Trusted AI and the Contribution of Trust Modeling in Multiagent Systems

Blue Sky Ideas Track

Robin Cohen, Mike Schaekermann, Sihao Liu, Michael Cormier
Computer Science; University of Waterloo; Waterloo, Canada
{rcohen,mschaeke,s367liu,m4cormier}@uwaterloo.ca

ABSTRACT

Researchers in the field of artificial intelligence today are increasingly concerned with whether the systems which they build will be "trusted AI", in other words, whether they will be accepted by their human users. The claim of this paper is that these researchers should be aware of the rich set of solutions being developed in the multiagent systems subfield of trust modeling. We propose a specific perspective on how to leverage trust modeling solutions towards assurances for trusted artificial intelligence. We conclude by advocating for greater dialogue between these AI communities.

ACM Reference Format:
Robin Cohen, Mike Schaekermann, Sihao Liu, Michael Cormier. 2019. Trusted AI and the Contribution of Trust Modeling in Multiagent Systems. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 5 pages.

1 INTRODUCTION

The topic of this Blue Sky paper is the future of multiagent systems trust modeling research and its community, within the broader AI landscape. The visionary idea is that trust modeling research can be leveraged in order to promote a more principled construction of efforts aimed at promoting trusted AI. A community of multiagent systems researchers have been developing for a long time now models to determine the trustworthiness of peers (typically based on a kind of probabilistic reasoning and learning over time (witness the beta reputation function of Josang and Ismail [15], which has become central to many solutions developed to date)). This modeling has been done in order to direct agent decision making in multiagent environments. Trusted AI researchers are focused on a set of issues which might lead users of their systems to lack trust in the solutions being delivered, ones which control decisions being made for those users. Such issues include: fairness, bias, explanability and appropriate consideration of ethics. One aim of this Blue Sky paper is to make the two communities more aware of the existence of the other [1]. The tremendous effort being devoted currently to examining the issue of trusted AI can be seen with the special emphasis at last year's AAAI conference, and the sudden emergence of a separate conference dedicated to AI, ethics and society, affiliated with that conference. Our point is that it would be beneficial for these researchers to make greater use of the effort that has been spent for over 20 years now by other researchers who have truly cared about formally defining the concept of trust (e.g. [5])[2] and developing frameworks for reasoning about peer trustworthiness (e.g [30, 38, 39, 42, 43, 46]), along with longstanding interest in the related concept of reputation (e.g. [33]). This paper begins to examine what the relationship between these two groups within the AI community should be.

Our suggested opportunities for synergy between the communities are preliminary but fit well, we believe, with the kinds of desiderata for trust modeling research outlined in Sen's [35] seminal Blue Sky paper at AAMAS 2013. That paper made the community aware that there need to be greater considerations for trust modeling than simply creating effective solutions for evaluating trust. How to make USE of the results of trust modeling is also important, as is the challenge of how to ENGENDER trust. A small number of researchers have emerged of late, examining the design of algorithms for agents to engender trust [1, 41]. Certainly work in the field that promotes incentives to honesty also in some sense contributes to the establishment of trust (e.g. [16, 28, 45, 47]). These papers begin to shed some light on why trust modeling is performed within multiagent systems, and how intelligent agents can set their behaviour so that when their trust is modeled, the expectation is that they will indeed be considered to be trustworthy. But in order to examine whether the same techniques carry over for the more general concerns with trusted AI in vogue today, the questions to ask include: what are human users cautious about? what would lead them to have insufficient trust in their intelligent solutions? Perhaps more strongly, would users lose trust if the designers of their AI systems failed to conduct some kind of trust modeling, in contexts where there are multiple actors (agents)? If so, trust modeling could be viewed as a kind of necessary condition to enable trusted AI. There is another way for trust modeling solutions to be used. Researchers in trusted AI have differing solutions for how to address fairness or bias or transparency within their systems. Thus, solutions could be compared using metrics adopted by the trust modeling community, when validating their models. This would then provide an avenue for gauging the value of the trusted AI solution. Yielding better performance under this kind of evaluation would engender improved trust in the system from users.

In this paper, we first discuss the topic of trusted AI and the primary concerns of researchers who are trying to develop solutions for this problem. We then reflect on how particular trust modeling solutions can be examined as part of the solution. We conclude with a proposal for next steps.

[1] Some small steps in this direction emerged just this year with a proposed combination of topic areas in a recent ACM TOIT Special Issue Call for Papers on Trust and AI, following the 2018 AAMAS Trust workshop [6]

[2] The concept of trusted delegation here was especially prescient.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

2 CURRENT CONCERNS WITH TRUSTED AI

One of this paper's authors was on the program committee for AAAI in 2018, where Trustable and Explainable AI was a possible category of submission. We examined all abstracts submitted to the conference which selected this submission category (50 of them) and stepped back to try to characterize the kinds of concerns expressed and the kinds of approaches offered for addressing the problem. The primary reason for being concerned about trusted AI appeared to fall into one of four primary categories which we term **fairness**, **transparency**, **dangers** and **collaboration**.

Transparency was by far the central concern motivating the research to enable trusted AI. Of the 23 papers in this category, about half were specifically addressing explaining machine learning and of these about half were explicitly focusing on the case of obscurity of deep learning for human users. The other topics under concerns for trusted AI due to transparency included human interpretability in general, explanations of ontologies and of recommender systems, and the need for transparency when advice is given. One paper even discussed how transparency may in fact cause harm. Fairness was another central concern and, as perhaps expected, a number of papers delved into how to address hidden bias in machine learning solutions. Yet others focused on a need to have the reasoning and output of intelligent systems adhering to expected ethics and moral behaviour, as another element of fairness. Dangers is the label we attached to those papers most concerned with negative impacts that could be delivered, if intelligent systems were provided with undue autonomy or if various attacks could compromise the performance of our AI solutions. The challenges anticipated spanned such scenarios as spammers, adversaries and dishonest agents. This latter case is certainly quite close to the core desiderata of multiagent trust modeling research. Safe reinforcement learning and awareness of bugs in the solution were also issues raised, for a somewhat distinct perspective on danger being derived from AI solutions failing to be aware of possible inherent shortcomings. The final category was collaboration. Papers here were primarily focused on whether partnerships between humans and agents would be effective or not, due to differing social practices, human expectations and human perception of the intended role of the agents (among other concerns that were also examined).

If we step back from this effort to identify what researchers are trying to find solutions for, in order to promote "trusted AI", we learn the following. One entire thread to this endeavour is to ensure that harmful behaviour does not compromise the performance of the AI system (**danger**); detecting harmful behaviour is often as well a primary motivation for modeling trust in multiagent systems.

With both a consideration of **fairness** and a desire to reach human standards within **collaboration**, the concern seems to be that the decisions promoted by the AI systems may fail to measure up, against considerations of acceptable social performance; this is related to the thread of trust modeling focused on identifying peers who should not be accepted into the community of agents [18].

The final and perhaps most prominent of desiderata for trusted AI is to enable human users to not only receive advice from their intelligent agents but also to promote a certain comfort level in the directions which the humans are asked to follow; being able to obtain an **explanation** for the actions that are proposed is intended to engender trust, since without these clarifications, the decisions are coming from a black box. This is perhaps the most intriguing path to trust modeling research but can be understood as follows. If the "brain dump" of the agent's reasoning is available for observation, the agent is then attempting to dispel any lack of trust in its decisions. Is this not in essence as if (in a multiagent system environment) agents desire their trustworthiness to be tracked and the opinions in their past behaviour to be scrutinized, so that this process has the opportunity to detect shortcomings? The conclusion is that modeling trustworthiness is to be encouraged. After all, simply displaying the rationale is insufficient; accompanying this must be a measure of whether these explanations are reasonable. If we could imagine automating that process could this not be achieved by modeling trustworthiness, where falling short leaves one to conclude that the behaviour doesn't measure up to standards?

It is important to note that concerns about trusted AI are not entirely new. A valuable survey of research on goal reasoning and trusted autonomy is presented in [14]. This discussion underscores the importance of designing multiagent solutions in a way that human users can approve of the autonomous decision making. These researchers in fact promote what they refer to as inverse trust: having intelligent agents reflect on their own trustworthiness, so that behaviour can be altered in order to earn more trust from their human users. In 2002, a set of researchers [31] examined as well how to enable robust autonomous decision making explicitly so that the underlying planning and task execution can be trusted. Countermeasures against misinformation is mentioned as an important element. These papers point to an important thread between designing trusted intelligent agents and including an explicit step of trust modeling. At the same conference where Sen promoted his blue sky vision for the field of trust modeling, Kaminka's Blue Sky paper [17] proposed trust modeling as an important step forward for new directions with robotics research (i.e. that the design should include elements which convince the designers that these robots will be accepted by their human users). How telling then that AI in general has begun to embrace the cause of trustable AI.

3 TOWARDS A COMMON GROUND

We begin with the following observation. Many multiagent trust modeling solutions are evaluated by simulations where a standardly used benchmark is the one where trust isn't being modeled at all, and perhaps random selection of partners from peers is then evoked (e.g. [11]). These are often the least favourable curves in those graphs, emphasizing that gains can be incurred if trust is modeled carefully. In a similar fashion, one could perhaps expect that a user of an AI system may be unwilling to accept the decisions promoted by an intelligent system that had not bothered at all to imagine deception (or unfairness or obscurity). Including some kind of trust modeling pass could then be viewed as a necessary condition to enabling trusted AI. We proceed to shed some light on possible ways for trust modeling techniques to be introduced into the design of AI systems by examining two distinct contexts: computer vision and supervised machine learning for labelling ground truth. Before we do so, we include brief reflection on what constitutes human acceptance of intelligent systems.

3.1 Exploring Human Acceptance

A recent special issue of ACM Transactions on Intelligent Interactive Systems focused on trust and influence in intelligent human-machine interaction, and challenged researchers to examine human emotional attachment to their agent collaborators. A possible motivating paper was that of Yuksel et al. [44], who explored affective trust: whether humans focus on reliability or attractiveness of anthroporphised agents. Several papers by special issue co-editor Gratch [9, 10, 26] revealed: concern when agents adopted different values; an aversion to envy of agents (though humans seem to experience less guilt when dealing with agent partners than humans); and creating agents to represent others causes selfishness while creating one to represent us promotes fairness. This last conclusion echoes well the value of developing algorithms for establishing trust and for gauging whether our own systems can meet desired reliability standards.

A seminal socio-economic model of trust was developed by Mayer et al. [24]. This integrative model of organizational trust decomposes perceived trustworthiness into three components: **i)** ability (belief in competencies) **ii)** benevolence (belief in wanting good) **iii)** integrity (belief in adhering to acceptable principles). The same model suggests that perceived trustworthiness and the trustor's propensities contribute to trust, surfacing in the form of risk-taking actions. Risk-taking actions are modulated not only by trust, but also by perceived risks, and the outcomes of risk-taking actions can influence future perceptions of trustworthiness. Various multiagent system researchers (e.g., [2, 37]) have drawn inspiration from Mayer et al.'s integrative trust model to take signals of ability, benevolence and integrity into account to quantify the trustworthiness of agents in online social networks [37] or to guide decision making in collaborative environments [2]. We seek to highlight that such systems can be combined with the idea of "inverse" trust [14], i.e., leveraging multiagent system approaches to have agents reflect about their own trustworthiness. Such combined models would allow trusted AI to reflect about its own perceived trustworthiness in terms of socio-economic dimensions like ability, benevolence and integrity, based on user behaviour (such as frequency of risk-taking actions like delegation of decisions to the AI system).

3.2 A Case of Trust for Computer Vision

In the field of computer vision, the topic of trusted AI arises in the context of users' trust that the algorithms correctly infer information about the world from images. This is a matter of trust in the reliability of the systems, rather than trust in their intentions, and it is an important factor where mistakes by an intelligent system caused by inaccurate vision algorithms may cause harm. Typically, computer vision algorithms are used in the context of a larger system with multiple AI components, which further complicates the problem of establishing trust in the system as a whole and in the individual components of the system. In some cases the vision and control systems form a single integrated network; this approach has seen considerable success in learning to control video games[25].

Reinforcement learning is a common method for training control systems, and is also used in trust modelling by systems such as that of [40]. One possible approach to establishing trust in control systems is to use training or testing logs as initial data for the trust modeling system. In this scenario, reliability information (in the form of the trust model) would be available immediately upon deployment of the system. An important consideration is when to begin training the trust model. For systems that are trained once, prior to deployment, it would be natural to use data only from tests conducted on the fully trained system. For systems capable of continuous, online learning, however, the problem is more complex. A system which eventually "forgives" early mistakes could learn from the entire training history of the system, and could continue to be used to learn separate trust models for each individual agent while it continues to learn after it is deployed.

The trust model of [40] can be used in this context. In this model, the reputation of agent a as perceived by agent b (represented by $r^b(a)$) is updated with feedback as follows:

$$r^b(\hat{s}) \leftarrow \begin{cases} r^b(a) + \mu(1 - r^b(a)) & \text{positive feedback, } r^b(a) \geq 0 \\ r^b(a) + \mu(1 + r^b(a)) & \text{positive feedback, } r^b(a) < 0 \\ r^b(a) + \nu(1 + r^b(a)) & \text{negative feedback, } r^b(a) \geq 0 \\ r^b(a) + \nu(1 - r^b(a)) & \text{negative feedback, } r^b(a) < 0 \end{cases}$$

The parameters μ and ν represent separate learning rate parameters for positive and negative experiences. The value of $r^b(a)$ is constrained to lie on the interval $(-1, 1)$; provided that the initial value is in this interval and $\mu, \nu < 1$, the update mechanism will guarantee that this property will be maintained. It is possible under this system for an agent to "recover" from a reputation $r^b(a)$ arbitrarily close to -1. More specifically, for $r^b(a) = -1 + \epsilon$, the reputation will require $\left\lceil \frac{\log(\frac{1-\epsilon}{\epsilon})}{\log(\mu+1)} \right\rceil$ iterations of consistent positive feedback to reach $r^b(a) \geq 0$. The parameters μ and ν can also be adjusted during the learning process such that the reputation is allowed to change quickly early in the system's training process, when the system's performance is low but improving rapidly, but the reputation is only allowed to change slowly once the performance of the system has become stable. This is very similar to the use of decreasing learning rates in training AI systems.

3.3 Trust and Argumentation

Argumentation is an approach to reasoning focusing not only on the conclusions reached, but also on the data and the inference steps involved in inferring conclusions from the data. Argumentation is therefore connected to the idea of establishing trusted AI by providing explanations for the output produced by intelligent agents. One area of AI in which explainability seems to be a necessary component for the development of trusted AI systems is supervised learning. The goal here is to learn a function mapping from input data to correct output labels based on a set of training examples for which the correct output labels are known (i.e., ground truth). These training labels are often generated by human experts, and the question arises as to how the judgment of individual experts should be trusted if there are legitimate reasons for why equally qualified experts happen to disagree on the correct output label. The problem of trusted AI in supervised learning can therefore be in part translated to the problem of trusted ground truth and of explaining ambiguous cases in terms of how different experts could arrive at possibly conflicting conclusions. Computational methods

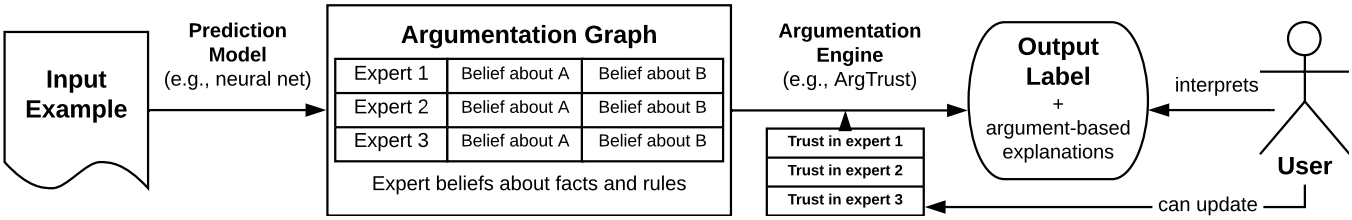

Figure 1: Use of trust-modulated argumentation engines to produce explanations in supervised learning regimes.

for aggregating noisy labels (e.g., [29]) fail to capture experts' arguments for diverging interpretations. Conversational techniques for resolving disagreements through an interactive exchange of arguments, however, were shown to increase label correctness [34].

Argumentation has a considerable history in multiagent systems and trust modeling research (e.g. [3, 23, 27, 36]). One of the models at the intersection of trust modeling and argumentation is ArgTrust [36]. ArgTrust is a trust-modulated argumentation engine serving as a decision-support tool for end users in complex scenarios where some information sources are more trustworthy than others. ArgTrust takes as input a pre-specified trust network (trustworthiness scores, derived by trust modeling), and a specification of each agent's beliefs about a set of facts and rules in the world. End users can then pose queries to the inference engine and receive output in the form of text reporting all arguments in support of their query along with each argument's level of trustworthiness. ArgTrust proposes a way of using existing trust values to modulate the influence of individual agents' arguments in order to help end users make informed judgments about complex scenarios.

We propose that future approaches to supervised learning may benefit from capturing the rationale of dissenting expert annotators in a format compatible with the belief specification syntax used by systems like ArgTrust. Assuming that, for each input training example in such a dataset, the possibly conflicting rationales from a panel of experts are known in the form of well structured belief specifications[3], these rationales could become prediction targets in and of themselves, as illustrated in Figure 1. In other words, supervised learning models would learn to predict these rationales for unknown data, as opposed to learning the aggregated output label directly. Trust-modulated inference engines like ArgTrust could then be used to produce argument-based explanations for what the AI system believes to be the correct output label. In cases where conflicting expert belief specifications are predicted by the model, the inference engine could offer an outline of conflicting arguments along with reasons why certain arguments may override others due to the underlying trust values assigned to individual experts. Finally, such systems could enable end users to adjust their own trust values for individual experts whose belief specifications are predicted by the model. This approach therefore not only sketches a possible solution to argumentation-based explainability in supervised learning regimes, but also opens up a way for users to communicate how much they trust certain types of reasoning, and thus could present an important milestone towards the development of trusted AI.

[3]Specifications about how the input example should be mapped to the output label, e.g. whether specific features are present in the input example (facts) and whether the presence of features should lead to certain conclusions about the output label (rules).

4 TOWARDS THE FUTURE

We first of all acknowledge that integrating trust modeling and AI ethics really seems to be the ideal overall goal. Human users will be disappointed if the AI system makes no effort to represent or reason about inherent social values that users would like to see reflected. At a workshop on AI and Ethics at IJCAI 2016, Rossi [32] posed several questions about how to address moral preferences in AI systems. Whether it is possible to integrate reasoning on both action preferences and moral preferences is one of the challenges we are urged to address within the field. Some recent efforts to advance specific reasoning frameworks for verifying morality have in fact emerged in the literature (e.g. [7, 8]) as promising first steps along this path. Trust modeling work on norms may also be a useful connected starting point to explore (e.g. [22]). We also fully acknowledge that it is impossible to dispel all fears simply by considering multiagent trust modeling. It is important for research on identifying vulnerabilities, attacks and collusive efforts should all continue, to identify failings in our models and to take steps forward. A useful starting point to this discussion is the work of [19] which chronicles some of the vulnerabilities of trust models. Kerr suggests that we can at least run our proposed algorithms through a testbed (e.g. [12, 20]), to begin to gauge our failings (and offers a method for addressing collusion by detecting lack of harm [21]). Related work on trying to gauge the relative benefits of different reputation mechanisms is offered in [13]: work like this may also provide crucial insights into how to calibrate trusted AI solutions. With various options for integrating trust modeling, efforts such as [4] which support interoperability and employ ontologies to facilitate explanation may also shed some light.

While we have only begun to sketch some ways in which trust modeling may be leveraged towards trusted AI, the primary take-home message of this paper is that we all have a responsibility, as multiagent system researchers, to enlighten our colleagues who are working on the thorny problem of trusted AI: the first step is **ours** to take. Tell them you know colleagues within your own subfield who have been devoting considerable effort to defining trust and reputation, to imagining alternate reasoning strategies for modeling trustworthiness most effectively, and to providing validations which yield metrics to gauge the performance of their solutions. Urge them to connect with these colleagues, to take advantage of our mutual interests. After all, if we cannot come up with truly effective solutions, then our poor human users will experience even greater disappointment, leading to even less trust in us and then no one will really be well served.

REFERENCES

[1] A. Aref and T. Tran. 2015. A trust establishment model in multi-agent systems. In *Proceedings of workshop on incentives and trust in e-communities (WIT-EC) at AAAI 2015.*

[2] Lucile Callebert, Domitile Lourdeaux, and Jean-Paul Barthès. 2016. A Trust-based Decision-making Approach Applied to Agents in Collaborative Environments. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence.* SCITEPRESS - Science and and Technology Publications, 287–295. https://doi.org/10.5220/0005825902870295

[3] Martin W. Caminada, Roman Kutlak, Nir Oren, and Wamberto Weber Vasconcelos. 2014. Scrutable Plan Enactment via Argumentation and Natural Language Generation. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '14).* International Foundation for Autonomous Agents and Multiagent Systems, 1625–1626.

[4] Sara Casare and Jaime Simão Sichman. 2005. Using a functional ontology of reputation to interoperate different agent reputation models. *Jour. Brazilian Comp Soc* 11, 2 (2005), 79–94.

[5] C. Castelfranchi and R. Falcone. 1998. Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In *Proceedings of ICMAS 1998.*

[6] Robin Cohen, Murat Sensoy, and Timothy J Norman (Eds.). 2018. *Proceedings of the 20th International Trust Workshop co-located with {AAMAS/IJCAI/ECAI/ICML} 2018, Stockholm, Sweden, July 14, 2018.* {CEUR} Workshop Proceedings, Vol. 2154. CEUR-WS.org.

[7] N. Cointe, G. Bonnet, and O. Boissier. 2016. Ethical judgment of agents' behaviors in mult-agent systems. In *Proceedings of AAMAS 2016.*

[8] V. Conitzer, W. Sinott-Armstrong, J. Borg, Y. Deng, and M. Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *Proceedings of AAAI 2017.*

[9] C. de Melo, S. Marsella, and J. Gratch. 2016. "Do As I Say, Not As I Do": Challenges in Delegating Decisions to Automated Agents. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16).* International Foundation for Autonomous Agents and Multiagent Systems, 949–956.

[10] C. de Melo, S. Marsella, and J. Gratch. 2016. People Do Not Feel Guilty About Exploiting Machines. *ACM Trans. Comput.-Hum. Interact.* 23, 2, Article 8 (May 2016), 17 pages. https://doi.org/10.1145/2890495

[11] H. Fang, J. Zhang, and N. Thalmann. 2014. Subjectivity Grouping: Learning from Users' Rating Behavior. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '14).* International Foundation for Autonomous Agents and Multiagent Systems, 1241–1248.

[12] K. Fullam, T. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, S. Barber, J. Rosenschein, L. Vercouter, and M. Voss. 2005. A Specification of the Agent Reputation and Trust (ART) Testbed: Experimentation and Competition for Trust in Agent Societies. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '05).* ACM, 512–518. https://doi.org/10.1145/1082473.1082551

[13] Christopher J Hazard and Munindar P Singh. 2013. Macau: A Basis for Evaluating Reputation Systems.. In *IJCAI.* 191–197.

[14] B. Johnson, A. Comman, B. Floyd, and D. Aha. 2017. *Goal reasoning and trusted autonomy.* Springer.

[15] A. Josang and R. Ismail. 2002. The beta reputation system. In *Proceedings of 15th Bled Electronic Commerce Conference.*

[16] R. Jurca and B. Faltings. 2003. An incentive compatible reputation mechanism. In *The Second International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2003, July 14-18, 2003, Melbourne, Victoria, Australia, Proceedings.* 1026–1027. https://doi.org/10.1145/860575.860778

[17] G. Kaminka. 2013. Curing robot autism: a challenge. In *Proceedings of AAMAS 2013.*

[18] G. Kastidou, K. Larson, and R. Cohen. 2009. Exchanging reputation information between communities: a payment-function approach. In *Proceedings of IJCAI 2009.*

[19] R. Kerr and R. Cohen. 2009. Smart cheaters do prosper: defeating trust and reputation systems. In *Proceedings of AAMAS 2009.*

[20] R. Kerr and R. Cohen. 2010. TREET: the Trust and Reputation Experimentation and Evaluation Testbed. *Electronic Commerce Research* 10, 3-4 (01 Dec 2010), 271–290. https://doi.org/10.1007/s10660-010-9056-y

[21] R. Kerr and R. Cohen. 2012. Detecting and identifying coalitions. In *Proceedings of AAMAS 2012.*

[22] M. Luck, L. Barakat, J. Keppens, S. Mahmoud, S. Miles, N. Oren, M. Shaw, and A. Taweel. 2009. Flexible behaviour regulation in agent based systems. In *Proceedings of CARE 2009.*

[23] Paul-Amaury Matt, Maxime Morge, and Francesca Toni. 2010. Combining Statistics and Arguments to Compute Trust. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1 (AAMAS '10).* International Foundation for Autonomous Agents and Multiagent Systems, 209–216. https://doi.org/10.1145/1838206.1838236

[24] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (7 1995), 709. https://doi.org/10.2307/258792

[25] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (26 02 2015), 529–533.

[26] Z. Nazari and J. Gratch. 2016. Predictive Models of Malicious Behavior in Human Negotiations. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16).* AAAI Press, 855–861.

[27] Simon Parsons, Elizabeth Sklar, Jordan Salvit, Holly Wall, and Zimi Li. 2013. ArgTrust: Decision Making with Information from Sources of Varying Trustworthiness. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '13).* International Foundation for Autonomous Agents and Multiagent Systems, 1395–1396.

[28] G. Radanovic and B. Faltings. 2015. Incentives for subjective evaluations with private beliefs. In *Proceedings of AAAI 2015.*

[29] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09.* ACM Press, 1–8. https://doi.org/10.1145/1553374.1553488

[30] K. Regan, P. Poupart, and R. Cohen. 2006. BLADE: Bayesian reputation modeling in e-marketplaces sensitive to subjetivity, deception and change. In *Proceedings of AAAI 2006.*

[31] J. Rose, M. Huhns, S. Roy, and W. Turkett. 2002. An agent architecture for long-term robustness. In *Proceedings of AAMAS 2002.*

[32] F. Rossi. 2016. Moral preferences. In *Ethics for AI workshop at IJCAI 2016.*

[33] J. Sabater and C. Sierra. 2005. Review on Computational Trust and Reputation Models. *Artif. Intell. Rev.* 24, 1 (Sept. 2005), 33–60. https://doi.org/10.1007/s10462-004-0041-5

[34] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. In *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'18).* https://doi.org/10.1145/3274423

[35] S. Sen. 2013. A comprehensive approach to trust management. In *Proceedings of AAMAS 2013.*

[36] Elizabeth I. Sklar, Simon Parsons, Zimi Li, Jordan Salvit, Senni Perumal, Holly Wall, and Jennifer Mangels. 2016. Evaluation of a trust-modulated argumentation-based interactive decision-making tool. *Autonomous Agents and Multi-Agent Systems* 30, 1 (1 2016), 136–173. https://doi.org/10.1007/s10458-015-9289-1

[37] Juan Pablo Soto, Aurora Vizcaíno, Javier Portillo-Rodríguez, and Mario Piattini. 2009. Why Should I Trust in a Virtual Community Member?. In *Groupware: Design, Implementation, and Use*, LuÃıs Carriço, Nelson Baloian, and Benjamim Fonseca (Eds.). Springer Berlin Heidelberg, 126–133.

[38] W. Teacy, A. Rogers M. Luck, and N. Jennings. 2012. An efficient and versatile approach to trust and reputation using hierarchical Bayesian modelling. *Artificial Intelligence* 193, Supplement C (2012), 149 – 185. https://doi.org/10.1016/j.artint.2012.09.001

[39] W. Teacy, J. Patel, N. Jennings, and M. Luck;. 2006. TRAVOS: Trust and Reputation in the Context of Inaccurate Information Sources. *Autonomous Agents and Multi-Agent Systems* 12, 2 (01 Mar 2006), 183–198. https://doi.org/10.1007/s10458-006-5952-x

[40] T. Tran and R. Cohen. 2004. Improving User Satisfaction in Agent-Based Electronic Marketplaces by Reputation Modelling and Adjustable Product Quality. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2 (AAMAS '04).* IEEE Computer Society, 828–835. https://doi.org/10.1109/AAMAS.2004.145

[41] T. Tran, R. Cohen, and E. Langlois. 2014. Establishing trust in multiagent environments: realizing the comprehensive trust dream. In *Proceedings of Trust workshop at AAMAS 2014.*

[42] D. Wang, T. Muller, J. Zhang, and Y. Liu. 2015. Quantifying robustness of trust systems against collusive unfair rating attacks using information theory. In *Proceedings of IJCAI 2015.*

[43] B. Yu and M. Singh. 2003. Detecting deception in reputation management. In *Proceedings of AAMAS 2003.*

[44] B. Yuksel, P. Collisson, and M. Czerwinski. 2017. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Trans. Internet Technol.* 17, 1, Article 2 (Jan. 2017), 20 pages. https://doi.org/10.1145/2998572

[45] J. Zhang and R. Cohen. 2007. Design of a mechanism for promoting honesty in e-marketplaces. In *Proceedings of AAAI 2007.*

[46] J. Zhang and R. Cohen. 2008. Evaluating the Trustworthiness of Advice About Seller Agents in e-Marketplaces: A Personalized Approach. *Electron. Commer. Rec. Appl.* 7, 3 (Nov. 2008), 330–340. https://doi.org/10.1016/j.elerap.2008.03.001

[47] J. Zhang, R. Cohen, and K. Larson. 2012. Combining trust modeling and mechanism design for promoting honesty in e-marketplaces. *Computational Intelligence* 28, 4 (2012), 549–578. https://doi.org/10.1111/j.1467-8640.2012.00428.x