Modelling Trust in Human-AI Interaction

Doctoral Consortium

Siddharth Mehrotra
Delft University of Technology
Delft, the Netherlands
s.mehrotra@tudelft.nl

ABSTRACT

Trust is an important element of any interaction, but especially when we are interacting with a piece of technology which does not think like we do. Therefore, AI systems need to understand how humans trust them, and what to do to promote appropriate trust. The aim of this research is to study trust through both a formal and social lens. We will be working on formal models of trust, but with a focus on the social nature of trust in order to represent how humans trust AI. We will then employ methods from human-computer interaction research to study if these models work in practice, and what would eventually be necessary for systems to elicit appropriate levels of trust from their users. The context of this research will be AI agents which interact with their users to offer personal support.

KEYWORDS

Trust; AI agents; Values; Value Similarity; Social Situations

ACM Reference Format:
Siddharth Mehrotra. 2021. Modelling Trust in Human-AI Interaction: Doctoral Consortium. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

Nowadays, many systems are being developed which have the potential to make a difference in people's lives, from health apps to robot companions. But to reach their potential, people need to have appropriate levels of trust in these systems. The surge of AI systems making decisions in real time makes trust a important factor to consider in AI. Therefore, it is important for an AI system to understand why and how we trust.

There are many unclear aspects of trust which are difficult to model with the tools available to us in literature on AI and Human-Computer Interaction (HCI). As AI systems gain more complexity and become ubiquitous in our life, it is important for them to blend in our society by eliciting appropriate trust from humans. We will study trust from two lenses namely social and formal lens. On the one hand, we need a formal perspective for AI itself to understand trust. Examples of such computational models are, for example [7] & [2] However, human trust is an inherently social concept, so if we want to understand human trust in AI systems, such as in [4], we also need to understand the social elements of trust. So additionally, we focus on social aspects that are important for any Human-AI interaction such as values, situational features etc.

Values represent what is important to people and hence our guide behavior. To understand trust between two entities an important aspect to consider is of Value Similarity (VS). In this project, we hypothesize that VS positively influence the trust in Human-AI agent interaction. We base our hypothesis on the premise of social science research where [12] show that people base their trust judgments on whether they feel that the agency shares similar values. We aim to formalize trust for AI systems to understand it, taking into account the social factors influencing trust, and how such formalism can be used when humans interact with the AI agent. For this, we try to understand trust from a more formal and computational point of view. We formulate an extensive literature review of former and current trust models in AI systems to scope the field. Based on the findings of the literature review, we will analyze how we could employ these trust models to promote appropriate trust in the agents, and what is still missing. To summarize, with the social lens we derive our knowledge of how to influence trust and, understanding trust models will allow us to embed appropriate trust in the agents. Combining both the lenses can provide us an overall bigger picture of how humans trust AI agents so that if we design interactions between humans and AI systems we can elicit an appropriate level of trust. Also, can comment and suggest human(s) to not over or under trust the AI system.

Researchers and designers of AI systems who wish to understand or promote user's (or agent's) trust can benefit from our research. We hope this project will encourage discussion within the community and offer new possibilities for modeling trust in the field of AI and multi-agent systems. The notion of formal and social perspective on trust will provide us an arena to look for future research directions and challenges identified by the scientific community.

2 RELATED WORK

We describe prior work in the area of modelling trust in AI agents. Trust in the AI agents has been explored modelled in context of decision making [11], examining/assessing user's trust [8], and improving the system performance [9]. Particularly, we are interested in those models that incorporate human values. Chhogyal and colleagues designed a formal value-based trust assessment model for Multi-agent systems [3]. In their work, they developed value-based trust assessment functions and showed how they lead to trust sequences. However, they did not validate their model with human participants leading to a lack of understanding of how humans trust AI agents.

We aim to understand the underlying importance of considering the intended purpose when developing a trust model. Much of

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

current work on modelling trust states that their motivation or intended purpose of their work was either related to making decisions for trust aware recommendations [6] or as a method for examining/assessing user trust [5]. We argue that to understand current trust models, their benefits and their shortcomings, we first need to understand: drivers, utilization, and the requirements of trust models. For this, we resonate with the suggestion of Anjomshoae et al. who provide a similar notion to understand agent explanations models in Explainable AI [1]. This notion of understanding purpose will guide both the research into current trust models, as well as the research on how to improve them.

In this project, we combine insights from formal and social perspective for modelling trust. We aim to understand humans trust in AI systems and to model those findings in AI agents. Combining these approaches allow us to use the knowledge about human's trust in an agent in interactions between them, as well as better understand the relationship between human trust and trustworthy AI. In a nutshell, we posit following research questions:

Table 1: Overall project research questions:

Social Lens
S1: What is the affect of the similarity of an agent & a human values on trust ? S2. What is the affect on trust if a human belief(s) an agent is taking into account their values?
Formal Lens
F1. What's the current state of the art in trust-modelling and within in the scope of computational constraints, what are the shortcomings in AI trust models? F2. Based on the findings of F1, how can we improve trust models to better enable AI agents to promote appropriate trust?
Application Areas
A1. How might we use knowledge about a human's trust in an agent in interactions between the two? A2. How can we guide the human(s) to appropriately trust the AI system(s), and how do we verify if this is the case?

3 PROPOSED APPROACH

We aim to design a framework which would allow a system to model and understand the trust of their users. To achieve this, we study trust through a combination of a social and formal lens.

First, we adopt a social lens to study the effect of the (non) similarity of human and agent values on trust in that agent *(S1, Table 1)*. We study this by designing an empirical study on how the VS influences human trust in a risk-taking task. Additionally, we are interested in how trust is influenced if the human believes the agent can take into account their values, irrespective of its own. Together, *S1* and *S2* should further illuminate the relationship between values and trust in human-AI interaction. Additionally, the results of both these studies will provide us with a social perspective in understanding how to build AI systems with understanding of human values, preferences and beliefs.

Second, we adopt a formal lens to study already developed models of trust in AI systems (*F1*). By doing a longitudinal systematic review study, we outline the drivers of trust models, their use, and the relevant literature's analysis. This helps us to understand how models make use of their results and different application scenarios they address. In addition, this will provide knowledge on the shortcomings and research directions in AI trust research. Here, we will study the primary concerns of researchers who are trying to develop solutions for trust in AI. Based on the concerns and computational constraints, we will reflect on how particular trust modelling solutions can be improved. Findings from *F1* will help us to analyze how trust models could be used to promote appropriate trust. Third, we put our attention in application areas for modelling trust. The results from combination of social and formal lens provide us a path to understand human values and belief(s) accompanied by how we can formally represent them in AI agents. Therefore, we utilize knowledge about formal models of human trust in AI agents for designing interactions between two (*A1*). These interactions can be for collaborative task or assisting AI agents. Finally, we will study how can we inform the users of the AI system to not over or under trust the AI systems based on the agent suggestions and notions from research on trustworthy AI.

4 PROGRESS

So far we have worked towards research questions S1 and F1. For S1 we have established preliminary theoretical foundations for designing agents with varying value similarity that use the approach from Schwartz [10]. In addition, we have conducted an user-study examining the affect of VS on trust. Our results indicate that:

- An overall positive affect of VS on trust was observed with correlation coefficient of 0.54, $p < 0.05$.
- There was a medium, positive correlation between benevolence and value similarity, which was statistically significant ($r = .47$, n = 436, $p < 0.05$). Similarly, for willingness, correlation was found to be positive ($r = .37$, n = 436, $p < 0.05$).
- VS is not the only essential prerequisite for trust, but it is most desirable.

For F1, we decompose the unstructured question into a set of four structured research questions:

- How have trust models been evolving over the years?
- Which are the main drivers demanding modeling of trust?
- What kinds of application scenarios have been addressed?
- What are the limitations of models as mentioned by authors?

We are in process to articulate a systematic review for answering above questions. We are currently reviewing 204 articles from last decade (2010-20) which include trust models for AI systems. Our preliminary results (56 articles) show that most trust models are either used to analyze a specific dataset (16) or to validate the model using a user study / an experiment based on the context (14). Additionally, the primary application scenarios highlighted in the surveyed articles include Gaming Applications (8), Recommender Systems (7), Social Networking (7), and Online Education (5). To extend the theoretical foundations of the developed approach to study trust and to further advance its applicability, we plan to explore our research questions in coming years.

ACKNOWLEDGMENTS

The author thanks Myrthe L. Tielman and Catholijn M. Jonker for their supervision and support.

REFERENCES

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

[2] Christiano Castelfranchi and Rino Falcone. 2010. *Trust Theory: A Socio-cognitive and Computational Model*. Vol. 18. John Wiley & Sons.

[3] Kinzang Chhogyal, Abhaya Nayak, Aditya Ghose, and Hoa K. Dam. 2019. A Value-based Trust Assessment Model for Multi-agent Systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 194–200. https://doi.org/10.24963/ijcai.2019/28

[4] Neta Ezer, Sylvain Bruni, Yang Cai, Sam J Hepenstal, Christopher A Miller, and Dylan D Schmorrow. 2019. Trust Engineering for Human-AI Teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 322–326.

[5] Siddharth Gulati, Sonia Sousa, and David Lamas. 2017. Modelling trust: An empirical assessment. In *IFIP Conference on Human-Computer Interaction*. Springer, 40–61.

[6] Arpit Merchant and Navjyoti Singh. 2017. Hybrid Trust-Aware Model for Personalized Top-N Recommendation. In *Proceedings of the Fourth ACM IKDD Conferences on Data Sciences*. 1–5.

[7] Lik Mui, Mojdeh Mohtashemi, and Ari Halberstadt. 2002. A computational model of trust and reputation. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*. IEEE, 2431–2439.

[8] Rui Ogawa, Sung Park, and Hiroyuki Umemuro. 2019. How Humans Develop Trust in Communication Robots: A Phased Model Based on Interpersonal Trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 606–607.

[9] Andisheh Partovi, Ingrid Zukerman, Kai Zhan, Nora Hamacher, and Jakob Hohwy. 2019. Relationship between Device Performance, Trust and User Behaviour in a Care-taking Scenario. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 61–69.

[10] Shalom H Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online readings in Psychology and Culture* 2, 1 (2012), 2307–0919.

[11] Maayan Shvo, Jakob Buhmann, and Mubbasir Kapadia. 2019. Towards Modeling the Interplay of Personality, Motivation, Emotion, and Mood in Social Agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019)*. 2195–2197.

[12] Michael Siegrist, George Cvetkovich, and Claudia Roth. 2000. Salient Value Similarity, Social Trust, and Risk/Benefit Perception. *Risk analysis* 20, 3 (2000), 353–362.