Exploring the Influence of Human-Likeness on Attributing Intentionality to Robot Errors*

Raffaella Esposito[1,*], Alessandra Rossi[1] and Silvia Rossi[1]

[1]University of Naples Federico II, Via Cintia, 80126 Naples, Italy

Abstract

Trusting a robot involves perceiving its actions as driven by a benevolent purpose, making intentionality attribution a psychological mechanism worthy of attention in Human–Robot Interaction (HRI).

By integrating findings from studies on intentionality bias in HRI, we highlight a gap in the literature and discuss implications for user expectations and trust. In particular, we know that people often explain robot mistakes as deliberate choices, but we do not yet know whether this judgment hinges on how human-like the robot appears.

We argue that a humanoid appearance will amplify attributions of agency and purpose, whereas a mechanical guise will steer observers toward design-based or accidental explanations. Demonstrating this effect would pinpoint when embodiment alone reshapes error interpretation, revealing when and how a robot's appearance alters the perceived intentionality behind its actions.

Keywords

robot errors, intentionality attribution in HRI, robot human-likeness

1. Introduction

People are driven by a spontaneous tendency to make sense of the world around them. For this reason, when we observe someone act, we almost automatically ask ourselves, "Why did they do that?". To explain someone's behavior, we assume the person had motives, goals, feelings, and other mental states, and those assumptions help us turn a stream of actions into a coherent story.

The human mind shows a strong intentionalistic bias, preferring explanations of others' behavior as intention-driven even when information is ambiguous. Indeed, when context does not offer immediate mechanical explanations, intention attribution reduces uncertainty and facilitates prediction of future events [1]. Once an act is framed as intentional, the door opens to evaluating the actor's motives, an appraisal that, in turn, underpins moral judgement and social expectations. In this way, intentionality attribution becomes the precondition for deciding whether another agent can be trusted [2].

According to the trust literature, trust matters precisely because the trustee possesses the capacity to choose actions that could either benefit or harm the trustor [3]. In fact, trust is defined as the willingness of the trustor to expose themselves to vulnerability in relation to the trustee [4]. For this reason, the trustee must be regarded as capable of deliberate, goal-directed behavior [5, 6]. In other words, a party can only be trusted if they are viewed as having agency; without that perception, the very concept of trust collapses [7, 3].

Investigating whether humans extend their intentionality bias to robots is therefore worthy of attention. Viewing robots as intentional agents can make users attribute benevolence to the robot's motives and can help trust repair [8, 9]. Understanding when and why intentionality attributions arise will clarify one of the hinges on which human–robot trust turns.

Here, we argue that a moment that may be particularly revealing of these psychological mechanisms is the moment when a robot fails. An error forces observers to decide whether the slip stems from blind mechanics or from the choices of an intentional agent, and such a claim is supported by experimental

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy

*This manuscript has been typeset using the official CEUR–WS template.

*Corresponding author.

raffaella.esposito3@unina.it (R. Esposito)

© 2025 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4074/paper8-8.pdf

evidence [10, 11]. We also argue that the robot's human-likeness could amplify the inference that an inner decision-making system lies behind a robot's error.

We believe that the degree of anthropomorphism may influence the extent to which a robot's errors are perceived as intentional or mere malfunctions. A humanoid form, with its familiar human-like features and expressive capabilities, may serve as a powerful cue for interpreting an error as a deliberate act with underlying motives. Studying how people explain robotic errors across a continuum of human-likeness may offer a natural test-bed for tracing when intentionality is conferred, and whether trust is ultimately eroded or repaired.

2. Intentionality Attribution to Robots

Understanding how the degree of anthropomorphism influences intentionality attributions for robot errors raises a broader question: do people attribute intentionality to robots in the first place?

Studies indicate that people use similar social-cognitive tools to explain both human and robot actions [12]. In particular, the correspondence bias (i.e., the tendency to explain behaviour in terms of dispositional choice despite situational constraints) operates for both human and robotic agents. In experiments with the humanoid robot *Pepper* [13], observers attributed volitional choice to the robot even after the experimental script made clear that its actions were externally programmed; the bias grew stronger when Pepper voiced a counter-normative stance, signalling an opinion that clashed with social expectations.

Functional-imaging work complements these behavioural findings: activity in classic Theory-of-Mind regions—the medial prefrontal cortex, temporoparietal junction and posterior superior temporal sulcus—rises linearly with a robot's human-likeness, from mechanical devices through zoomorphic platforms to fully anthropomorphic embodiments [14]. This graded neural response suggests that perceived agency is neurally encoded well before any explicit judgement is made.

When people confront complex robot behaviours whose internal logic they cannot fully parse, they seem to default to inferring intentions; if the behavioural pattern then breaks, they interpret the deviation as a deliberate act of opposition [15]. Consistent with this, Short [10] and Ullman *et al.* [16] showed that humanoid robots programmed to cheat during games drew markedly stronger attributions of intent. Further, Ciardo *et al.* [11] demonstrated that the type of error matters: observers framed a clearly mechanical malfunction as a design glitch, whereas a more human-like slip sustained their mentalistic reading of the robot's behaviour. Taken together, these findings indicate that deviations from normative scripts or user expectations may amplify intentionality attributions.

Beyond errors and cheating, subtler non-verbal cues also invite mental-state inferences. Human partners read intention into robots' gaze shifts [17] and reactive micro-movements [18] using the same heuristics deployed in human–human interaction.

Overall, while people may not view robots as fully equivalent to humans in terms of intentionality, they do ascribe mental states and intentions to robots to varying degrees depending on the robot's design and behaviour. Whether, and in what way, these two factors interact remains an open question.

3. Robot Errors and Anthropomorphism: An Open Question

Although numerous studies have shown that errors enacted by humanoid robots prompt attributions of purpose and mental states [10, 11], no work to date has directly compared an anthropomorphic-looking robot with a mechanical one making the same error.

Figuring out whether the anthropomorphic envelope alone is enough to cast an error as an intentional act means deciding whether the robot body functions as a magnifying glass for any future moral and relational assessment. If the blame falls on the plastic face rather than the robot's programming, then aesthetics becomes ethics. Thus, we would expect that the more deliberate an error appears to be, the more our trust in the robot will swing—upward when we infer benevolent motives, downward when we deem them malevolent or negligent.

4. Conclusions

This paper set out to bridge two strands of research: (i) studies showing that humanoid appearance alone can trigger an intentionality bias, and (ii) studies demonstrating that certain robot errors invite mental-state attributions. Bringing these lines together highlights an untested junction: does the very same error elicit different intentional readings solely because of the body that performs it?

Robots will disappoint us. The question is not *if* but *how* we will explain those disappointments.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA8655-23-1-7060. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] B. F. Malle, Attribution theories: How people make sense of behavior, Wiley (2022) 93–120.
[2] B. Vanneste, P. Puranam, Artificial intelligence, trust, and perceptions of agency, Acad. Manage. Rev. (2024). doi:10.5465/amr.2022.0041.
[3] R. Hardin, Trust and Trustworthiness, Russell Sage Fdn., New York, NY, 2002.
[4] Schilke, Trust in social relations, Annual Review of Sociology 47 (2021) 239–259. doi:10.1146/annurev-soc-082120-082850.
[5] C. Taylor, Human Agency and Language, Cambridge Univ. Press, Cambridge, UK, 1985.
[6] M. W. Morris, T. Menon, D. R. Ames, Culturally conferred conceptions of agency: A key to social perception of persons, groups, and other actors, in: Lay Theories and Their Role in the Perception of Social Groups, Psychology Press, 2003, pp. 169–182.
[7] D. M. Rousseau, S. B. Sitkin, R. S. Burt, C. Camerer, Not so different after all: A cross-discipline view of trust, Acad. Manage. Rev. 23 (1998) 393–404.
[8] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon, M. L. Tielman, Taxonomy of trust-relevant failures and mitigation strategies, in: Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction (HRI), 2020, pp. 3–12. doi:10.1145/3319502.3374793.
[9] Z. Rezaei Khavas, A review on trust in human-robot interaction, arXiv (2021) arXiv:2105.???
[10] E. Short, J. Hart, M. Vu, B. Scassellati, No fair!! an interaction with a cheating robot, in: Proceedings of the 5th ACM/IEEE International Conference on Human–Robot Interaction (HRI), 2010, pp. 219–226. doi:10.1109/HRI.2010.5453193.
[11] F. Ciardo, D. De Tommaso, A. Wykowska, Effects of erring behavior in a human-robot joint musical task on adopting intentional stance toward the icub robot, in: Proceedings of the 2021 IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2021, pp. 698–703.
[12] M. M. A. De Graaf, B. F. Malle, People's explanations of robot behavior subtly reveal mental state inferences, in: Proceedings of the 2019 ACM/IEEE International Conference on Human–Robot Interaction (HRI), 2019, pp. 239–248.
[13] A. Edwards, C. Edwards, Does the correspondence bias apply to social robots?: Dispositional and situational attributions of human versus robot behavior, Front. Robot. AI 8 (2022) 788242. doi:10.3389/frobt.2021.788242.

[14] S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski, T. Kircher, Can machines think? interaction and perspective taking with robots investigated via fmri, PLoS ONE 3 (2008) e2597. doi:10.1371/journal.pone.0002597.

[15] Y. Imamura, K. Terada, H. Takahashi, Effects of behavioral complexity on intention attribution to robots, in: Proceedings of the 3rd International Conference on Human–Agent Interaction (HAI), 2015, pp. 65–72.

[16] D. Ullman, L. Leite, J. Phillips, J. Kim-Cohen, B. Scassellati, Smart human, smarter robot: How cheating affects perceptions of social agency, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 36, 2014.

[17] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, N. Hagita, Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior, in: Proceedings of the 4th ACM/IEEE International Conference on Human–Robot Interaction (HRI), 2009, pp. 69–76. doi:10.1145/1514095.1514110.

[18] K. Terada, T. Shamoto, A. Ito, H. Mei, Reactive movements of non-humanoid robots cause intention attribution in humans, in: Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2007, pp. 3715–3720.