Artificial Trust as a Tool in Human-AI Teams

Centeio Jorge, Carolina; Tielman, Myrthe L.; Jonker, Catholijn M.

Publication date
2022

Document Version
Final published version

Published in
HRI '22

Citation (APA)
Centeio Jorge, C., Tielman, M. L., & Jonker, C. M. (2022). Artificial Trust as a Tool in Human-AI Teams. In *HRI '22: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 1155-1157). IEEE Press. https://dl.acm.org/doi/abs/10.5555/3523760.3523956

Artificial Trust as a Tool in Human-AI Teams

Carolina Centeio Jorge
Intelligent Systems
Delft University of Technology
Delft, The Netherlands
C.Jorge@tudelft.nl

Myrthe L. Tielman
Intelligent Systems
Delft University of Technology
Delft, The Netherlands
M.L.Tielman@tudelft.nl

Catholijn M. Jonker
Intelligent Systems[1] *& LIACS*[2]
[1]*Delft University of Technology*
[2]*Leiden University*
[1]Delft & [2]Leiden, The Netherlands
C.M.Jonker@tudelft.nl

***Abstract*—Mutual trust is considered a required coordinating mechanism for achieving effective teamwork in human teams. However, it is still a challenge to implement such mechanisms in teams composed by both humans and AI (human-AI teams), even though those are becoming increasingly prevalent. Agents in such teams should not only be trustworthy and promote appropriate trust from the humans, but also know when to trust a human teammate to perform a certain task. In this project, we study trust as a tool for artificial agents to achieve better team work. In particular, we want to build mental models of humans so that agents can understand human trustworthiness in the context of human-AI teamwork, taking into account factors such as human teammates', task's and environment's characteristics.**

***Index Terms*—trust, trustworthiness, human-robot teams, human-agent, human-AI, hybrid intelligence, HART**

I. Introduction

As technology advances, the understanding that artificial agents should collaborate with humans, instead of ultimately replacing them, becomes more corroborated and important. The idea that humans and Artificial Intelligence (AI) should work together comes from the understanding that both entities have a set of strengths and limitations, that can complement each other. Consequently, they can cover each other's weaker points, becoming stronger together. Hopefully, humans and AI can work as teammates, interdependently, helping each other. For this to become possible, it is important to explore mechanisms that contribute and allow effective teamwork and interdependence of *human-AI teams*. In particular, mutual trust is one key driver of effective teamwork in human teams [1]. In this project, we want to explore how we can use the notion of trust as a tool for prediction for artificial agents, when interacting with human teammates. If an agent would know how to estimate trustworthiness, it could know what to expect from a teammate regarding a task. More specifically, the agent would be able to decide when to rely on someone (we see reliance as the resulting behaviour of trust evaluation). We call artificial trust [2] to the artificial agent's belief in trustworthiness (in particular, human trustworthiness).

In a dyadic relation between two *cognitive agents* [3] (artificial or human), trust involves two parties, the *trustor* and the *trustee*, and an action (trusted by the trustor to the trustee) that affects a goal (of the trustor) [4]. Trust is dynamic and it is affected by several factors, from individual properties (both trustor's and trustee's characteristics) to environmental properties (such as challenges and limitations). Trust can be seen as the perceived trustworthiness, where trustworthiness is a property of the trustee. In several contexts, including human-AI teams, it is not only important that there is trust among teammates, but also that this trust is appropriate, i.e., that trust corresponds to actual trustworthiness (avoiding understrust and overtrust) [5]. Trustworthiness is a complex concept, and following the literature it can consist of a set of dimensions that range from the trustee's competence to its intentions [6]. Models in slightly different settings propose that trust depends on how one perceives another's 1) Ability, Benevolence and Integrity [7] (in human organizations), 2) Willingness, Competence and Dependence [4] (in multi-agent systems), and 3) Performance, Process and Purpose [8] (when the human is the trustor and an artificial agent is the trustee). The way trustworthiness is perceived can also depend on trustor's characteristics [7] and is usually influenced by external factors, which are contextual conditions determining the situation in which the task is executed [9], such as environmental configuration, emotional state, workload, etc. When studying trust in human-robot teams, we particularly need to take into account that the perception (from a human) of robot's trustworthiness may be influenced by its specific robotic characteristics, such as embodiment [10], which may also affect how the agent should trust the human. Moreover, trust is dynamic and in these teams we also need to consider how trust develops. Particularly, teammates may not possess the time to deepen their knowledge regarding other's trustworthiness dimensions, making use of swift trust [11], for example.

Trust has been vastly explored in several contexts in human teams (see e.g. [12]–[16]), and recently starts to being investigated also for human-AI teams (see e.g. [5], [17]–[19]). When diving into the perspective of an artificial agent's trust towards other entities, multi-agent systems community has addressed several important aspects, mostly when the other entity is also an artificial agent (see e.g. [20]–[25]). In particular, it is relevant for this work to take into account the models that distinguish internal qualities (krypta) of the agents from their observable signs (manifesta), to estimate trustworthiness, such as done in Falcone et al. [9]. Although there are several contributions in 1) how humans trust humans, 2) how agents can trust other agents, 3) how humans trust artificial agents (see e.g. [26], [27]), and 4) team trust (still recent but growing

AI*MAN of Delft University of Technology

in human-AI contexts), there is little research on how an artificial agent should trust its human teammates. However, there is some work in this direction, for instance on how an artificial agent can detect that a situation requires trust [28], [29] and also how an artificial agent can detect whether a human is being trustworthy, based on episodic memory [30] and social cues [31]. Also, Azevedo-Sa et al. [2] has recently proposed a model for trusting tasks in human-robot teams, making a clear distinction between natural trust (when the trustor is a human) and artificial trust (when the trustee is an artificial agent). The focus of the authors' model is capabilities, whereas in this current project we hypothesise that we should take more dimensions into account when determining trust.

Research on how an artificial agent should use the concepts of trust and trustworthiness in human-AI teams, as to understand better their human teammates' mental models, is still preliminary, and this research project aims at filling a part of that gap. The main research question of this PhD project is: "How can an artificial agent make use of trust in human teammates regarding tasks, in order to achieve the team's goals?". Although we aim at providing general frameworks for human-AI teams, our goal is to apply our research to robots, such as drones for search and rescue scenarios.

II. Proposed Approach

To answer our research question, we want to develop methods that will allow the artificial agent to both ask for help and initiate assistance when teaming up with humans, through reasoning about trust. Imagine there is a task (e.g. identifying an image that the agent captured). Which teammate would do it? How well? Would they need help? Which factors should the agent take into account? What should the agent do? To allow an agent to answer these questions, we will go from conceptually defining our model to later on tuning it from data. In particular, we want to use hybrid AI techniques, bridging formal (e.g. mental models, beliefs) and machine learning models (e.g. Machine Theory of Mind [32]), to decide on when and who to trust for a certain task. We want to apply these techniques to robots that can update the models based on interactions.

We start by defining human trustworthiness (i.e. what is a trustworthy human teammate?) and its dimensions (i.e. what influences human trustworthiness, e.g. integrity), in the context of human-AI teams, given a task. After knowing which dimensions are related to trustworthiness, we can form artificial trust (which can computationally unfold into other beliefs, such as competence and willingness belief [33]). For this, we want build machine learning models which, based on behaviour than hint to such dimensions, can estimate trustworthiness (e.g., learn integrity of a human teammate from observations and estimate whether a human teammate will perform a task). With such models, we can detect critical points (such as very low trustworthiness, meaning a human will likely be unreliable regarding a certain sub-task) in the process of a human teammate performing tasks. When detecting critical points, the artificial agent can act accordingly, adjusting its actions to the actions of its human teammate, ensuring as much as possible the achievement of the team goal (e.g., if the agent knows a human will not be able to perform a certain part of the process, then it can decide to help the human, ask some other human to do it, etc). Consequently, our agent should be provided with models that recognize when and who to ask for help as well as when its human teammates may need its help. This model should be used on robots and learn from mistakes of the interactions with human teammates, updating itself. Finally, we want to update this model to a real scenario, such as drones on urban search and rescue (USAR) or medical domains.

III. Progress

To define trustworthiness for this project, we started by investigating the general dynamics of trust in human-AI teams. In such teams, there are several dyadic trust relationships (human-human, human-agent, agent-human, agent-agent). More important than dyadic trust in teams, is *appropriate* dyadic trust, i.e. when one teammate's trust in another actually corresponds to the latter's trustworthiness. We looked at the specific beliefs in trust and trustworthiness that affect 1) an agent's appropriate trust in a human teammate and 2) a human's appropriate trust in an agent teammate, and how these beliefs are nested, in [34]. All of these trust beliefs contribute to the overall team trust, which we have been further investigating in a collaboration with psychology researchers and recently submitted a paper.

To form artificial trust (i.e. the artificial belief in human's trustworthiness, which usually unfolds into competence and willingness beliefs [33] when computing trust) regarding a human teammate, the agent needs to understand which human internal features (the krypta [9]) make a human trustworthy (i.e. ability, benevolence and integrity (ABI)), and how these can be observed through human behaviour (the manifesta [9]). To explore the relationships among these concepts, we designed, implemented, and ran a study with 54 human subjects in which people teamed up with artificial agents for collecting products from a supermarket, in a 2D grid online world. We have submitted a paper with the results, where we present a mental model of human trustworthiness, defending that an artificial agent can form artificial trust from behaviours that manifest ABI. Results also suggest that humans follow different strategies, depending on effort and reward, which also needs to be considered when assessing human trustworthiness for a certain task, in human-AI teams.

Moving forward, we hope to use the mental model of the first experiment, to learn how to interactively estimate human's trustworthiness in teamwork. For this, we will start by exploring machine learning models, such as Machine Theory of Mind [32] for this problem. We will also further explore which social signals may serve as relevant observable behaviour to estimate trustworthiness dimensions, so we can apply these models to human-robot teams.

REFERENCES

[1] E. Salas, D. E. Sims, and C. Burke, "Is there a "big five" in teamwork?," *Small Group Research*, vol. 36, pp. 555 – 599, 2005.

[2] H. Azevedo-Sa, X. J. Yang, L. P. Robert, and D. M. Tilbury, "A unified bi-directional model for natural and artificial trust in human-robot collaboration," *IEEE Robotics Autom. Lett.*, vol. 6, no. 3, pp. 5913–5920, 2021.

[3] C. Castelfranchi and R. 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.

[4] C. Castelfranchi and R. Falcone, *Trust & Self-Organising Socio-technical Systems*. Springer International Publishing, 2010.

[5] M. Lewis, H. Li, and K. Sycara, "Deep learning, transparency, and trust in human robot teamwork," in *Trust in Human-Robot Interaction*, pp. 321–352, Elsevier, 2020.

[6] N. Griffiths, "Task delegation using experience-based multi-dimensional trust," in *AAMAS '05*, 2005.

[7] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Source: The Academy of Management Review*, vol. 20, pp. 709–734, 1995.

[8] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of Human Factors and Ergonomics Society*, vol. 46, pp. 50 – 80, 2004.

[9] R. Falcone, M. Piunti, M. Venanzi, and C. Castelfranchi, "From manifesta to krypta: The relevance of categories for trusting others," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, 3 2013.

[10] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pp. 55–60, 2003.

[11] K. S. Haring, E. Phillips, E. H. Lazzara, D. Ullman, A. L. Baker, and J. R. Keebler, "Chapter 17 - applying the swift trust model to human-robot teaming," in *Trust in Human-Robot Interaction* (C. S. Nam and J. B. Lyons, eds.), pp. 407–427, Academic Press, 2021.

[12] C. Breuer, J. Hüffmeier, F. Hibben, and G. Hertel, "Trust in teams: A taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams," *Human Relations*, vol. 73, pp. 3 – 34, 2020.

[13] H. Huynh, C. E. Johnson, and H. S. Wehe, "Humble coaches and their influence on players and teams: The mediating role of affect-based (but not cognition-based) trust," *Psychological Reports*, vol. 123, pp. 1297 – 1315, 2019.

[14] A. M. Naber, S. C. Payne, and S. S. Webber, "The relative influence of trustor and trustee individual differences on peer assessments of trust," *Personality and Individual Differences*, vol. 128, pp. 62–68, 7 2018.

[15] A. Y. Lee, G. D. Bond, D. C. Russell, J. Tost, C. González, and P. S. Scarbrough, "Team perceived trustworthiness in a complex military peacekeeping simulation," *Military Psychology*, vol. 22, no. 3, pp. 237–261, 2010.

[16] B. D. Adams, S. Waldherr, and J. Sartori, "Trust in teams scale, trust in leaders scale: Manual for administration and analyses," 2008.

[17] A.-S. Ulfert and E. Georganta, "A model of team trust in human-agent teams," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, (New York, NY, USA), p. 171–176, Association for Computing Machinery, 2020.

[18] K. E. Schaefer, B. S. Perelman, G. M. Gremillion, A. R. Marathe, and J. S. Metcalfe, "A roadmap for developing team trust metrics for human-autonomy teams," in *Trust in Human-Robot Interaction*, Academic Press, 2021.

[19] E. J. D. Visser, Marieke, M. M. Peeters, Malte, F. Jung, S. Kohn, Tyler, H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human-robot teams," *International Journal of Social Robotics*, vol. 12, pp. 459–478, 2020.

[20] J. Urbano, A. P. Rocha, and E. Oliveira, "A socio-cognitive perspective of trust," in *Agreement Technologies*, pp. 419–429, Springer, 2013.

[21] J. Sabater-Mir and L. Vercouter, "Trust and reputation in multiagent systems," *Multiagent systems*, p. 381, 2013.

[22] A. Herzig, E. Lorini, J. F. Hübner, and L. Vercouter, "A logic of trust and reputation," *Logic Journal of the IGPL*, vol. 18, pp. 214–244, 12 2009.

[23] C. Burnett, T. J. Norman, and K. Sycara, "Stereotypical trust and bias in dynamic multiagent systems," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, 3 2013.

[24] K. Chhogyal, A. C. Nayak, A. Ghose, and K. H. Dam, "A value-based trust assessment model for multi-agent systems," *28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019.

[25] C. Cruciani, A. Moretti, and P. Pellizzari, "Dynamic patterns in similarity-based cooperation: An agent-based investigation," *Journal of Economic Interaction and Coordination*, vol. 12, no. 1, 2017.

[26] M. Winikoff, "Towards trusting autonomous systems," *Lecture Notes in Computer Science*, vol. 10738 LNAI, pp. 3–20, 2018.

[27] C. Nam, P. Walker, H. Li, M. Lewis, and K. Sycara, "Models of trust in human control of swarms with varied levels of autonomy," *IEEE Transactions on Human-Machine Systems*, vol. 50, pp. 194–204, 6 2020.

[28] A. R. Wagner and R. C. Arkin, "Recognizing situations that demand trust," in *2011 RO-MAN*, pp. 7–14, IEEE, 2011.

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

[30] S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi, "Would a robot trust you? developmental robotics model of trust and theory of mind," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, 4 2019.

[31] V. 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)*, pp. 1–8, 2019.

[32] N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine theory of mind," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (J. G. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 4215–4224, PMLR, 2018.

[33] R. Falcone and C. Castelfranchi, "Trust dynamics: How trust is influenced by direct experiences and by trust itself," in *AAMAS*, pp. 740–747, IEEE Computer Society, 2004.

[34] C. C. Jorge, S. Mehrotra, C. M. Jonker, and M. L. Tielman, "Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams," in *Proceedings of the International Workshop in Agent Societies*, 2021.