Calibrating AI Trust in Complementary Human-AI Collaboration

Hanjiang Hu, Yifan Sun, Changliu Liu

***Abstract*— Human-AI collaboration is a powerful paradigm in decision-making systems, where humans and AI contribute different strengths with clear complementarity. Yet, achieving optimal team performance depends critically on proper trust in AI, ensuring humans rely on AI appropriately. In real-world scenarios, humans often lack the expertise or performance transparency to judge AI accuracy directly, creating a gap in appropriate trust calibration. In this paper, we address this challenge through three key contributions: (1) we propose a theoretical framework modeling the evolution of human trust in AI over time under AI performance uncertainty, (2) we investigate two self-calibrating trust methods, an instance-based cognitive model and a reinforcement learning (RL) model that learns trust calibration policies from experience, and (3) we conduct simulations comparing both approaches against a rule-based baseline under dynamically varying AI performance. Results show that RL-based trust calibration outperforms others in cumulative performance, while instance-based calibration offers interpretability and sample efficiency. These findings offer pathways for safe and adaptive trust alignment in human-AI collaboration toward trustworthy autonomy.**

I. Introduction

Human-AI collaboration and complementarity are essential for achieving superior decision-making in complex real-world safety-critical scenarios such as healthcare, scientific discovery, finance, and law [1], [2]. AI can process data at scale and offer consistent predictions as an advisor, while humans bring contextual reasoning, ethical judgment, and take responsibility for team decision-making. To improve team performance with distinct expertise of humans and AI, humans must appropriately calibrate their trust in AI systems. Overtrust may lead to blind acceptance of incorrect AI outputs, while undertrust can prevent beneficial collaboration from AI expertise.

However, trust calibration becomes highly challenging when human users lack access to AI performance or domain expertise to assess AI reliability. Black-box AI systems like deep neural newtorks have poor interpretability, and in human-AI complementarity settings, the human cannot verify every decision made by the AI. Recent works have explored trust calibration by humans assuming humans can observe AI accuracy or audit performance [3], [4], [5], but these assumptions do not hold given distinct human-AI expertise gap in many high-stakes applications.

Specifically, accuracy-based trust calibration by humans are prevalent in the previous work. For example, [6] and [7] show that observed system accuracy strongly influences trust. Yet, this breaks down when accuracy is hidden from humans, as explored in [8], which demonstrates how humans

The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. hanjianh@andrew.cmu.edu

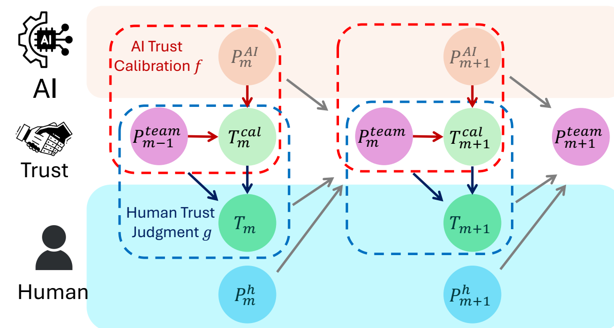

Fig. 1. Trust evolution dynamics in human-AI collaboration. Red dashed box represent AI trust self-calibration policy f, while blue dashed boxes represent human trust judgment g. The team performan P_m^{team} is determined by AI performance P_m^{AI}, human performance P_m^h and human trust in AI T_m at each step m.

overtrust AI when deprived of performance indicators, even when explanations are shown. To address AI trust misalignment, Trust Repair Strategies (TRS) [9] have been proposed, drawing from literature in social psychology [10]. Apology, denial, promise, and model update strategies have been tested in human-robot interaction settings [11], [12], [9]. However, even though TRS mechanisms can help recover the trust in AI without access AI performance, it suffers from the risk of overtrust if not calibrated through human feedback of team performance. Prior studies rarely consider adaptive trust adjustment in dynamic environments or under asymmetric expertise [1], [13], [2].

To this end, our research tackles this challenge through three key contributions:

1) We model trust evolution dynamics in human-AI collaboration as a feedback system linking calibrated trust, human trust judgment, and resulting team performance.
2) We propose two self-calibrating trust models: an instance-based method rooted in cognitive science and a reinforcement learning method trained to align trust with AI performance.
3) We validate both approaches in a simulated environment with dynamically changing AI performance, demonstrating their advantage over heuristic rule-based trust calibration baseline.

II. Trust Evolution and Calibration in Human-AI Collaboration

A. Problem Formulation

Under the AI-as-advisor configuration in human-AI Collaboration, at each step m, the AI has a varying performance score P_m^{AI}, and the human has a more consistent but low performance P_m^{human}. The team performance $P_m(T_m)$ is determined by:

$$P_m(T_m) = T_m \cdot P_m^{AI} + (1 - T_m) \cdot P_m^{human} \tag{1}$$

We assume there exists a calibrated trust score T_m^{cal} given by some self-calibrating policy f , which further influences the human's actual trust T_m through judgment dynamics g. As shown in Fig. 1 the trust evolves as:

$$T_m^{cal} = f(T_{m-1}, P_m^{AI}) \tag{2}$$

$$T_m = g(T_m^{cal}, P_{m-1}(T_{m-1})) \tag{3}$$

The goal is to find a trust calibration policy f that maximizes cumulative team performance $\max \sum_{m=1}^{M} P_m(T_m)$.

B. Instance-Based Trust Calibration

We first propose a trust self-calibration method based on Instance-Based Learning Theory (IBLT) [14], [15], where the model retrieves past experiences to estimate and calibrate trust dynamically. With rich background in cognitive science and psychology, it follows the memory-based dynamic decision-making with instances I_i of AI performance situation P_i^{AI}, decision of calibrated trust T_i^{cal} and utility of team performance P_i based on Eq. (1), i.e. $I_i = (P_i^{AI}, T_i^{cal}, P_i)$. Given the situation of AI performance P_m^{AI} at step m, similar instances are retrieved and blended via:

$$T_m^{cal} = \frac{1}{m} \sum_{i=1}^{m} d_i \cdot T_i^{cal} \tag{4}$$

where $d_i = \mathbf{1}\{|P_i^{AI} - P_m^{AI}| < \delta\}$ measures situation similarity with threshold of δ.

To mimic human trust judgment dynamics g with feedback of the last-step team performance, the self-calibrated final trust used by the human is simplified as the linear combination of T_m^{cal} and previous team performance $P_{m-1}(T_{m-1})$:

$$T_m = (1 - \alpha) T_m^{cal} + \alpha P_{m-1}(T_{m-1}) \tag{5}$$

Each step updates memory for future blending via $P_m^{AI} \leftarrow P_m^{AI}, T_m^{cal} \leftarrow T_m, P_m \leftarrow P_m(T_m)$.

C. Reinforcement Learning-Based Trust Calibration

We then explore reinforcement learning-based AI to learn an optimal policy to adjust its calibrated trust over time to maximize overall team performance.

We model the calibration process as an Markov decision process (MDP): state is AI performance P_m^{AI}, action is calibrated trust and reward is $\sum_m r_m, r_m = P_m(T_m)$ from Eq. (1) by assuming $T_m = T_m^{cal}$ as identity human judgment dynamics. We adopt Q-learning to optimize discrete Q table via ϵ-greedy exploration:

$$T_m^{cal} = \begin{cases} \text{random in } [0,1], & \text{w.p. } \epsilon \\ \arg\max_T Q(P_m^{AI}, T), & \text{otherwise} \end{cases}$$

Then the tabular Q-values are updated as: $Q(P_m^{AI}, T_m^{cal}) \leftarrow Q(P_m^{AI}, T_m^{cal}) + \alpha[r_m + \gamma \max_T Q(P_{m+1}^{AI}, T) - Q(P_m^{AI}, T_m^{cal})]$, where $\epsilon = 0.1, \gamma = 0.9$.

III. Experiment

A. Experimental Setup

We simulate the AI's performance via fluctuating accuracy between 0 and 1 with a decaying sine curve shape controlled by decaying factor η and random noise $\varepsilon \in [-0.05, 0.05]$:

$$P_m^{AI} = \text{clip}(\exp(-\eta m) \cdot \frac{\sin(m) + 1}{2} + \varepsilon, 0, 1) \in [0, 1] \tag{6}$$

The human performance is a constant accuracy of 0.5. We choose the rule-based trust calibration baseline where trust increases or decreases by 0.1 based on better or worse last-step past performance. The decision-making horizon consists of 100 steps, and the RL policy is trained over 100k episodes. Situation similarity threshold is $\delta = 0.1$ and the seed is fixed.

B. Team Performance Comparison

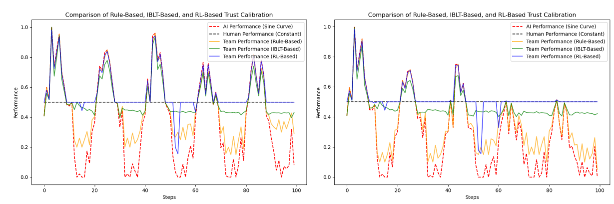

Fig. 2. Team performance for $\eta = 0.01$ (left) and $\eta = 0.03$ (right)

RL-based calibration consistently outperforms other methods across both slow and fast AI performance decay (Fig. 2). It adapts trust to match AI reliability. IBLT shows conservative yet stable behavior, while the rule-based baseline shows poor adaptability and cannot adapt to AI performance changes.

C. Trust Trajectories

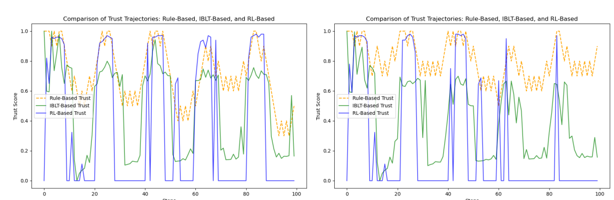

Fig. 3. Trust trajectories under $\eta = 0.01$ (left) and $\eta = 0.03$ (right)

Rule-based trust calibration baseline always suffers from overtrust, even when AI performs worse than humans. IBLT shows similar fluctuations but tends to be conservative due to the blended memory of earlier similar situations. RL shows clearer patterns by adapting trusts dynamically and avoiding overtrust even when AI significantly degrades.

IV. Conclusion

We proposed a framework for trust evolution in human-AI collaboration, supported by two learning-based methods: an interpretable IBLT cognitive learning model and a high-performing RL agent. Simulations confirm that RL consistently achieves strong performance under dynamic AI accuracy. Future work includes user studies and hybrid trust calibration models combining learning with explicit social strategies.

REFERENCES

[1] C. Gonzalez, P. Fakhari, and J. Busemeyer, "Dynamic decision making: Learning processes and new research directions," *Human factors*, vol. 59, no. 5, pp. 713–721, 2017.

[2] J. Li, Y. Yang, Q. V. Liao, J. Zhang, and Y.-C. Lee, "As confidence aligns: Exploring the effect of ai confidence on human self-confidence in human-ai decision making," *arXiv preprint arXiv:2501.12868*, 2025.

[3] Q. Zhang, M. L. Lee, and S. Carter, "You complete me: Human-ai teams and complementary expertise," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–28, 2022.

[4] M. Nourani, J. King, and E. Ragan, "The role of domain expertise in user trust and the impact of first impressions with intelligent systems," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, pp. 112–121, 2020.

[5] K. Inkpen, S. Chappidi, K. Mallari, B. Nushi, D. Ramesh, P. Michelucci, V. Mandava, L. H. Vepřek, and G. Quinn, "Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making," *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 5, pp. 1–29, 2023.

[6] K. Yu, S. Berkovsky, D. Conway, R. Taib, J. Zhou, and F. Chen, "Trust and reliance based on system accuracy," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pp. 223–227, 2016.

[7] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12, 2019.

[8] S. Pareek, E. Velloso, and J. Goncalves, "Trust development and repair in ai-assisted decision-making during complementary expertise," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 546–561, 2024.

[9] S. Pareek, N. van Berkel, E. Velloso, and J. Goncalves, "Effect of explanation conceptualisations on trust in ai-assisted credibility assessment," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–31, 2024.

[10] E. J. De Visser, R. Pak, and T. H. Shaw, "From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction," *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, 2018.

[11] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. de Vries, "Trust repair in human-agent teams: the effectiveness of explanations and expressing regret," *Autonomous agents and multi-agent systems*, vol. 35, no. 2, p. 30, 2021.

[12] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, "Toward an understanding of trust repair in human-robot interaction: Current research and future directions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 4, pp. 1–30, 2018.

[13] C. Gonzalez, "Building human-like artificial agents: A general cognitive algorithm for emulating human decision-making in dynamic environments," *Perspectives on Psychological Science*, vol. 19, no. 5, pp. 860–873, 2024.

[14] C. Gonzalez, J. F. Lerch, and C. Lebiere, "Instance-based learning in dynamic decision making," *Cognitive Science*, vol. 27, no. 4, pp. 591–635, 2003.

[15] T. Lejarraga, V. Dutt, and C. Gonzalez, "Instance-based learning: A general model of repeated binary choice," *Journal of Behavioral Decision Making*, vol. 25, no. 2, pp. 143–153, 2012.