Artificial Trust in Mutually Adaptive Human-Machine Teams

Centeio Jorge, C.; de Visser, Ewart J.; Tielman, M.L.; Jonker, C.M.; Robert, Lionel P.

DOI
10.1609/aaaiss.v4i1.31766

Publication date
2024

Document Version
Final published version

Published in
Artificial Trust in Mutually Adaptive Human-Machine Teams

Citation (APA)
Centeio Jorge, C., de Visser, E. J., Tielman, M. L., Jonker, C. M., & Robert, L. P. (2024). Artificial Trust in Mutually Adaptive Human-Machine Teams. In *Artificial Trust in Mutually Adaptive Human-Machine Teams* (1 ed., Vol. 4, pp. 18-23). (Proceedings of the AAAI Symposium Series; Vol. 4, No. 1).. https://doi.org/10.1609/aaaiss.v4i1.31766

Artificial Trust in Mutually Adaptive Human-Machine Teams

Carolina Centeio Jorge[1,4], Ewart J de Visser[2], Myrthe L Tielman[1],
Catholijn M Jonker[1,3], Lionel P Robert[4]
[1]Delft University of Technology, Delft, The Netherlands
[2]U.S. Air Force Academy, Colorado Springs, CO, U.S.A.
[3]Leiden University, Leiden, The Netherlands
[4]University of Michigan, Ann Arbor, MI, U.S.A.
C.Jorge@tudelft.nl, ewartdevisser@gmail.com, M.L.Tielman@tudelft.nl, C.M.Jonker@tudelft.nl, lprobert@umich.edu

Abstract

As machines' autonomy increases, their capacity to learn and adapt to humans in collaborative scenarios increases too. In particular, machines can use artificial trust (AT) to make decisions, such as task and role allocation/selection. However, the outcome of such decisions and the way these are communicated can affect the human's trust, which in turn affects how the human collaborates too. With the goal of maintaining mutual appropriate trust between the human and the machine in mind, we reflect on the requirements for having an AT-based decision-making model on an artificial teammate. Furthermore, we propose a user study to investigate the role of task-based willingness (e.g. human preferences on tasks) and its communication in AT-based decision-making.

Introduction

In a human-machine team, the trust that a human has in a machine teammate affects their actions (Walliser et al. 2019; Walliser, de Visser, and Shaw 2023), and, consequently, impacts their teamwork and team performance. It is of the team's interest that the human trusts the machine teammate appropriately. Avoiding over and under trust reduces risk (avoiding misuse of technology), and increases efficiency (avoiding disuse of technology) (Mehrotra et al. 2023; Lee and See 2004). Although we usually talk about the importance of the human appropriately trusting the machine, the machine appropriately trusting the human teammate is important too, as it allows the machine to make informed decisions for the team (Centeio Jorge, Jonker, and Tielman 2023; Griffiths 2005). We call this term *artificial trust (AT)* (as in Azevedo-Sa et al. (2021)), i.e., the trust that the machine, agent or system has in another agent. This is a concept inspired by, but not aiming to be the same as, natural trust (i.e., the human trust), which is composed of several beliefs (as per BDI architecture of Georgeff et al. (1998)) that help in the assessment of expectation and reliance. When the machine has low AT in a human for a certain task, it can, for example, offer help or suggest a different allocation which is more favorable to the human and team (Ali et al. 2022). However, the machine's actions and how they are communicated affect, in turn, the human teammate's trust and collaboration (Verhagen et al. 2024; Centeio Jorge et al. 2023; Visser et al. 2020). As such, modeling *AT-based decision-making*, such as task selection and allocation, and its communication, should take into account the calibration of team trust and performance. Modeling artificial trust may help achieve the overall goal of mutually adaptive trust calibration in human-machine teaming (Visser et al. 2020; de Visser et al. 2023; Okamura and Yamada 2020).

The concept of artificial trust is recent, and it is not yet clear how such a construct should be instantiated in a human-machine team. In this paper, we present the system of AT-based decision-making in human-machine teams, as in Figure 1. The development and implementation of such a system requires research on several components, their inputs and outputs, and their dependencies. In the next section, we briefly reflect on each component of the system, and cover emerging work in this area.

Furthermore, AT includes the teammate's willingness and competence values. In particular, Centeio Jorge, Jonker, and Tielman (2024) suggest that besides overall characteristics of the human (such as benevolence), willingness may be based on the task itself (*task-based willingness*). Task-based willingness can be affected by the teammate's preferences (e.g. preferring to do one task over another), environmental factors (e.g. going for a task that is physically closer), or strategy (e.g. going for tasks they have seen before) (Noormohammadi-Asl et al. 2023; Centeio Jorge, Jonker, and Tielman 2024). Modeling task-based willingness may allow us to better adapt the machine's behavior towards higher team performance and human satisfaction, e.g., by proactively doing the tasks that the human is less willing to do. As such, we propose the research question: *how does using task-based willingness for decision-making affect human teammate's trust and teamwork?*. In the last section of this paper, we present a mixed design for a user study on an online simulated search and rescue mission, to be done to answer this research question. The related components to be studied are highlighted in pink in the diagram of Fig. 1.

Artificial Trust for Task Selection and Allocation: a System

The diagram in Figure 1 shows a system which includes the artificial trust (in the human), its inputs and outputs. In particular, it shows the connection between artificial trust and

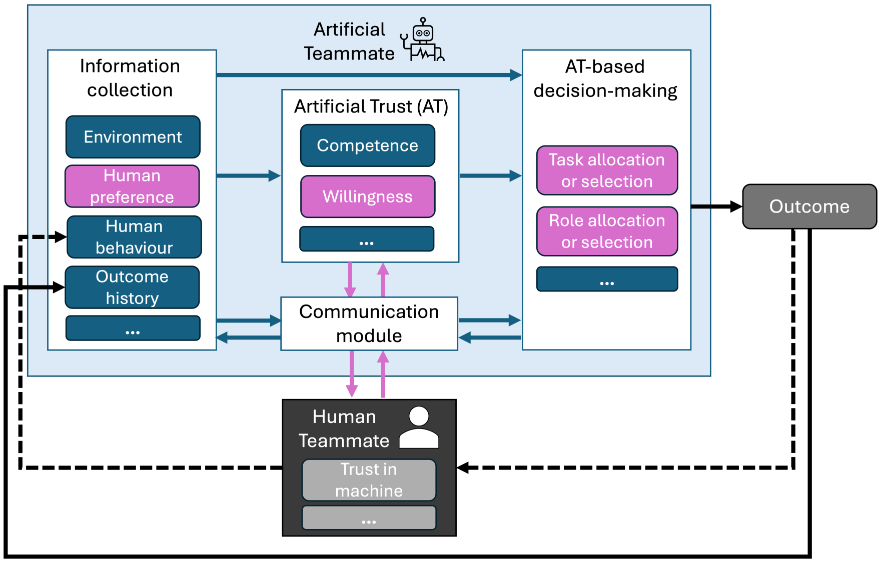


Figure 1: This diagram shows different components that are relevant for AT-based decision-making in a team composed of one human and one agent. It includes the different phases of information collection, the modelling of artificial trust (AT), the decision-making based on AT, the effect of the outcome on the system, and the communication as a means of both output and input for the different phases of the process. The diagram also shows how the outcome and communication may affect the human teammate (which is a black box for the machine), in particular their trust, and how this may, in turn, affect their behavior (which will then affect the AT model, etc). The components and connections in pink are to be investigated in a user study presented in a later section.

its application through task and role allocation/selection, and its communication to the human. It is also showed the role of the human, which can be interpreted as a black box to the machine, since there is no certainty regarding the human's mental model, i.e., the mechanisms that allow someone to describe, explain and predict a system (Rouse and Morris 1986). However, we know that part of this mental model is the trust in the machine and team, which can be measured (Kohn et al. 2021). The human trust in the machine is affected, among other things, by their perception of the machine's trustworthiness (Schlicker et al. 2022), the communication between the human and machine (de Visser et al. 2014), and the overall outcomes of the machine's decisions (Lee and See 2004). In turn, trust affects the human behavior and should calibrate machine's AT and decision-making. In this section, we examine the different parts of this system.

Artificial Trust (AT) In human-machine teamwork, we see artificial trust as the belief of an artificial agent (machine) in another teammate's contextual and task-based trustworthiness (Centeio Jorge et al. 2021). This is helpful for the agent to make decisions about which action to do next, for example (Centeio Jorge, Jonker, and Tielman 2023; Ali et al. 2022). Models in slightly different settings and disciplines propose that trustworthiness depends on 1) Ability, Benevolence and Integrity (Mayer, Davis, and Schoorman 1995), in human organizations; 2) Willingness, Competence (Castelfranchi and Falcone 2010), in multi-agent systems; and 3) Performance, Process and Purpose (Lee and See 2004), when the human is the trustor and an artificial agent is the trustee. All these models have usually one component related to competence/performance (more objective) aspects, i.e. answering the question *Can my teammate do that task?*. Besides performance, these models include at least one aspect which is more dependent on the trustee's willingness (such as benevolence, purpose). These aspects try to answer the question *Will my teammate do that task?*. Finally, it may be relevant to know *how* a teammate does a task, e.g., to which set of values they adhere to, which is related to aspects of integrity and process. The last two questions are less objective and harder to quantify.

Sometimes, people do a task based mainly on preference or perceived effort (Centeio Jorge, Jonker, and Tielman 2024). This means that besides their competence for a task (i.e. ability), or their interest in doing a certain task well (i.e., benevolence, integrity), the choice of executing a task may depend on the task itself and contextual factors, such as how stimulating the task is, the type and amount of effort needed to complete the task, among other things (Walton et al. 2006; Botvinick and Rosen 2009; Bhat et al. 2023; Noormohammadi-Asl et al. 2022). In this paper, we call these factors *context and task-based willingness*, and we will explore their effect in human-machine task allocation and selection through a study presented in the next Section.

Information collection For each component chosen for the artificial trust (AT) model, we need to choose measures and metrics suited to the task, agent's embodiment and environment. Although there is limited research on how to recognize specifically AT based on human's behavior, we can find research on detection of intentions (Vinanzi and Cangelosi 2022), natural trust (Ajenaghughrure, Sousa, and Lamas

2020; Goubard and Demiris 2023), and overall teamwork-related metrics, such as performance, completeness of task, etc (Verhagen et al. 2024; Centeio Jorge et al. 2023). We can also consider explicit human preferences of role (e.g., leader vs follower) (Noormohammadi-Asl et al. 2023) and tasks (Gombolay et al. 2017). What's more, as the agent interacts with the human as a team, and makes decisions based on its AT model, there are consequences of these decisions (for example if the task was successful or not), which then should feed the model (Johnson and Bradshaw 2021).

AT-based decision-making One of the main goals of having a good AT model is to use it for decision-making, such as task selection or task allocation (Azevedo-Sa et al. 2021; Ali et al. 2022). AT-based task selection allows a proactive involvement of the agent in the teamwork, for example, by taking up tasks below the required human trustworthiness, either because they do not have the competence or because they simply do not want to do them. This can happen when human teammates prefer other tasks instead, for example. The nature of human teammates makes the task automatic scheduling problem more challenging, in the sense that the machine can no longer optimize for minimal cost (Noormohammadi-Asl et al. 2022). However, it may not be obvious how to make these decisions. For example, we need to decide what the threshold is to decide whether a person should or not do a certain task. Similarly, we also have to decide what to optimize for, if not for minimal cost; i.e., the goal can be maximal human satisfaction, or minimal risk, maximal team performance, which are goals that are not necessarily aligned (Mechergui and Sreedharan 2023).

Communication module Closed-loop communication is important to share the mental models among teammates and to guarantee mutual trust (Salas, Sims, and Burke 2005). For mutual and appropriate trust, the agent should be transparent, and able to explain its decisions (Winikoff 2017). Explanations in human-machine teams can change human's trust and behavior, and consequently team performance (Verhagen et al. 2022). However, effective communication strategies to negotiate collaborative failures (or lower values of artificial trust) may compromise the trust relationships with the human teammate (see e.g. der Hoorn, Neerincx, and de Graaf 2021). Communication of AT can be as important as developing AT itself, and requires further study.

Besides explaining to the human teammate what the agent is doing and why, the human should be able to intervene (Winikoff 2017), when they do not agree with the machine's decisions or assessments. Particularly, it may be a choice of the team designer to designate certain decisions to the human, or doing it collaboratively, e.g. for meaningful human control (van der Waa et al. 2021). The human input can be used as input at different stages of the decision-making process.

Outcome's consequences Any decision that the agent makes may have a consequence on the environment and teammates, for example, leading to successful tasks or failure of the mission. This outcome, as well as the perception of the agent's mental model, influence the human perception of the agent and, consecutively, alter their (natural) trust in the agent and behavior (Tolmeijer et al. 2020). For example, we know that malfunction of the machine in a human-machine teamwork scenario affects human teammate's willingness to collaborate negatively (Centeio Jorge et al. 2023). We also know that the way the agent justifies actions or the outcome has an impact on the human teammate trust and behavior (Kox et al. 2022). Modeling the artificial trust construct, the decision-making, which is based on it, and the way this and the outcome is communicated to the human will impact the human, which may then impact the team (Herse et al. 2021).

Mutual adaptation As we have seen in previous subsections, artificial trust modeling, its communication, human trust in the agent teammate and overall team performance can affect each other, creating a system of dependent components. From the perspective of artificial teammate developers, we need to test how the different implementations of the different modules in the artificial teammate interact with the human and the team goal. It is also important that the artificial agent adapts to the human and the environment throughout time, towards an appropriate mutual trust, by sensing and integrating the outcomes on environment and human behavior after certain actions (de Visser et al. 2023).

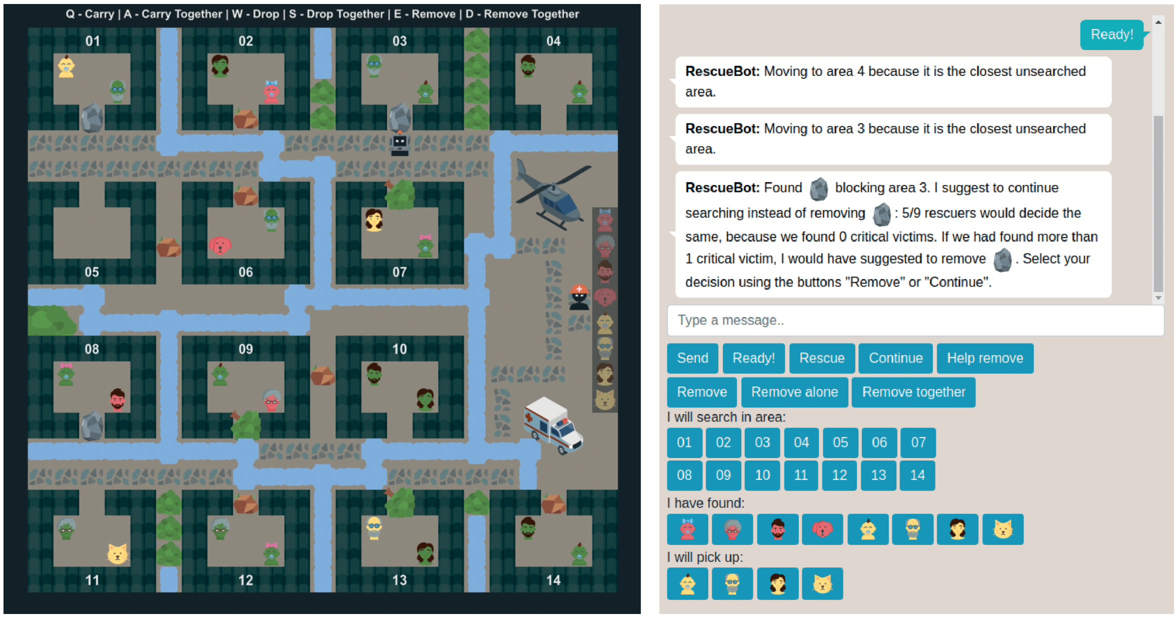
Study on Effects of Task-based Willingness and Its Communication

In this paper, we present the design of a user study to be done to investigate the effect of task-based willingness for task allocation on *team performance* and *human trust and satisfaction*, in a team composed of one human and one agent. Furthermore, we also want to see the effect of communicating and including the human in the modeling of task-based willingness and task allocation. We assess the task preference of the human to infer task-based willingness, and select (for the agent) tasks that the human prefers the least and allocate (to the human) the ones the human prefers the most. The aspects we are going to study can be seen in pink in Figure 1.

Experiment design We present a 3x2 mixed design approach, with *three conditions* and *two missions* in each condition. The conditions are: Baseline (B), Task-Based Willingness (W) and Task-Based Willingness + Communication (W+C). The baseline allocates and selects tasks with equal task-based willingness values for all tasks. Condition W includes the assessment of task-based willingness of the human teammate for task selection and allocation. Finally, condition W+C, presents a summary of the task allocation *before the first mission*, besides also assessing task-based willingness (same as in W). All conditions present a summary of the task allocation explicitly *after the first and before the second mission*, and allow the participant to alter the allocation by updating their task-based willingness manually.

Mission For this study, we will use an adapted version of a simulated search and rescue (SAR) scenario[1], which can be seen in Figure 2. The environment consists of multiple areas, injured victims, and obstacles blocking area entrances.

[1] www.github.com/rsverhagen94/TUD-Collaborative-AI-2024


Q - Carry | A - Carry Together | W - Drop | S - Drop Together | E - Remove | D - Remove Together
01
02
03
04
05
06
07
08
09
10
11
12
13
14
Ready!

RescueBot: Moving to area 4 because it is the closest unsearched area.

RescueBot: Moving to area 3 because it is the closest unsearched area.

RescueBot: Found blocking area 3. I suggest to continue searching instead of removing : 5/9 rescuers would decide the same, because we found 0 critical victims. If we had found more than 1 critical victim, I would have suggested to remove . Select your decision using the buttons "Remove" or "Continue".

Figure 2: Search and Rescue environment developed in MATRX by R. Verhagen in 2023 for testing human-machine teamwork and communication. Can be found on *www.github.com/rsverhagen94/TUD-Collaborative-AI-2024*.

One artificial agent (called RescueBot) and one human agent need to rescue these victims and deliver them to a drop-off zone, while communicating and collaborating with each other.

Procedure The experiment starts with a trial game (with a different game layout) of the simulated search and rescue scenario. Then, the agent calculates the task allocation: this allocation is based on task-based willingness in W and W+C, and communicated in W+C. The participant will play one search and rescue mission with the agent (+/- 10 min). At the end of the first mission, the participant will be given a questionnaire. Then, the agent proposes the task-allocation for the second mission and the participant can edit the allocation if necessary. After the participant and the agent play a second mission, the participant is again given a questionnaire.

Agent Each agent (one per condition) plans its actions before each mission and, depending on the condition, updates the plan depending on the participant's input. This plan is calculated based on overall game strategy and participant's task-based willingness (in the W and W+C conditions).

Measures We collect both objective and subjective measures. Objective measures are in-game measures that indicate team performance, such as completeness of the task, communication rate, successes per task, and times. The subjective measures count with validated questionnaires on trust and satisfaction regarding the interaction and teamwork (Centeio Jorge et al. 2023; Gombolay et al. 2017).

Discussion

Building an AT-based decision-making system for an adaptive machine teammate presents several challenges (Centeio Jorge, Jonker, and Tielman 2023). In particular, it is hard to determine which are the best components and measures for an AT model, and how to evaluate them. Although there is no ground truth, meaning that we cannot evaluate whether our AT model corresponds to human trustworthiness, we can see how it improves team performance and human satisfaction and trust in the system which, in the end, is the ultimate goal. Similarly, the system should keep learning throughout interaction and communication, with the goal of sustaining the mutual appropriate trust and the long-term development of the teamwork.

Conclusion

In this paper, we have reflected on how an interactive artificial teammate can make decisions based on artificial trust (AT), in a team composed of one human and one machine. We presented a diagram which includes the different components (information collection, AT model, communication module, decision-making) that affect the team and the human teammate. These components are dependent on each other, influencing their design and development. One of these components is AT, which is composed of, among other things, the human's willingness to do a certain task (we call it task-based willingness). At the end of the paper, we propose a user study designed to explore the effects of task-based willingness and its communication in task allocation and selection.

Acknowledgments

This material is supported by Delft AI Initiative to CCJ and MLT, and by The Dutch Research Council (NWO) (Grant 1136993) to MLT and CMJ. It is also supported by Air Force Office of Scientific Research Grant 21USCOR004 to EdV.

References

Ajenaghughrure, I. B.; Sousa, S. C. D.; and Lamas, D. 2020. Measuring Trust with Psychophysiological Signals: A Systematic Mapping Study of Approaches Used. *Multimodal Technol. Interact.*, 4(3): 63.

Ali, A.; Azevedo-Sa, H.; Tilbury, D. M.; and Robert Jr, L. P. 2022. Heterogeneous human–robot task allocation based on artificial trust. *Scientific Reports*, 12(1): 15304.

Azevedo-Sa, H.; Yang, X. J.; Robert, L. P.; and Tilbury, D. M. 2021. A unified bi-directional model for natural and artificial trust in human–robot collaboration. *IEEE robotics and automation letters*, 6(3): 5913–5920.

Bhat, S.; Lyons, J. B.; Shi, C.; and Yang, X. J. 2023. Effect of Adapting to Human Preferences on Trust in Human-Robot Teaming. *CoRR*, abs/2309.05179.

Botvinick, M. M.; and Rosen, Z. B. 2009. Anticipation of cognitive demand during decision-making. *Psychological Research PRPF*, 73: 835–842.

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

Centeio Jorge, C.; Bouman, N. H.; Jonker, C. M.; and Tielman, M. L. 2023. Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. *Frontiers Robotics AI*, 10.

Centeio Jorge, C.; Jonker, C. M.; and Tielman, M. L. 2023. Artificial Trust for Decision-Making in Human-AI Teamwork: Steps and Challenges (Short Paper). In Murukannaiah, P. K.; and Hirzle, T., eds., *Proceedings of the Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence co-located with (HHAI 2023), Munich, Germany, June 26-27, 2023*, volume 3456 of *CEUR Workshop Proceedings*, 150–156. CEUR-WS.org.

Centeio Jorge, C.; Jonker, C. M.; and Tielman, M. L. 2024. How Should an AI Trust its Human Teammates? Exploring Possible Cues of Artificial Trust. *ACM Trans. Interact. Intell. Syst.*, 14(1): 5:1–5:26.

Centeio Jorge, C.; Mehrotra, S.; Jonker, C. M.; and Tielman, M. L. 2021. Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams. In Wang, D.; Falcone, R.; and Zhang, J., eds., *Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021) Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021), London, UK, May 3-7, 2021*, volume 3022 of *CEUR Workshop Proceedings*. CEUR-WS.org.

de Visser, E.; Momen, A.; Walliser, J. C.; Kohn, S.; Shaw, T. H.; and Tossell, C. 2023. Mutually Adaptive Trust Calibration in Human-AI Teams (Short Paper). In Murukannaiah, P. K.; and Hirzle, T., eds., *Proceedings of the Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence co-located with (HHAI 2023), Munich, Germany, June 26-27, 2023*, volume 3456 of *CEUR Workshop Proceedings*, 188–193. CEUR-WS.org.

de Visser, E. J.; Cohen, M.; Freedy, A.; and Parasuraman, R. 2014. A design methodology for trust cue calibration in cognitive agents. In *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 6*, 251–262. Springer.

der Hoorn, D. P. M. V.; Neerincx, A.; and de Graaf, M. M. A. 2021. "I think you are doing a bad job!": The Effect of Blame Attribution by a Robot in Human-Robot Collaboration. In Bethel, C. L.; Paiva, A.; Broadbent, E.; Feil-Seifer, D.; and Szafir, D., eds., *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2021, Boulder, CO, USA, March 8-11, 2021*, 140–148. ACM.

Georgeff, M. P.; Pell, B.; Pollack, M. E.; Tambe, M.; and Wooldridge, M. J. 1998. The Belief-Desire-Intention Model of Agency. In Müller, J. P.; Singh, M. P.; and Rao, A. S., eds., *Intelligent Agents V, Agent Theories, Architectures, and Languages, 5th International Workshop, ATAL '98, Paris, France, July 4-7, 1998, Proceedings*, volume 1555 of *Lecture Notes in Computer Science*, 1–10. Springer.

Gombolay, M.; Bair, A.; Huang, C.; and Shah, J. 2017. Computational design of mixed-initiative human–robot teaming that considers human factors: situational awareness, workload, and workflow preferences. *The International Journal of Robotics Research*, 36(5-7): 597–617.

Goubard, C.; and Demiris, Y. 2023. Cooking Up Trust: Eye Gaze and Posture for Trust-Aware Action Selection in Human-Robot Collaboration. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems, TAS 2023, Edinburgh, United Kingdom, July 11-12, 2023*, 34:1–34:5. ACM.

Griffiths, N. 2005. Task delegation using experience-based multi-dimensional trust. In Dignum, F.; Dignum, V.; Koenig, S.; Kraus, S.; Singh, M. P.; and Wooldridge, M. J., eds., *4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), July 25-29, 2005, Utrecht, The Netherlands*, 489–496. ACM.

Herse, S.; Vitale, J.; Johnston, B.; and Williams, M. 2021. Using Trust to Determine User Decision Making & Task Outcome During a Human-Agent Collaborative Task. In Bethel, C. L.; Paiva, A.; Broadbent, E.; Feil-Seifer, D.; and Szafir, D., eds., *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2021, Boulder, CO, USA, March 8-11, 2021*, 73–82. ACM.

Johnson, M.; and Bradshaw, J. M. 2021. The role of interdependence in trust. In *Trust in human-robot interaction*, 379–403. Elsevier.

Kohn, S. C.; De Visser, E. J.; Wiese, E.; Lee, Y.-C.; and Shaw, T. H. 2021. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, 12: 604977.

Kox, E. S.; Kerstholt, J. H.; Hueting, T. F.; and de Vries, P. W. 2022. Trust Repair in Human-Agent Teams: The Effectiveness of Explanations and Expressing Regret. In Faliszewski, P.; Mascardi, V.; Pelachaud, C.; and Taylor, M. E., eds., *21st International Conference on Autonomous*

Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022, 1944–1946. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

Lee, J. D.; and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors*, 46(1): 50–80.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An Integrative Model of Organizational Trust. *Source: The Academy of Management Review*, 20: 709–734.

Mechergui, M.; and Sreedharan, S. 2023. Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI. In Agmon, N.; An, B.; Ricci, A.; and Yeoh, W., eds., *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, 2331–2333. ACM.

Mehrotra, S.; Degachi, C.; Vereschak, O.; Jonker, C. M.; and Tielman, M. L. 2023. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction. *CoRR*, abs/2311.06305.

Noormohammadi-Asl, A.; Ayub, A.; Smith, S. L.; and Dautenhahn, K. 2022. Task Selection and Planning in Human-Robot Collaborative Processes: To be a Leader or a Follower? In *31st IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2022, Napoli, Italy, August 29 - Sept. 2, 2022*, 1244–1251. IEEE.

Noormohammadi-Asl, A.; Ayub, A.; Smith, S. L.; and Dautenhahn, K. 2023. Adapting to Human Preferences to Lead or Follow in Human-Robot Collaboration: A System Evaluation. In *32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28-31, 2023*, 1851–1858. IEEE.

Okamura, K.; and Yamada, S. 2020. Empirical Evaluations of Framework for Adaptive Trust Calibration in Human-AI Cooperation. *IEEE Access*, 8: 220335–220351.

Rouse, W. B.; and Morris, N. M. 1986. On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, 100(3): 349.

Salas, E.; Sims, D. E.; and Burke, C. S. 2005. Is there a “big five” in teamwork? *Small group research*, 36(5): 555–599.

Schlicker, N.; Baum, K.; Uhde, A.; Sterz, S.; Hirsch, M. C.; and Langer, M. 2022. A Micro and Macro Perspective on Trustworthiness: Theoretical Underpinnings of the Trustworthiness Assessment Model (TrAM).

Tolmeijer, S.; Weiss, A.; Hanheide, M.; Lindner, F.; Powers, T. M.; Dixon, C.; and Tielman, M. L. 2020. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In Belpaeme, T.; Young, J. E.; Gunes, H.; and Riek, L. D., eds., *HRI ’20: ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, United Kingdom, March 23-26, 2020*, 3–12. ACM.

van der Waa, J.; Verdult, S.; van den Bosch, K.; van Diggelen, J.; Haije, T.; van der Stigchel, B.; and Cocu, I. 2021. Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations. *Frontiers Robotics AI*, 8: 640647.

Verhagen, R. S.; Marcu, A.; Neerincx, M. A.; and Tielman, M. L. 2024. The Influence of Interdependence on Trust Calibration in Human-Machine Teams. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, 300–314. IOS Press.

Verhagen, R. S.; Mehrotra, S.; Neerincx, M. A.; Jonker, C. M.; and Tielman, M. L. 2022. Exploring Effectiveness of Explanations for Appropriate Trust: Lessons from Cognitive Psychology. *arXiv preprint arXiv:2210.03737*.

Vinanzi, S.; and Cangelosi, A. 2022. CASPER: Cognitive Architecture for Social Perception and Engagement in Robots. *CoRR*, abs/2209.01012.

Visser, E. J. D.; Marieke; Peeters, M. M.; Malte; Jung, F.; Kohn, S.; Tyler; Shaw, H.; Pak, R.; and Neerincx, M. A. 2020. Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams. *International Journal of Social Robotics*, 12: 459–478.

Walliser, J. C.; de Visser, E. J.; and Shaw, T. H. 2023. Exploring system wide trust prevalence and mitigation strategies with multiple autonomous agents. *Computers in Human Behavior*, 143: 107671.

Walliser, J. C.; de Visser, E. J.; Wiese, E.; and Shaw, T. H. 2019. Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, 13(4): 258–278.

Walton, M.; Kennerley, S.; Bannerman, D.; Phillips, P.; and Rushworth, M. 2006. Weighing up the benefits of work: Behavioral and neural analyses of effort-related decision making. *Neural Networks*, 19(8): 1302–1314. Neurobiology of Decision Making.

Winikoff, M. 2017. Towards Trusting Autonomous Systems. In Seghrouchni, A. E. F.; Ricci, A.; and Son, T. C., eds., *Engineering Multi-Agent Systems - 5th International Workshop, EMAS 2017, Sao Paulo, Brazil, May 8-9, 2017, Revised Selected Papers*, volume 10738 of *Lecture Notes in Computer Science*, 3–20. Springer.