Artificial Trust for Decision-Making in Human-AI Teamwork

Steps and Challenges

Carolina Centeio Jorge [a] Catholijn M. Jonker [a,b] Myrthe L. Tielman [a]
[a] *Interactive Intelligence, Delft University of Technology, The Netherlands*
[b] *LIACS, Leiden University, The Netherlands*

Abstract. Human-AI teams count on both humans and artificial agents to work together collaboratively. In human-human teams, we use trust to make decisions. Similarly, our work explores how an AI can use trust (in human teammates) to make decisions, while ensuring the team's goal and mitigating risks for the humans involved. We present the several steps and challenges towards the development of an artificial-trust-based decision-making model.

Keywords. artificial trust, human trustworthiness, human-AI teamwork

1. Introduction

Our work focuses on how an artificially intelligent agent (also referred to as AI) can understand its human teammates in a wide range of teams, from search and rescue to healthcare, cooking, etc. In particular, we explore how an AI can use artificial trust to predict and understand whether a human will do a certain task and, if so, how well. With artificial trust, the agent might be able to make informed decisions which will improve team efficiency and reduce risks. Although trust in human-AI teams has counted on several contributions over the recent years, see e.g., [1,2,3,4,5,6], literature which includes notions of artificial trust in humans is limited. However, it includes important works such as [7,8,9,10,11,12,13].

Artificial trust (term described to the process of trust where the trustor is an AI [12]), similarly to natural trust (when the trustor is human), can be seen as a construct of aspects that are relevant when a human teammate is asked to collaborate on a certain task, such as ability, benevolence, integrity, capacity, preference, etc, see e.g. [14,15]. However, it is still an open question how relevant these aspects are in human-AI teams. Most of these constructs come from human-human studies and need to be tested in scenarios where humans and AIs are teammates. On the other hand, multi-agent systems (MAS) community has since long used beliefs of trust and trustworthiness for decision-making, see e.g., [16,17,18,19]. Our work aims at computing trust beliefs for the agent, as in MAS literature, while having a human as trustee (the entity being trusted). This requires us to reach inspiration from social sciences and run user studies to explore and validate our possibilities. In this short paper, we present the steps and challenges towards our artificial-trust-based decision-making model.

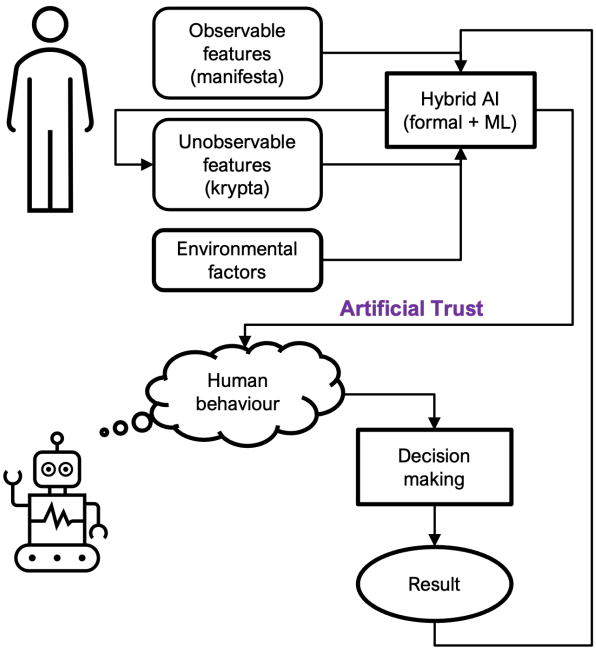

Figure 1. Overview of the model of artificial trust for decision-making. The AI can observe the human teammate and, by estimating their features, model artificial trust hybridly. With beliefs of artificial trust, the AI can estimate human behaviour and make a decision. Finally, it can update its model with the outcome of such decision.

2. Towards beliefs of Artificial Trust in Human Teammates

The goal of this work is to enable an AI teammate to make decisions taking its human teammates' trustworthiness into account. For this, the AI teammate needs to be able to form beliefs of artificial trust regarding humans. In [20], the authors present an artificial agent's trust in another as a construct of *Competence belief*, *Willingness belief* and *Dependence belief*. Competence belief deals with believing the trustee has the necessary abilities to perform a task, whereas willingness translates into believing the trustee will do a task *given the context*, independently of their abilities. Finally, dependence belief lies on the trustor's side, and it is crucial for the decision-making process as it tells how much the trustor depends on the trustee for the execution of a certain task. When we consider a team and joint goals, not only the trustor's (the one making the decision to trust) dependence belief is important but also the trustee's dependence on the trustor (e.g., the AI may choose to help a human because it believes the human is in a risky situation and depends on the AI for success). To the best of our knowledge, these beliefs are yet to be implemented and tested in human-AI interaction. It is challenging to do so as we do not know how to form such beliefs from humans, i.e., as said before, which aspects constitute human trustworthiness and how an artificial agent can perceive them. However, we aim at having such beliefs as a starting point for our artificial trust model which will be used to make decisions.

Figure 1 presents the overview of the goal model. The human presents a manifesta (i.e., set of behavioural cues) which represent a *krypta* (i.e., set of characteristics of the

human) [21,22]. The AI can use the *manifesta* to model beliefs of artificial trust in a hybrid way, i.e., with both data-driven and knowledge-based techniques, from the *manifesta* along with *environmental factors* (which give context). Beliefs of artificial trust can be, as mentioned before *competence*, *willingness* and *dependence* beliefs. With some of these beliefs, such as competence and willingness, the AI should be able to predict some human behaviour and then, keeping the context in mind (and the *(inter)dependencies*), make a decision. Finally, given the outcome of this action, the agent should update its model of artificial trust.

In summary, the steps towards developing a model of artificial trust for AI teammates' decision making are:

1. *Investigating krypta and manifesta of human trustworthiness towards an artificial teammate.* This included exploring which unobservable characteristics (the krypta) constitute human trustworthiness in human-AI teams and how they can be observed (the manifesta) through a user study in an online 2D grid-world supermarket environment. The task consisted of helping two artificial agents by collecting the necessary products in the supermarket (inspired in online supermarket orders that increased during pandemic). The agents would ask for different products and the human could choose which agent to help. After choosing to help one of the agents, the subject could either complete the task, lie about it, or give up on that task, and then go to the next one. We included metrics and conditions based on Mayer's ABI model [14], which proposes that someone's trustworthiness depends on their ability, benevolence and integrity. The results of the experiment are currently in the publishing process, but it is possible to find a preliminary report in [23].
2. *Updating artificial trust based on interaction given a context.* After studying which aspects may play a role in human's trustworthiness towards an artificial teammate, and grasp how it may be possible to perceive them, it is important to explore how this trust can be updated. Trust is dynamic[24], and an artificial teammate should be able to constantly update its trust values throughout interaction. Here, we can integrate existing models such as [11].
3. *Using artificial trust to make decisions.* The goal of this step is to propose a decision-making model for an AI agent which takes into account values of artificial trust for each sub-task of a joint goal and, at the same time, the interdependencies. We use principles of Coactive Design [25], including Interdependence Analysis.
4. *Evaluating artificial trust and decision-making models.* It is, in general, hard to evaluate artificial trust models, since we do not have ground truth. We work on defining metrics to compare different artificial trust models with baselines. The values of artificial trust can lead to decisions which may impact the environment and possibly the human teammate. It is then easier to compare the outcome of the decisions, rather than the trust values alone. The baselines include never-trusting models, always-trusting models or random models.

3. Challenges

Although the steps towards having an artificial-trust-based decision-making model are defined, they present several challenges. Overall, designing and evaluating experiments

presents methodological and theoretical issues, such as trying to explore complex real-world scenarios in controlled environments [26,27]. Our research is no exception, and we have learned along the way about the main obstacles.

The first and main challenge is the lack of ground truth. Human trustworthiness has no available ground truth, making it extremely hard to evaluate our artificial trust models. When we propose objective measures of trustworthiness, which may be manifestations of the krypta, we cannot prove that these are correct, or good, as there is nothing to compare them with. There is even the risk that, when we propose what human trustworthiness in a certain context may be, define the measures and then design the study, and all of this without other models to compare with or ground truth, we fall into a self-filling prophesy. We have tried to compare our measures with human's perception of their own trustworthiness, but this is far from the ground truth, as their perception can be very far from reality. This challenge affects enormously the first step. This being said, we develop metrics and baselines (mentioned in step 4) that help us evaluate our model by comparison.

Another challenge every time we start designing an experiment within this topic is the task design. The task is the platform in which we want to explore several aspects which usually come from theory, but it is usually hard to find one task which can accommodate all aspects nicely. Furthermore, to study human-AI teamwork, we need to ensure the human understands the collaborative aspect of the environment. Otherwise, it can happen that participants focus on solving the task rather than caring about the interaction (which, we believe, would not necessarily be the same in a real-world situation). For example, one may present different narratives to different participants, such as context or background story of the AI or their (human and AI) relationship, without integrating them in the task. This can lead to the participants focusing mainly on completing the task and ignoring the rest of the information they were given. For this reason, we have been investing on more visual representations of the teamwork as well as more levels of interdependencies, so that the participants understand that the AI is collaborative and that they can team up.

4. Conclusion

This paper presents an overview of the work that is being developed towards an artificial-trust-based decision-making model. The goal of such model is to enable an AI teammate to make informed decisions within a team context, taking into account its human teammate's trustworthiness and the context. The model should be updated throughout the interactions. We explain how we try to bridge multi-agent systems with social sciences and present the main steps that we take to develop such model. Finally, we conclude with some of the challenges we have encountered throughout our research, mainly related to the user studies and reflect on tentative mitigation strategies. Although there is still a long gap to fill, we believe this model will be of utmost importance for human-AI teams, improving their efficiency and mitigating possible risks.

References

[1] Anna-Sophie Ulfert and Eleni Georganta. A model of team trust in human-agent teams. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 171–176, New York, NY, USA, 2020. Association for Computing Machinery.

[2] Michael Lewis, Huao Li, and Katia Sycara. Deep learning, transparency, and trust in human robot teamwork. In *Trust in Human-Robot Interaction*, pages 321–352. Elsevier, 2020.

[3] Kristin E. Schaefer, Brandon S. Perelman, Gregory M. Gremillion, Amar R. Marathe, and Jason S. Metcalfe. A roadmap for developing team trust metrics for human-autonomy teams. In *Trust in Human-Robot Interaction*. Academic Press, 2021.

[4] Ewart J De Visser, Marieke, M M Peeters, Malte, F Jung, Spencer Kohn, Tyler, H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, 12:459–478, 2020.

[5] Beau G. Schelble, Caitlin Lancaster, Wen Duan, Rohit Mallick, Nathan J. McNeese, and Jeremy Lopez. The effect of AI teammate ethicality on trust outcomes and individual performance in human-ai teams. In Tung X. Bui, editor, *56th Hawaii International Conference on System Sciences, HICSS 2023, Maui, Hawaii, USA, January 3-6, 2023*, pages 322–331. ScholarSpace, 2023.

[6] Nathan J. McNeese, Mustafa Demir, Erin K. Chiou, and Nancy J. Cooke. Trust and team performance in human-autonomy teaming. *Int. J. Electron. Commer.*, 25(1):51–72, 2021.

[7] Alan R Wagner and Ronald C Arkin. Recognizing situations that demand trust. In *2011 RO-MAN*, pages 7–14. IEEE, 2011.

[8] Alan R. Wagner, Paul Robinette, and Ayanna Howard. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems*, 8, 11 2018.

[9] Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. Would a robot trust you? developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374, 4 2019.

[10] Vidullan Surendran and A. Wagner. Your robot is watching: Using surface cues to evaluate the trustworthiness of human actions. *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–8, 2019.

[11] Arsha Ali, Hebert Azevedo-Sa, Dawn M Tilbury, and Lionel P Robert. Heterogeneous human–robot task allocation based on artificial trust. *Scientific Reports*, 12(1):1–15, 2022.

[12] Hebert Azevedo-Sa, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. A unified bi-directional model for natural and artificial trust in human-robot collaboration. *IEEE Robotics Autom. Lett.*, 6(3):5913–5920, 2021.

[13] Samuele Vinanzi, Angelo Cangelosi, and Christian Goerick. The collaborative mind: intention reading and trust in human-robot interaction. *iScience*, 24(2):102130, 2021.

[14] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Source: The Academy of Management Review*, 20:709–734, 1995.

[15] Matthew Johnson and Jeffrey M Bradshaw. The role of interdependence in trust. In *Trust in Human-Robot Interaction*, pages 379–403. Elsevier, 2021.

[16] Jordi Sabater-Mir and Laurent Vercouter. Trust and reputation in multiagent systems. *Multiagent systems*, page 381, 2013.

[17] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. Computing confidence values: Does trust dynamics matter? In *Portuguese Conference on Artificial Intelligence*, pages 520–531. Springer, 2009.

[18] Cristiano Castelfranchi and Rino Falcone. Trust is much more than subjective probability: Mental components and sources of trust. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*. IEEE, 2000.

[19] Rino Falcone, Giovanni Pezzulo, and Cristiano Castelfranchi. A fuzzy approach to a belief-based trust computation. volume 2631, pages 73–86. Springer Verlag, 2003.

[20] Rino Falcone and Cristiano Castelfranchi. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *AAMAS*, pages 740–747. IEEE Computer Society, 2004.

[21] Rino Falcone, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. From manifesta to krypta: The relevance of categories for trusting others. *ACM Transactions on Intelligent Systems and Technology*, 4, 3 2013.

[22] Michael Bacharach and Diego Gambetta. *Trust as Type Detection*, pages 1–26. Springer Netherlands, Dordrecht, 2001.

[23] Carolina Centeio Jorge, Myrthe L. Tielman, and Catholijn M. Jonker. Assessing artificial trust in human-agent teams: a conceptual model. In Carlos Martinho, João Dias, Joana Campos, and Dirk Heylen, editors, *IVA '22: ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, September 6 - 9, 2022*, pages 24:1–24:3. ACM, 2022.

[24] Lixiao Huang, Nancy J Cooke, Robert S Gutzwiller, Spring Berman, Erin K Chiou, Mustafa Demir, and Wenlong Zhang. Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In *Trust in human-robot interaction*, pages 301–319. Elsevier, 2021.

[25] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69, 2014.

[26] Ann L Brown. Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The journal of the learning sciences*, 2(2):141–178, 1992.

[27] Allan Collins, Diana Joseph, and Katerine Bielaczyc. Design research: Theoretical and methodological issues. *Journal of the Learning Sciences*, 13(1):15 – 42, 2004. Cited by: 1090.