Autonomy, Agency, and Trust: Towards Integrated Calibration in Human-Robot Interaction

Ali Fallahi[1], Patrick Holthaus[1], Farshid Amirabdollahian[1], Gabriella Lakatos[1]

***Abstract*— This position paper argues that perceived agency is the key mediator linking robot autonomy and user trust in human-robot interaction (HRI). The main focus of this work is on autonomy and agency as two important robot-related elements. Building on our previous work where participants interacted with a Pepper robot framed as either autonomous or remotely controlled, this paper emphasises the need to integrate nuanced trust calibration mechanisms in HRI. The future direction of this research includes analysing behavioural video recordings and participants' open-ended responses to further understand how trust is behaviourally and cognitively manifested. This position paper proposes that a more holistic analysis—incorporating behavioural, verbal, and task-specific indicators will advance our understanding of trust dynamics in HRI.**

I. Introduction

Trust significantly influences HRI, and serves as a critical indicator of both functional and social acceptance, particularly in decision-making scenarios under uncertainty [3]. Trust in HRI often builds on foundational organisational trust frameworks, such as Mayer et al.'s [13] integrative model, which emphasises ability, benevolence, and integrity. Hancock et al. [7] classify factors influencing trust into human-related, robot-related, and environmental categories. This paper specifically addresses robot-related factors—autonomy and perceived agency arguing that perceived agency plays a critical mediating role in shaping user trust. Our position, grounded in empirical research, is that effective trust calibration in HRI requires a comprehensive understanding of how robot autonomy influences perceived agency and, consequently, trust.

II. Literature Review

The literature distinguishes between autonomy and agency in HRI. Autonomy is defined as the robot's capacity to act independently without direct human control [12], while perceived agency refers to the extent to which humans attribute intentionality and decision-making capacity to robots [21]. Previous research has shown that robots framed as more autonomous are consistently perceived as more competent and capable collaborators in mixed-initiative tasks [17].Where autonomy is a technical property of the robot, agency is a human-attributed quality: the sense that the robot has a mind of its own. Mind-perception theory posits two orthogonal dimensions—experience and agency—both of which drive moral judgment [5]. In HRI, social-agency theory argues that users infer agency from interactivity, adaptability, and autonomy [10]. Trafton et al. [20] in their work showed manipulating a robot's decision latitude (none vs. shared vs. full) had a monotonic increase in cognitive and behavioural trust via perceived agency ratings. Chanseau et al. [1] explored whether robot appearance influences people's perceptions of task criticality, finding that while appearance can affect user expectations about correct task performance, the classification of task criticality often depends more on task type and perceived safety than on robot appearance. Salem et al. [15] investigated how robot errors and task types influence human trust and cooperation in HRI. Their findings show that robot performance strongly shapes the trust participants report; however, this change in attitude does not always translate into behaviour, as participants often do not follow the robot's advice or instructions which known as behavioural compliance [6], [8]. This discrepancy between perceived and behavioural trust underscores the importance of integrating objective behavioural analysis into trust research, which is a key direction of this project. Additionally, Salem et al. [14] discussed the ethical and practical challenges in designing trustworthy social robots, highlighting the critical risk of overreliance on robots, even when participants recognise system faults. Their work emphasises the importance of task criticality, perceived risk, and long-term trust calibration in HRI, reinforcing the need for behavioural measures and careful experimental design in trust research.

Although a few number of the earlier studies like Salem et al. [15] and Strohkorb Sebo et al. [19] have already combined behavioural video coding with qualitative analysis of participants' free-text explanations, further work is needed to systematically relate these multimodal data to specific cues of autonomy and agency. Our planned inclusion of fine-grained video annotations and open-ended responses therefore represents an important next step toward closing this gap. To complement behavioural observations, trust measurement scales like the Trust Perception Scale-HRI developed by Schaefer [16] offer validated tools for quantifying user trust in HRI. Holthaus et al. [9] investigated the relationship between perceived agency and trust across different age groups, finding that although priming robot autonomy did not significantly affect trust across conditions, age appeared to influence participants' attitudes toward robots. Their findings support the complexity of trust calibration and the influence of individual differences on trust in HRI. Related research has shown that specific zoomorphic embodiments shape the emotions [11] and intentions people attribute to a robot

[1]Robotics Research Group, University of Hertfordshire, Hatfield, United Kingdom {a.fallahi, p.holthaus, f.amirabdollahian2, g.lakatos}@herts.ac.uk

signals of benevolence and predictability that directly increase perceived trustworthiness [18]. Nevertheless, this PhD focuses primarily on trust and agency rather than on embodiment itself. While the present research focuses on trust and agency rather than emotional expression, it is noteworthy that Ghafurian et al. [4] demonstrated how zoomorphic robots' affective expressions can shape user perceptions and engagement. Their work highlights the importance of considering social cues and user anthropomorphism tendencies when evaluating robot interactions.

III. Research Questions and Analytical Focus

The aim of this PhD research is to explore how trust in robots is shaped by autonomy, perceived agency, embodiment, and task type. The research addresses the following questions:

- **RQ-1:** Does the autonomy of a robot affect its perceived agency?
- **RQ-2:** How might people's perception of robot agency influence their trust towards companion robots?
- **RQ-3:** Does the embodiment of a robot affect its perceived agency and trust?
- **RQ-3.1:** How do various types of tasks affect user trust in robots, considering the robots' embodiment?

This research aims to explore these questions through a series of experimental interactions and behavioural and qualitative analyses. It is expected that various framings of autonomy may influence how users interpret agency and form trust judgments. Additionally, characteristics such as task type and robot embodiment may play subtle but meaningful roles in shaping these dynamics, building on insights from prior literature on social presence, agency attribution, and trust calibration in HRI.

Rationale:

This investigation builds on theoretical and empirical work indicating that how robots are framed can influence users' perceptions of their autonomy and decision making capacity [21]. However, trust in HRI is not a simple consequence of perceived autonomy—it is shaped by multiple interacting factors [7]. To better understand these complexities, future phases of the research will employ behavioural video coding to examine trust-related cues such as response latencies, task engagement, verbal confirmations, and physical interactions. Thematic analysis of participants' open-ended responses will further provide nuanced insights into how trust, agency, and autonomy are interpreted and rationalised in real-time interactions.

IV. Summary of Previous Work

Our previous study [2], investigated how autonomy framing affected perceived agency and behavioural trust using a Pepper robot. Thirty-three participants interacted with the robot under autonomous or remotely controlled framing across tasks including ID verification, feeding a cat, playing Sudoku, and dancing the robot. The study found that while autonomy framing influenced perceived sincerity (a subdimension of trust), it did not consistently predict behavioural trust. Further analysis of behavioural video data and qualitative responses will enhance our understanding of these trust mechanisms.

V. Future Work

The next phase of this PhD research will focus on:

- Detailed behavioural video coding to quantify trust-related behaviours such as task completion, response speed, hesitation, physical contact, and verbal confirmations.
- Thematic analysis of participants' open-ended responses to better understand their reasoning behind trust-related decisions.
- Expansion to new experiments examining how robot embodiment influences perceived agency and trust.
- Comparative task analysis to explore how social versus functional tasks differentially impact trust formation.

VI. Conclusion

Robot autonomy influences trust primarily through perceived agency, but trust calibration in HRI is complex and cannot rely on framing effects alone. This research proposes that a multi-modal analysis (integrating quantitative, behavioural, and qualitative data) will provide a more comprehensive understanding of trust formation in human-robot interaction. Future studies focusing on embodiment and task specificity will further refine trust calibration strategies in social robotics.

References

[1] Chanseau, A., Dautenhahn, K., Walters, M.L., Koay, K.L., Lakatos, G., Salem, M.: Does the appearance of a robot influence people's perception of task criticality? In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 1057–1062. IEEE (2018). https://doi.org/10.1109/ROMAN.2018.8525663

[2] Fallahi, A., Holthaus, P., Amirabdollahian, F., Lakatos, G.: Effects of Perceived Robot Autonomy and Personal Differences on Trust in Human-Robot Interactions. In: International Conference on Social Robotics (ICSR 2025). **Springer**, Naples, Italy (in press)

[3] Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., Ivaldi, S.: Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to icub answers. Computers in Human Behavior **61**, 633–655 (2016). https://doi.org/10.1016/j.chb.2016.03.057

[4] Ghafurian, M., Lakatos, G., Dautenhahn, K.: The zoomorphic miro robot's affective expression design and perceived appearance. International journal of social robotics pp. 1–18 (2022). https://doi.org/10.1007/s12369-021-00832-3

[5] Gray, H.M., Gray, K., Wegner, D.M.: Dimensions of mind perception. science **315**(5812), 619–619 (2007). https://doi.org/10.1126/science.1134475

[6] Gurney, N., Pynadath, D.V., Wang, N.: Comparing psychometric and behavioral predictors of compliance during human-ai interactions. In: International Conference on Persuasive Technology. pp. 175–197. Springer (2023). https://doi.org/10.1007/978-3-031-30933-5_12

[7] Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. Human factors **53**(5), 517–527 (2011). https://doi.org/10.1177/0018720811417254

[8] Haring, K.S., Satterfield, K.M., Tossell, C.C., De Visser, E.J., Lyons, J.R., Mancuso, V.F., Finomore, V.S., Funke, G.J.: Robot authority in human-robot teaming: Effects of human-likeness and physical embodiment on compliance. Frontiers in Psychology **12**, 625713 (2021). https://doi.org/10.3389/fpsyg.2021.625713

[9] Holthaus, P., Fallahi, A., Förster, F., Menon, C., Wood, L., Lakatos, G.: Agency Effects on Robot Trust in Different Age Groups. In: International Conference on Human-Agent Interaction (HAI 2024). ACM, Swansea, UK (2024). https://doi.org/10.1145/3687272.3690903

[10] Jackson, R.B., Williams, T.: A theory of social agency for human-robot interaction. Frontiers in Robotics and AI **8**, 687726 (2021). https://doi.org/10.3389/frobt.2021.687726

[11] Lakatos, G., Holthaus, P., Sharma, P., Velmurugan, V., Hamilton-Holbrook, T., Riches, L., Moros, S., Wood, L.: Does a "robot dog" need legs, ears, and tail? a comparative analysis of intention-and emotion-attribution to miro-e and unitree go1. Biologia Futura pp. 1–15 (2025). https://doi.org/10.1007/s42977-025-00263-5

[12] Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human factors **46**(1), 50–80 (2004). https://doi.org/10.1518/hfes.46.1.50$_3$0392

[13] Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Academy of management review **20**(3), 709–734 (1995). https://doi.org/10.5465/amr.1995.9508080335

[14] Salem, M., Lakatos, G., Amirabdollahian, F., Dautenhahn, K.: Towards safe and trustworthy social robots: ethical challenges and practical issues. In: Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7. pp. 584–593. Springer (2015). https://doi.org/10.1007/978-3-319-25554-5$_5$8

[15] Salem, M., Lakatos, G., Amirabdollahian, F., Dautenhahn, K.: Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction. pp. 141–148 (2015). https://doi.org/10.1145/2696454.2696497

[16] Schaefer, K.E., Chen, J.Y., Szalma, J.L., Hancock, P.A.: A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human factors **58**(3), 377–400 (2016). https://doi.org/10.1177/001872081663422

[17] Schermerhorn, P., Scheutz, M.: Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task. In: Proceedings of the 2011 international conference on advances in computer-human interactions. pp. 236–241 (2011)

[18] Song, Y., Tao, D., Luximon, Y.: In robot we trust? the effect of emotional expressions and contextual cues on anthropomorphic trustworthiness. Applied Ergonomics **109**, 103967 (2023). https://doi.org/10.1016/j.apergo.2023.103967

[19] Strohkorb Sebo, S., Traeger, M., Jung, M., Scassellati, B.: The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In: Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction. pp. 178–186 (2018). https://doi.org/10.1145/3171221.3171275

[20] Trafton, J.G., McCurry, J.M., Zish, K., Frazier, C.R.: The perception of agency. ACM Transactions on Human-Robot Interaction **13**(1), 1–23 (2024). https://doi.org/10.1145/3640011

[21] Waytz, A., Heafner, J., Epley, N.: The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of experimental social psychology **52**, 113–117 (2014). https://doi.org/10.1016/j.jesp.2014.01.005