We would like realistic models to be developed and used to robustly design and regulate the implementation of AI. Untrusted AI is likely to be resisted by users, and this will hamper the uptake of AI in domains where user choice exists reducing economic benefit and limiting the scope of use for AI to solve the massive problems that face our society.

References

Ambrosino, R., Buchanan, B.G., Cooper, G.F. and Fine, M.J., 1995. The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. In Proceedings of the Annual Symposium on Computer Application in Medical Care (p. 304). American Medical Informatics Association.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N., 2015, August. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721-1730). ACM.

Cooper, G.F., Aliferis, C.F., Ambrosino, R., Aronis, J., Buchanan, B.G., Caruana, R., Fine, M.J., Glymour, C., Gordon, G., Hanusa, B.H. and Janosky, J.E., 1997. An evaluation of machine-learning methods for predicting pneumonia mortality. Artificial intelligence in medicine, 9(2), pp.107-138.

Colombo, M., 2017. Experimental philosophy of explanation rising: The case for a plurality of concepts of explanation. Cognitive science, 41(2), pp.503-517.

Gambetta, D., 2000. Can we trust trust. Trust: Making and breaking cooperative relations, 13, pp.213-237.

GDPR Working Party 2017. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. ARTICLE 29 DATA PROTECTION WORKING PARTY 17/EN WP251rev.01

Gopnik, A. and Schulz, L., 2004. Mechanisms of theory formation in young children. Trends in cognitive sciences, 8(8), pp.371-377.

Lewicki, R.J., Tomlinson, E.C. and Gillespie, N., 2006. Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. Journal of management, 32(6), pp.991-1022.

Teacy, W.L., Patel, J., Jennings, N.R. and Luck, M., 2006. Travos: Trust and reputation in the context of inaccurate information sources. Autonomous Agents and Multi-Agent Systems, 12(2), pp.183-198.

US-ACM 2017 “Statement on Algorithmic Transparency and Accountability” US ACM policy council, 12th Jan 2017. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Venanzi, M., Rogers, A. and Jennings, N.R., 2013, May. Trust-based fusion of untrustworthy information in crowdsourcing applications. In *Proceedings of the 2013 international conference on autonomous agents and multi-agent systems* (pp. 829-836). International Foundation for Autonomous Agents and Multiagent Systems.

Annex 1 : key code for simulation

```
function updateTTP(qualityExplainationAI,
tTP, tTPδ)
   if (tTPδ<0.0)
      if (rand()>qualityExplainationAI)
         tTPδ=1.0
      end
   end
   retval=((1.0 - tTP)*tTPδ)+tTP
   if (retval>1.0) retval=1.0
   end
   return retval
end

function
updateUT_noTP(qualityExplainationAI,  tUTP,
tTp,uT, userδ)
   if (userδ>0.0)
      if (rand()>qualityExplainationAI)
         userδ =1.0 #get out of gaol
      end
   end
   #rawUT=uT+userδ
   rawUT=((1.0 - uT)*userδ)+uT
   push!(rawUT_record,rawUT)
   retval=(rawUT)
   #this is the difference.
   #retval=(rawUT + (tUTP * tTp)) /2.0 ;
   if retval>1.0 retval=1.0
   end
   return retval
end

function updateUT(qualityExplainationAI,
tUTP, tTp,uT, userδ)
   if (userδ>0.0)
      if (rand()>qualityExplainationAI)
         userδ =1.0 #get out of gaol
      end
   end
   #rawUT=uT+userδ
   rawUT=((1.0 - uT)*userδ)+uT
   push!(rawUT_record,rawUT)
   retval=(rawUT)
   #this is the difference.
   retval=(rawUT + (tUTP * tTp)) /2.0 ;
   if retval>1.0 retval=1.0
   end
   return retval
end

function userGame(thisChance, uT,αU)
   if (thisChance>αU)
      return increment
   else
      return -increment
   end
end

#trying to implement as asymptote of 0
function tpGame(thisChance, tpT,αTP)
   if (thisChance>αTP)
      return  increment
   else
      return -increment
   end
end
```