Mutually Adaptive Trust Calibration in Human-AI Teams

Ewart de Visser [a], Ali Momen[a], James Walliser[a], Spencer Kohn[b], Tyler Shaw[b], Chad Tossell[a]
[a] *U.S. Air Force Academy, Colorado Springs, CO, USA*
[b] *George Mason University, Fairfax, VA, USA*
ORCiD ID: Ewart de Visser https://orcid.org/0000-0001-9238-9081
Tyler Shaw https://orcid.org/0000-0002-4202-1120
Chad Tossell https://orcid.org/0000-0003-1662-9308

Abstract. We present the idea of mutually adaptive trust calibration in Human-AI Teams (HATs). Mutually adaptive trust calibration in HATs is established when both the human agent and the machine agents continuously adapt to one another, in terms of beliefs, attitudes and behaviors, to optimize trust and team performance. This goal requires new concepts, definitions, models, and measures. We highlight our past and recent studies that advance this important objective.

Keywords. Trust, Human-Autonomy Teams, Trust Calibration

1. Introduction

Much current work with regards to establishing “common ground” in relation to trust has been to either 1) enhance the human’s understanding of the machine or 2) enhance the machine’s understanding of the human [1,2]. The first approach, to create *human-readable* machines, encompasses efforts to enhance the transparency of the machine and the explainability of its actions to the human by enhancing interface design and training materials, amongst other methods. The second approach, to create *machine-readable* humans, attempts to measure the intent of a human to then provide adaptive feedback and adjustments to enhance overall understanding and collaboration (see **Fig. 1**). Yet, these methods overlook an important factor that is naturally evident in human-human communication: flexible real-time and fluent coordination. For example, real-time communication is naturally messy and requires quick adjustments, adaptation to errors, and recognition of mutual understanding. A more flexible and dynamic approach is needed that can account for errors, can adjust on the fly and can calibrate in real-time. The ultimate version of success with these approaches is to establish bi-directional and adaptive mutual trust calibration.

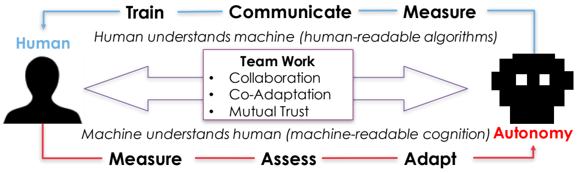
Train
Communicate
Measure
Human
Human understands machine (human-readable algorithms)
Team Work
Collaboration
Co-Adaptation
Mutual Trust
Machine understands human (machine-readable cognition)
Autonomy
Measure
Assess
Adapt

Figure 1. Approach for human-autonomy collaboration.

2. Mutually Adaptive Trust Calibration

2.1. Trust Calibration Definitions

Mutual trust is a fundamental property and predictor of good performing teams. We define trust as *"the continuous process of setting and updating a discrete interaction policy towards another agent in risky situations*." Some people trust AI a lot which can lead to over-trusting, leading to misuse and potentially disastrous outcomes. Others distrust AI which can cause under-trust, leading to disuse and unnecessary additional workload [3]. For instance, people tend to apply broad heuristics to trusting other agents such as the system-wide strategy, which occurs when one faulty system "bleeds over" negatively into the perception of similar, but well performing systems [4]. Ideally, people have *calibrated trust*, an optimized state when the perceived trust matches actual machine trustworthiness (see **Fig. 2**). Early on, we identified and studied teams composed of people and unmanned vehicles and found that trust increased with experience even though the robot made many mistakes [5]. Over-trust can be adjusted downwards by dampening expectations and under-trust can be adjusted upwards through trust repair strategies [6,7].

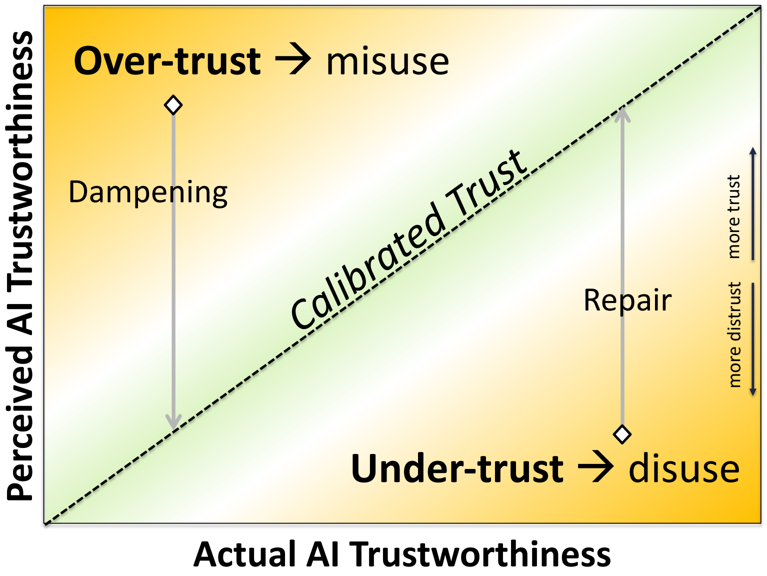

Figure 2. Trust calibration for over-trust and distrust with autonomy.

2.2. A Model for Mutually Adaptive Trust Calibration

Recently, we have established a model for mutually adaptive trust calibration in HATs [8]. This work presents a new model explaining the role and process of establishing *mutually adaptive longitudinal social trust calibration* throughout the life cycle of a HAT (see **Fig. 3**). The HAT Trust Model describes the development and role of trust calibration in HAT collaboration. HAT consists of four parts including 1) Relationship Equity, 2) Social Collaborative Processes, 3) Perceptions of Team Partner, and 4) Perceptions of Self. Central to our model is the idea of *relationship equity* which describes the cumulative result of the cost and benefit relationship acts that are exchanged during repeated collaborative experiences (including social and/or emotional interactions) between two actors. The middle part of the model describes the collaborative task performance between the teammates. Together, they perform a joint activity with the purpose of achieving a common goal. Collaboration is risky: actions may fail and circumstances may change. Therefore, the individual actors monitor the behavior and collaboration of themselves and their teammates. Based on their observations, actors aim to establish appropriate trust stances towards one another (A in B and B in A) to mitigate the potential risks involved in accomplishing the joint task. One part of the model includes *passive trust calibration process*: Based on team members' perceptions of one another, actors predict one another's trustworthiness. Taking into account their current formal work agreements and informal way of collaboration, they then (sub)consciously assess the risk involved in the collaboration as it currently is, and decide upon a trust stance towards one another. Another process of the model is the *active trust calibration process*: This process is based on an actor's awareness concerning their involvement in team trust calibration. This awareness enables both actors to engage in deliberate attempts to influence, aid, or hamper the trust calibration process.

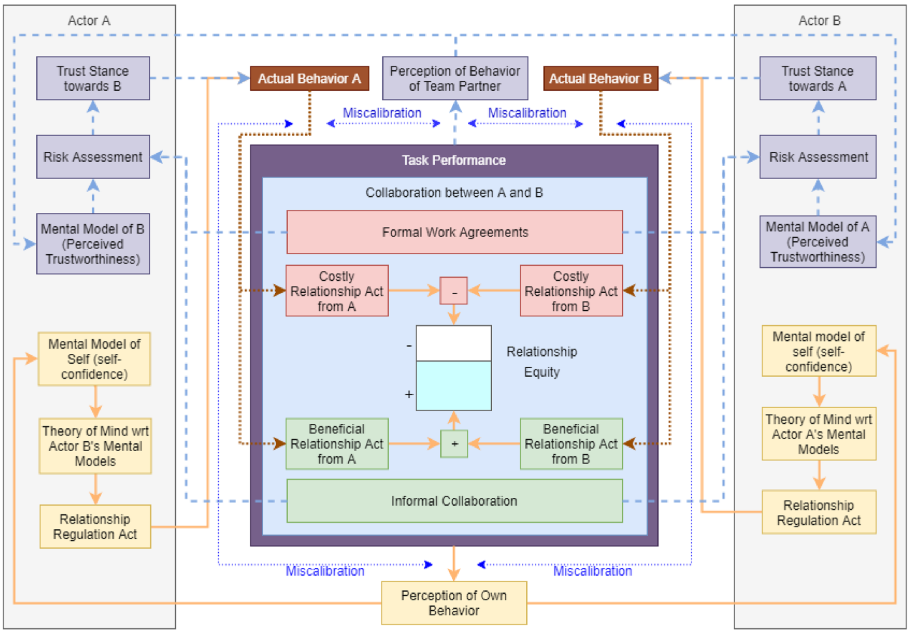
Actor A
Trust Stance towards B
Risk Assessment
Mental Model of B (Perceived Trustworthiness)
Mental Model of Self (self-confidence)
Theory of Mind wrt Actor B's Mental Models
Relationship Regulation Act
Actual Behavior A
Perception of Behavior of Team Partner
Actual Behavior B
Miscalibration
Miscalibration
Task Performance
Collaboration between A and B
Formal Work Agreements
Costly Relationship Act from A
-
Costly Relationship Act from B
-
Relationship Equity
+
Beneficial Relationship Act from A
+
Beneficial Relationship Act from B
Informal Collaboration
Miscalibration
Miscalibration
Perception of Own Behavior
Actor B
Trust Stance towards A
Risk Assessment
Mental Model of A (Perceived Trustworthiness)
Mental model of self (self-confidence)
Theory of Mind wrt Actor A's Mental Models
Relationship Regulation Act

Figure 3. The Human-AI Team (HAT) Trust Model.

3. Measurement of Trust in Human-Autonomy Teams

To validate models of mutually adaptive trust calibration, it is essential to develop and empirically assess trust building, development and repair in human-autonomy teams. Essential in this endeavor is good measures of trust and trustworthiness [9-12]. Recently, we have catalogued the state of the art of trust measurement [13] and we have mapped this to Mayer's original trust model [14]. We divide measures of trust into self-report, behavioral and physiological measures and show some examples of how these measures are used in the next sections.

3.1. Trusting the Moral Judgments of a GPT-Enabled Robot

We explored how a machine agent trained to respond to moral queries is perceived and trusted by human questioners [15]. Participants were tasked with querying the human-like agent with the goal of figuring out whether the agent, presented as a humanlike robot or a web client, was morally competent and could be trusted. Participants rated the moral competence and perceived morality of both agents as high yet found it lacking because it could not provide justifications for its moral judgments. While both agents were also rated highly on trustworthiness, participants had little intention to rely on such an agent in the future. This work presents an important evaluation of a morally competent algorithm integrated with a human-like platform that could advance how moral robot advisors are trusted in the future [16-18].

3.2. System-Wide Trust Effects in Human-Autonomy Teams

To demonstrate the effectiveness of interventions on trust calibration we provide an example of a recent UAV supervisory control study that countered a bias known as the *system-wide trust effect* [4]. This bias is the tendency for operators to apply trust broadly rather than exhibiting specific trust in each component of the system when one of these systems is faulty. Our study assessed the effectiveness of two trust calibration interventions to counter this bias including transparency feedback and scenario-based training. Results showed that by providing both system transparency feedback and training resulted in the most optimal verification rates and response times (see **Fig. 4**). This finding affects how HATs should be designed and trained [19-20].

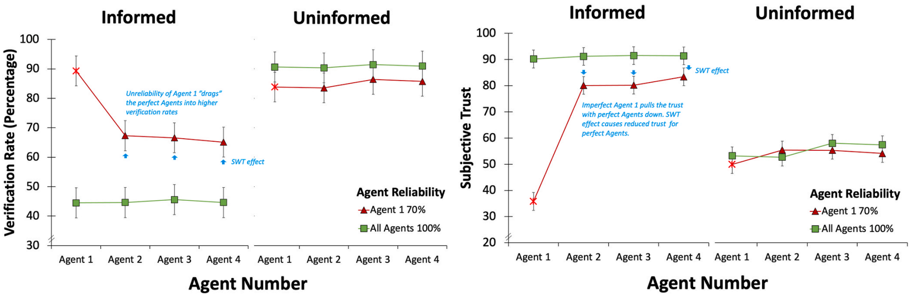

Figure. 4. Verification rates (left) and subjective trust (right) monitoring automation recommendations in a UAV supervisory setting. The Faulty UAV1 is indicated by a red asterisk. By providing transparency feedback (informed condition), participants were more calibrated but showed the system-wide trust effect.

3.3. Neural Correlates of Automated Agents

We also investigated the neural underpinnings and mechanisms of trust in automated agents [21]. We used two event-related potentials measured by electroencephalography as an indicator of trustworthiness (see **Fig. 5**). We demonstrated that this marker could distinguish between high and low reliability.

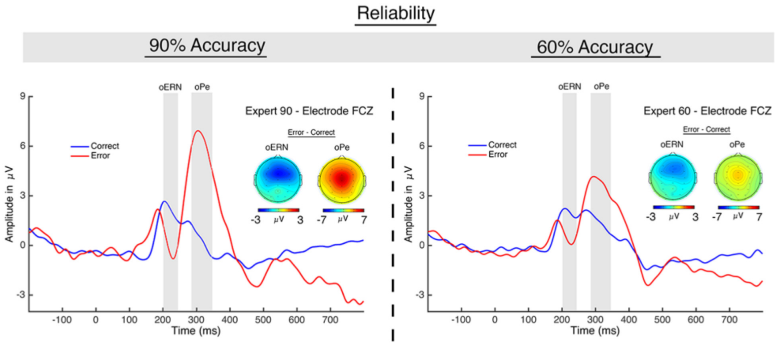

Fig. 5. Observed error related negativity (oERN) and observed positivity (oPE) as indicators of trustworthiness of an automated agent.

4. Planned Future Work

New efforts are under way to realize the vision of mutually adaptive trust calibrating HATs. New developments in generative AI will enable novel types of communication that can allow us to improve human-AI teaming, which is not as good as human-human teaming [22]. We are also investigating how trust calibration efforts trade-off with situation awareness and workload to improve teaming performance. Other efforts are focusing on quantifying the relationship equity construct that some have found useful for the prediction of cross-task and long-term AI teaming [23]. Furthermore, recent reviews have found trust repair, dampening and explanation efforts are not always effective and determined there is a need to develop better, predictive models to enhance such interventions [24, 25]. Lastly, we are examining how AI teammates can *align* with human decision-makers in terms of values and decision styles, which focuses on the integrity (process) and benevolence (purpose/intent) aspects of trustworthiness (as opposed to the ability / performance dimension of trustworthiness). Combined, these efforts may help to further advance and improve trust calibration in human-AI teams.

References

[1] Hancock, P.A., Billings, D.R., Schaefer, K. E., Chen, J. Y., de Visser, E. J., Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517-527.

[2] Juvina, I., Collins, M. G., Larue, O., Kennedy, W. G., Visser, E. D., & Melo, C. D. (2019). Toward a unified theory of learned trust in interpersonal and human-machine interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *9*(4), 1-33.

[3] de Visser, E.J., Marvin Cohen, Amos Freedy & Raja Parasuraman (2014). A design methodology for trust cue calibration in cognitive agents. *In Proceedings of the 16th International Conference on Human-Computer Interaction*. Crete, Greece.

[4] Walliser, J. C., de Visser, E. J., & Shaw, T. H. (2023). Exploring system wide trust prevalence and mitigation strategies with multiple autonomous agents. *Computers in Human Behavior*, *143*, 107671.

[5] de Visser, E.J., R. Parasuraman, A. Freedy, E. Freedy, and G. Weltman, "A Comprehensive Methodology for Assessing Human-Robot Team Performance for Use in Training and Simulation," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 50, no. 25, pp. 2639–2643, Oct. 2006.

[6] de Visser, E.J., Pak. R. & Shaw, T.H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human-machine interaction. *Ergonomics*, 1-19.

[7] Textor, C., Zhang, R., Lopez, J., Schelble, B., McNeese, N., Freeman, G., Pak, R., de Visser, E. J. & Tossell, C. (2022). Exploring the Relationship Between Ethics and Trust in Human-AI Teaming: A Mixed Methods Approach. Submitted to the *Journal of Cognitive Engineering and Decision Making*.

[8] de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, *12*(2), 459-478.

[9] Freedy, A., de Visser, E. J., & Weltman, G., Coeyman, N.(2007). Measurement of trust in human-robot collaboration. In *Proceedings of International Symposium on Collaborative Technologies and Systems*, Orlando, FL.

[10] Tenhundfeld, N. L., de Visser, E. J., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2020). Trust and distrust of automated parking in a Tesla Model X. *Human factors*, 001872081986541.

[11] Tenhundfeld, N. L., de Visser, E. J., Haring, K. S., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2019). Calibrating Trust in Automation Through Familiarity With the Autoparking Feature of a Tesla Model X. *Journal of Cognitive Engineering and Decision Making*, 1555343419869083.

[12] Tenhundfeld, N., Demir, M., & de Visser, E. (2022). Assessment of Trust in Automation in the "Real World": Requirements for New Trust in Automation Measurement Techniques for Use by Practitioners. *Journal of Cognitive Engineering and Decision Making*, *16*(2), 101-118.

[13] Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in psychology*, *12:* 604977.

[14] Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

[15] Momen, A., de Visser, E., Cooley, K., Walliser, J. & Tossell, C. (2023). Trusting a Robot With Moral Questions: Perceptions of Moral Competence and Humanlikeness in a GPT-3 Enabled Moral AI. In Proceedings of the *Hawaii International 56th Conference on Systems Sciences.*

[16] de Visser, E.J., Monfort, S.S., Goodyear, K., Lu, L., O'Hara, M., Lee, M.R., Parasuraman, R., Krueger, F. (2017). A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance and team performance with automated agents. *Human Factors,* 59, 116-133.

[17] de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331-349.

[18] de Visser, E., Krueger, F., McKnight, P., Steve Scheid, Melissa Smith, Stephanie Chalk & Parasuraman, R. (2012). The world is not enough: Trust in cognitive agents. In *Proceedings of the the 56th Annual Meeting of the Human Factors and Ergonomics Society,* Boston, MA.

[19] de Visser, E.J., & Parasuraman, R. (2011) Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making, 5*(2), 209-231.

[20] Coovert, M. D., Arbogast, M. S., & de Visser, E. J. (2020). The cognitive Wingman: considerations for trust, humanness, and ethics when developing and applying AI systems. In *Fields of Practice and Applied Solutions within Distributed Team Cognition* (pp. 191-217). CRC Press.

[21] de Visser, E.J., Beatty, P., Estepp, J.R., Kohn, S., Abubshait, A., Fedota, J.R. & McDonald, C.G. (2018). Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in Neuroscience,* 12, 309.

[22] Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team Structure and Team Building Improve Human–Machine Teaming With Autonomous Agents. *Journal of Cognitive Engineering and Decision Making*, 1555343419867563.

[23] Sharp, W. H., Jackson, K. M., & Shaw, T. H. (2023). The frequency of positive and negative interactions influences relationship equity and trust in automation. *Applied Ergonomics*, *108*, 103961

[24] Esterwood, C., & Robert, L. P. (2022, August). A literature review of trust repair in hri. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1641-1646). IEEE

[25] Pak, R., & Rovira, E. (2023). A Theoretical Model to Explain Mixed Effects of Trust Repair Strategies in Autonomous Systems