The Effects of Social Intelligence on Trust in Human-AI Teams

Morgan Bailey[a,1], Benjamin Gancz[b] and Frank Pollick[a]
[a] *University of Glasgow*
[b] *Qumodo, London*

Abstract. This study investigates trust calibration in Human-AI teams, specifically focusing on design choices that enhance the perception of AI's human-like qualities in a workplace setting. Participants engaged in a geolocation task, collaborating with an AI teammate, a human teammate, and themselves as the team leader. The AI's responses varied in style, either resembling human responses with logical explanations or providing one-word locations. The reliability of the AI's answers was manipulated, with high (90% correct) or low (60% correct) levels. Overall, participants exhibited greater trust in AI with low humanness and high reliability. However, when reliability was low, participants trusted the humanized AI more. These findings highlight the complex nature of the relationship of Human-AI teams and offer insights into trust calibration and team dynamics.

Keywords. Human-AI Teams, Human-AI Dynamic Team Trust, Social AI, Hybrid intelligence, Trust

1. Introduction

In recent decades, the rapid advancement of Artificial Intelligence (AI) has had a transformative impact on various aspects of society. Within the realm of work, AI has demonstrated superior performance in tasks involving extensive data analysis, high precision, and prolonged cognitive workload. However, research has consistently highlighted the effectiveness of collaboration between humans and AI, known as hybrid intelligence, in achieving optimal outcomes[1, 5]. Consequently, there is growing interest in understanding the dynamics of Human-AI teams (HATs) to implement AI in the workforce effectively.

Trust emerges as a crucial variable in the design of HATs, as it underlies critical team dynamics. While an appropriate level of trust is essential in HATs, excessive trust can lead to over-reliance on AI systems, causing users to overlook mistakes and errors [3]. Conversely, insufficient trust may result in team members underutilizing the capabilities of AI, leading to diminished team performance[1].

Calibrating trust in HATs involves transitioning from black-box AI methods to explainable AI, enabling successful trust calibration. Presenting AI outputs in a humanized manner, infused with elements of Social Intelligence (SI) [5], serves as a means to explain AI and facilitate trust calibration. SI encompasses social categories such as understanding, memory, perception, creativity, and knowledge [5], and has effectively calibrated trust in human teams[4].

[1] Corresponding Author: Morgan Bailey, m.bailey.1@research.gla.ac.uk.

A key objective in forming HATs is to elevate the role of AI from a mere tool to that of a teammate. To establish HATs as social entities, the presence of multiple individuals, shared goals, interdependency, and distinct roles and functions are essential drivers[6]. Additionally, fostering a sense of team spirit, cohesion, and a collective identity among team members is crucial for HATs to develop as social entities.

Researchers have employed anthropomorphism to enhance AI's human-like and social aspects, aiming to calibrate trust. However, the effectiveness of anthropomorphism remains subject to constraints and ongoing debates [7]. Furthermore, the embodiment of AI can vary, ranging from complete embodiment to complete disembodiment. This study specifically investigates the creation of feelings of SI through written feedback. Building upon existing literature, we propose the following hypothesis:

H1: Increased humanness in AI responses will influence the decision-making process when determining which teammate to trust.

2. Method

The study employed a Wizard of Oz experimental method to facilitate development convenience and ensure optimal control. Participants were led to believe they were collaborating with an AI and a human teammate when instead they were actually interacting with pre-written responses. The task involved presenting participants with random locations extracted from Google Earth. Participants were tasked with determining the continent, country, and city associated with each location, with the final decision resting on the participant, who assumed the role of the 'team leader'. A time constraint of 90 seconds per location was enforced, meaning participants had to rely on their teammates' responses to submit the location in time. Notably, the AI and human teammates provided conflicting answers 90% of the time, necessitating participants to discern which teammate they trusted more. A total of 30 locations were identified by each participant across three blocks, comprising 10 trials per block. Following each trial, the correct answer was revealed, enabling participants to assess the performance of the human and AI teammates.

The experimental design employed a 2x2 between-subject design. Participants were assigned to groups based on their interaction with an AI with either high or low levels of humanness and exhibiting either high reliability (90% accuracy) or low reliability (60% accuracy) of correct answers. The human teammate consistently achieved 30% accuracy and gave the same responses in all conditions.

In the low humanness group, the AI was introduced as an AI, emphasizing technical aspects and delivering only the answer. Conversely, in the high humanness conditions, the AI presented itself as 'Pixie' and expressed enthusiasm about being a teammate, thereby enhancing perceptions of humanness. Additionally, the AI offered responses similar to those of the human teammate, adopting an anthropomorphic writing style.

For each trial, participants rated how much each team member influenced them. This rating was performed using a slider tool that assessed the degree of influence exerted by each teammate. The influence rating was utilised as an implicit measure of trust, drawing on prior research suggesting that greater influence corresponds to higher levels of trust. To further explore the influence results, participants completed the Godspeed Questionnaire[8].

3. Results

A total of 44 participants from the University of Glasgow were recruited. The group consisted of 16 males, 25 females, 2 non-binary, and one participant who chose not to disclose their gender.

We conducted a two-way ANOVA with interactions to compare trust levels towards AI. The quantitative variable was the influence score given during the experiment, while the independent variables were the assigned reliability and humanness levels. The analysis revealed a significant difference in influence ratings based on reliability level ($F(1, 1) = 48.20$, $p < 0.001$) and an interaction effect between reliability and humanness ($F(1) = 19.34$, $p < 0.001$). Further pairwise comparisons using a Tukey post hoc test demonstrated significant differences, as shown in Table 1. The mean trust scores by group were High Humanness High Reliability (μ=4.13), High Humanness Low Reliability (μ=3.91), Low Humanness High Reliability (μ=4.61) and Low Humanness Low Reliability (μ=3.69).

Table 1. Tukey Multiple Comparisons of Means for ANOVA

Group	Mean Diff	Lower	Upper	P Adjusted
Humanness	0.12	-0.03	0.21	0.120
Reliability	**-0.55**	**-0.71**	**-0.40**	**0.000**
LH:HR – HH:HR	**0.48**	**0.19**	**0.77**	**0.000**
HH:LR – HH:HR	-0.22	-0.50	0.07	0.196
LH:LR – HH:HR	**-0.44**	**-0.73**	**-0.15**	**0.001**
HH:LR – LH:HR	**-0.70**	**-0.99**	**-0.41**	**0.000**
LH:LR – LH:HR	**-0.91**	**-1.22**	**-0.62**	**0.000**
LH:LR- HH:LR	-0.22	-0.51	0.07	0.200

*Bold results indicate significance. HH – High Humanness, LH – Low Humaneness, HR – High Reliability, LR- Low Reliability.

In the anthropomorphism sub-section of the Godspeed Questionnaire, we observed a significant difference based on humanness level ($F(1,1) = 15.49$, $p < 0.001$). Similarly, in the likeability sub-section, a significant difference was found based on humanness ($F(1,1) = 11.98$, $p < 0.001$). Additionally, in the perceived intelligence subsection, we identified statistically significant results based on reliability ($F(1, 1) = 10.26$, $p < 0.01$).

4. Discussion

The hypotheses are generally supported, although the relationship is intricate. Trust levels differed based on the conditions of humanness and reliability. Participants exhibited higher trust in AI with low humanness and high reliability, these findings collaborate previous research which indicates a reduced cognitive load can increase trust[9]. Conversely, when reliability was low, trust levels were higher in AI with high humanness. This indicates a complex and dynamic relationship between humanness, trust, and HATs.

Moreover, participants demonstrated greater trust in humanized AI compared to non-humanized AI when reliability was low. This suggests that as AI performance declined, participants relied on the human-like responses to understand the AI's decision-making process, leading to increased trust. These findings have implications for trust calibration in team settings, highlighting the dynamic nature of the relationship. However,

caution must be exercised as a humanized AI may potentially foster overreliance in a system with subpar performance.

Additionally, the questionnaire data revealed that participants found humanized AI more likable than non-humanized AI. The preference for humanized AI is significant, as likability can have an immediate impact on short-term trust[10] and facilitate the initial adoption of a system. These results also imply that HATs can emulate non-AI teams, where likability influences team performance[11]. Future research should explore how team members rate the likability of AI with varying degrees of humanness, which will inform design choices for trust calibration. It is advisable to corroborate these findings with more comprehensive qualitative methods to gain deeper insights into users' ratings of AI with social intelligence (SI).

Further investigations should focus on gaining a comprehensive understanding of trust dynamics in teams. This can be achieved by deliberately manipulating levels of reliability and humanness and exposing participants to multiple AI versions to capture their preferences. Conducting qualitative research will be instrumental in obtaining an in-depth understanding of team dynamics.

References

[1] Kamar E. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In-IJCAI 2016 Jul 9, 4070-4073.

[2] Wang D, Weisz JD, Muller M, Ram P, Geyer W, Dugan C, et al. Human-AI collaboration in Data Science. Proceedings of the ACM on Human-Computer Interaction. 2019;3(CSCW):1–24.

[3] Wang N, Pynadath DV, Hill SG. Trust calibration within a human-robot team: Comparing automatically generated explanations. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2016Aug14;

[4] Alonso V, de la Puente P. System transparency in shared autonomy: A mini review. Frontiers in Neurorobotics. 2018;12.

[5] Williams J, Fiore SM, Jentsch F. Supporting artificial social intelligence with theory of mind. Frontiers in Artificial Intelligence. 2022;5.

[6] Rix J. From Tools to Teammates: Conceptualising Humans' Perception of Machines as Teammates with a Systematic Literature Review.

[7] Mou W, Ruocco M, Zanatto D, Cangelosi A. When would you trust a robot? a study on trust and theory of mind in human-robot interactions. In 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) 2020 Aug 31 (pp. 956-962). IEEE.

[8] Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics. 2009 Jan;1:71-81

[9] Zhou J, Arshad SZ, Luo S, Chen F. Effects of uncertainty and cognitive load on user trust in Predictive decision making. Human-Computer Interaction – INTERACT2017. 2017 Sept 21;23–39. doi:10.1007/978-3-319-68059-0_2

[10] Nagel DM, Giunipero L, Jung H, Salas J, Hochstein B. Purchaser perceptions of early phase supplier relationships: The role of similarity and likeability. Journal of Business Research. 2021May;128:174–86.

[11] Nowlin EL, Walker D, Anaza N. The impact of manager likeability on sales performance. Journal of Marketing Theory and Practice. 2019 Apr 3;27(2):159-73.

Acknowledgements

Morgan Bailey's work was supported by the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents, Grant Number EP/S02266X/1.