Assessing artificial trust in human-agent teams

A conceptual model

Centeio Jorge, C.; Tielman, M.L.; Jonker, C.M.

DOI
10.1145/3514197.3549696

Publication date
2022

Document Version
Final published version

Published in
IVA 2022 - Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents

Citation (APA)
Centeio Jorge, C., Tielman, M. L., & Jonker, C. M. (2022). Assessing artificial trust in human-agent teams: A conceptual model. In *IVA 2022 - Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents* (pp. 1-3). Article 24 (IVA 2022 - Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents). https://doi.org/10.1145/3514197.3549696

Assessing Artificial Trust in Human-Agent Teams

A Conceptual Model

Carolina Centeio Jorge
Delft University of Technology
Delft, Netherlands
c.jorge@tudelft.nl

Myrthe L. Tielman
Delft University of Technology
Delft, Netherlands
m.l.tielman@tudelft.nl

Catholijn M. Jonker
Delft Univ. of Tech. & Leiden Univ.
Delft & Leiden, Netherlands
c.m.jonker@tudelft.nl

ABSTRACT

As intelligent agents are becoming human's teammates, not only do humans need to trust intelligent agents, but an intelligent agent should also be able to form artificial trust, i.e. a belief regarding human's trustworthiness. We see artificial trust as the beliefs of competence and willingness, and we study which internal factors (krypta) of the human may play a role when assessing artificial trust. Furthermore, we investigate which observable measures (manifesta) an agent may take into account as cues for the human teammate's krypta. This paper proposes a conceptual model of artificial trust for a specific task during human-agent teamwork. Our model proposes observable measures related to human trustworthiness (ability, benevolence, integrity) and strategy (perceived cost and benefit) as predictors for willingness and competence, based on literature and a preliminary user study.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**.

KEYWORDS

trust, trustworthiness, human-agent teamwork, hybrid teams

ACM Reference Format:

Carolina Centeio Jorge, Myrthe L. Tielman, and Catholijn M. Jonker. 2022. Assessing Artificial Trust in Human-Agent Teams: A Conceptual Model. In *ACM International Conference on Intelligent Virtual Agents (IVA '22), September 6–9, 2022, Faro, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3514197.3549696

1 INTRODUCTION

Artificial agents are becoming more intelligent and able to execute relevant tasks for our daily lives, including in work environments, home assistance, battlefield and crisis response [15]. This holds for chat-based agents, but also for intelligent virtual agents and even robots. These tasks should complement human's sensorial and cognitive capabilities. For example, an intelligent agent can quickly process large quantities of data, but it may require a human to make ethical decisions. In these cases, humans and intelligent agents should learn to cooperate, coordinate and collaborate with people, forming human-agent teams. A key driver for achieving effective teamwork is *mutual trust* [24], i.e. teammates should trust each other. While literature usually considers the human's trust in Artificial Intelligence, we also need to look at how an artificially intelligent agent can trust a human teammate, i.e. when the agent is the trustor and the human the trustee (we call this artificial trust, as in [2]). This is important for agent decision-making [23, 28], e.g. to know who to rely on for a particular task (we see reliance as the resulting behaviour of trust evaluation), or who to help with a certain task, mitigating the risks and ensuring the team's goal. However, for an agent to be able to trust, it needs to form beliefs regarding a human's trustworthiness [11]. Literature so far explores how artificial agents can form beliefs regarding other artificial agent's trustworthiness, but not how they can form these beliefs regarding a human teammate. Thus an open question is: how can an artificial agent estimate trustworthiness of a human teammate from observations, given a specific task? This paper explores how these beliefs could be formed and proposes a conceptual model of artificial trust in human-agent teamwork context, based on theory and a preliminary user study.

IVA '22, September 6–9, 2022, Faro, Portugal

ACM ISBN 978-1-4503-9248-8/22/09.
https://doi.org/10.1145/3514197.3549696

2 CONCEPTUAL MODEL

Most research has focused on how humans trust agents, and not vice versa. Although there is some work in this direction, see e.g. [2, 27, 29], none of these works has tried to deconstruct trustworthiness, but rather looked at it as a simple metric. Instead, we hypothesise that we should take several dimensions into account when determining trustworthiness. Our proposed conceptual model of artificial trust in human-agent teams can be found in Figure 1. This model is based on existing concepts within the literature and it was explored in a preliminary user study where two artificial agents requested human to help them with simple tasks.

We start by exploring how artificial agents can form trust, from a computational perspective. We propose that artificial trust can be deconstructed in two beliefs regarding trustee's trustworthiness, i.e. *competence* belief, and *willingness* belief [8]. The competence belief reflects an evaluation of the trustee's abilities, meaning that the trustee can produce the expected results (i.e. can perform an action as expected). On the other hand, the willingness belief translates to whether the trustor believes the trustee will do the task (independently of competence belief). These beliefs may be affected by *external factors* like opportunities and interferences [8], which can be part of activity context and process [10, 14]. In general, factors such as propensity, preference, and perceived risk are often mentioned as elements that affect the trustor [10, 17]. We claim that some of these factors form human *strategy*. Strategy is mainly related to the goal, the task, and the consequence of taking the task. It

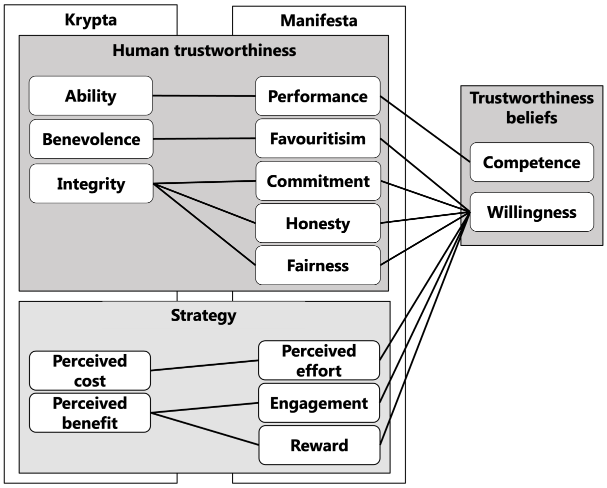

Figure 1: Conceptual model of human artificial trust in human-agent teams.

also plays a role in the decision-making of the trustee, determining whether a task will be performed, thereby affecting the trustee's trustworthiness.

2.1 Human trustworthiness

2.1.1 Krypta. Krypta is the set of internal features of an agent [3] that make them more or less trustworthy. We base our human krypta on the *ABI* model of trust [17], which has been widely used to study trustworthiness in organizational psychology. *ABI* says that human trustworthiness depends on their internal features of *ability* (group of skills and competencies of the trustee), *benevolence* (whether the trustee wants to do good to the trustor) and *integrity* (trustee's set of principles and the extent to which the trustor finds them acceptable). We associate ability with competence and benevolence and integrity with willingness by definition.

2.1.2 Manifesta. We looked into literature to find possible ways of observing ability, benevolence and integrity, so that we can have cues of the human krypta, and finally form beliefs regarding human's trustworthiness (i.e. willingness and competence). Ability in the context of human teams can be observed through *performance* (e.g. based on time or score of some kind), which reflects how much effort was put to do a task well, and also in how appropriately the tools (such as technology) were used [6]. Benevolence, however, can take its time to meaningfully develop [17], since it is connected to the relationship between the trustor and trustee. This makes the process of observing it in first-time interactions hard. A benevolent teammate accepts the requests of other agents, i.e. the one that voluntarily helps another agent, without this serving or harming its own goal [6, 16, 19]. In a team with several agents, we propose we can estimate benevolence through observing *favouritism* towards a certain agent, e.g. if a human is helping more one agent than others. Integrity, finally, is by definition related to values and moral principles. These principles can be such as honesty, truthfulness, sincerity, fairness, and ability to keep commitments (i.e. reliability, dependability) [18, 22]. As such, we can observe it through credible communications, a strong sense of justice, consistency of word and action, and availability [1, 6, 17]. Overall, in human-agent teamwork we propose that an agent can observe integrity through measures of *commitment* (e.g. how many times a human gave up their task), *honesty* (e.g. how many times the actions of a human mismatched their actions) and *fairness* (e.g. how many times a human broke commitment or lied to a specific agent when compared to how they treated the rest). It differs from benevolence, since it is related to general principles and values of the trustee, rather than trustee's attitude towards one particular trustor.

2.2 Strategy

2.2.1 Krypta. During our user study we could observe that participants might be following a *strategy*, i.e., to select their perceived advantageous alternatives from the beginning and persist on these lines of options [5]. Supported by literature [4, 5, 30] and the results of our study, we propose we should consider human's *cost-benefit evaluation*, i.e. participants choose whether or not the *perceived benefit* (e.g. score increment) is worth the *perceived effort* (e.g. time it takes to complete a task), and this affects their decision of which task to take or proceed with. In particular, the law of least effort plays a central role in decision-making, i.e. when presented to two tasks with equal rewards, one will choose the least effortful [26]. Perceived cost-benefit is affected by several factors, including goals, motivation, engagement, perceived risk, perceived effort, difficulty, time, utility, and overall cognitive characteristics [13, 21, 25]. Overall, what is effort and how a certain effort is rewarding to us depends on our characteristics (krypta) [9] (e.g. a person with good photographic memory may find it effortless to collect a new product).

2.2.2 Manifesta. How an agent can observe the perceived cost and perceived benefit is still an open question, as well as the relationship with the three trustworthiness dimensions. We do speculate, however, that the agent might be able to calculate perceived effort, engagement and reward, through observation of repeated human behaviour (see e.g. [7, 12, 20]).

3 CONCLUSION

In this paper, we presented a conceptual model of artificial trust, for human-agent teams, which is based on literature and a preliminary user study. We proposed a set of measures for observable human behaviours (manifesta), representing their krypta in order to relate them to artificial trust. This krypta is not only made of trustworthiness dimensions, i.e. ability, benevolence and integrity, but also strategy, i.e. perceived-cost and perceived-benefit which are used for human's cost-benefit analysis. We propose that with such measures it is possible for an artificial agent to build artificial trust, forming beliefs of competence and willingness. This conceptual model is a first step towards building an intelligent agent capable of building trust in human teammates and therefore capable of estimating human's behaviour and making informed decisions to achieve the team's goal.

ACKNOWLEDGMENTS

AI*MAN lab of TU Delft AI initiative.

REFERENCES

[1] Barbara D. Adams and R. Webb. 2002. Trust in Small Military Teams. *Command and Control Research Program* (2002).

[2] Hebert Azevedo-Sa, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. 2021. A Unified Bi-Directional Model for Natural and Artificial Trust in Human-Robot Collaboration. *IEEE Robotics Autom. Lett.* 6, 3 (2021), 5913–5920. https://doi.org/10.1109/LRA.2021.3088082

[3] Michael Bacharach and Diego Gambetta. 2001. *Trust as Type Detection.* Springer Netherlands, Dordrecht, 1–26. https://doi.org/10.1007/978-94-017-3614-5_1

[4] Matthew M. Botvinick and Zev B. Rosen. 2009. Anticipation of cognitive demand during decision-making. *Psychological Research PRPF* 73 (2009), 835–842.

[5] Matthias Brand, Kirsten Labudda, and Hans J Markowitsch. 2006. Neuropsychological correlates of decision-making in ambiguous and risky situations. *Neural Networks* 19, 8 (2006), 1266–1276.

[6] Christina Breuer, Joachim Hüffmeier, Frederike Hibben, and G. Hertel. 2020. Trust in teams: A taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams. *Human Relations* 73 (2020), 3 – 34.

[7] Mohammadreza Esfandiari, Senjuti Basu Roy, and Sihem Amer-Yahia. 2018. Explicit Preference Elicitation for Task Completion Time. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* 1233–1242.

[8] Rino Falcone and Cristiano Castelfranchi. 2004. Trust Dynamics: How Trust Is Influenced by Direct Experiences and by Trust Itself. In *AAMAS.* IEEE Computer Society, 740–747. https://doi.org/10.1109/AAMAS.2004.10084

[9] Michael Inzlicht, Amitai Shenhav, and Christopher Y. Olivola. 2018. The Effort Paradox: Effort Is Both Costly and Valued. *Trends in Cognitive Sciences* 22, 4 (2018), 337–349. https://doi.org/10.1016/j.tics.2018.01.007

[10] Matthew Johnson and Jeffrey M. Bradshaw. 2021. Chapter 16 - The role of interdependence in trust. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, 379–403. https://doi.org/10.1016/B978-0-12-819472-0.00016-2

[11] Carolina Centeio Jorge, Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams. In *Proceedings of the International Workshop in Agent Societies.*

[12] Monika Kaczorowska, Paweł Karczmarek, Małgorzata Plechawska-Wójcik, and Mikhail Tokovarov. 2021. On the Improvement of Eye Tracking-Based Cognitive Workload Estimation Using Aggregation Functions. *Sensors* 21, 13 (2021), 4542.

[13] Daniella Laureiro-Martínez, Stefano Brusoni, and Maurizio Zollo. 2010. The neuroscientific foundations of the exploration- exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics* 3, 2 (2010), 95.

[14] J. D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society* 46 (2004), 50 – 80.

[15] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. *The Role of Trust in Human-Robot Interaction.* Springer International Publishing. 135–159 pages. https://doi.org/10.1007/978-3-319-64816-3_8

[16] Michael Luck and Mark D'Inverno. 1996. Engagement and cooperation in motivated agent modelling. In *Distributed Artificial Intelligence Architecture and Modelling*, Chengqi Zhang and Dickson Lukose (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 70–84.

[17] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An Integrative Model of Organizational Trust. *Source: The Academy of Management Review* 20 (1995), 709–734. Issue 3.

[18] D. Harrison McKnight, Vivek Choudhury, and Charles J. Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Inf. Syst. Res.* 13, 3 (2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81

[19] Abdulla M. Mohamed and M. Huhns. 2001. Multiagent Benevolence as a Societal Norm.

[20] Shofiyati Nur Karimah, Teruhiko Unoki, and Shinobu Hasegawa. 2021. Implementation of Long Short-Term Memory (LSTM) Models for Engagement Estimation in Online Learning. In *2021 IEEE International Conference on Engineering, Technology Education (TALE).* 283–289. https://doi.org/10.1109/TALE52509.2021.9678909

[21] James Onken, Reid Hastie, and William Revelle. 1985. Individual differences in the use of simplification strategies in a complex decision-making task. *Journal of Experimental Psychology: Human Perception and Performance* 11, 1 (1985), 14.

[22] Michael E Palanski and Francis J Yammarino. 2007. Integrity and leadership:: clearing the conceptual confusion. *European Management Journal* 25, 3 (2007), 171–184.

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

[24] E. Salas, Dana E. Sims, and C. Burke. 2005. Is there a "Big Five" in Teamwork? *Small Group Research* 36 (2005), 555 – 599.

[25] Johannes Schiebener and Matthias Brand. 2015. Self-reported strategies in decisions under risk: role of feedback, reasoning abilities, executive functions, short-term-memory, and working memory. *Cognitive processing* 16, 4 (2015), 401–416.

[26] Richard L Solomon. 1948. The influence of work on behavior. *Psychological Bulletin* 45, 1 (1948).

[27] Vidullan Surendran and A. Wagner. 2019. Your Robot is Watching: Using Surface Cues to Evaluate the Trustworthiness of Human Actions. *RO-MAN* (2019), 1–8.

[28] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. 2013. A socio-cognitive perspective of trust. In *Agreement Technologies.* Springer, 419–429.

[29] Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. 2019. 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). Issue 1771. https://doi.org/10.1098/rstb.2018.0032

[30] M.E. Walton, S.W. Kennerley, D.M. Bannerman, P.E.M. Phillips, and M.F.S. Rushworth. 2006. Weighing up the benefits of work: Behavioral and neural analyses of effort-related decision making. *Neural Networks* 19, 8 (2006), 1302–1314. https://doi.org/10.1016/j.neunet.2006.03.005 Neurobiology of Decision Making.