Can robots trust us? Neuropsychophysiological insights into honesty towards Artificial Agents

Martini, M.; Hooft, D. van; Kret, M.E.

Can Robots Trust *Us*? Neuropsychophysiological Insights into Honesty Towards Artificial Agents

Fabiola Diana
Department of Cognitive Psychology
Leiden University
Leiden, The Netherlands
d.fabiola@fsw.leidenuniv.nl

Ruud Hortensius
Department of Psychology
Utrecht University
Utrecht, The Netherlands
r.hortensius@uu.nl

Mariska E. Kret
Department of Cognitive Psychology
Leiden University
Leiden, The Netherlands
m.e.kret@fsw.leidenuniv.nl

2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI) | 979-8-3503-7893-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/HRI61500.2025.10974055

***Abstract*— While research often examines human trust in robots, humans frequently cheat in interactions—even simple ones. This raises a neglected question: can robots trust us? Our interdisciplinary, cross-cultural research explores honesty in human-robot interaction by capturing real-time psychophysiological signals through a comparative robotics framework. We investigate whether social cues, like pupil size, that influence human-human decisions similarly impact human-robot interactions. Future work will expand this to study how human physiological and neural responses shape behavior in social dilemmas with human vs artificial partners, guiding the development of agents attuned to human signals.**

Keywords—psychophysiology, HRI, honesty, network analysis

I. Introduction

Trust underpins human collaboration and societal progress. This is no different when the interaction partner is artificial: to collaborate or have a meaningful relationship with social robots, chatbots, or other forms of artificial agents, people must trust them in the first place, a question widely explored in literature [1]. Yet, trust is a two-way street. Our research explores a complementary, overlooked question: can artificial agents "trust" us? People are often tempted to cheat, [2] as evidenced by dishonest behaviors that entail a cost to society, like tax fraud, downloading illegal software and music, or bypassing self-checkouts. As artificial agents play an increasingly important role in our society, a deep understanding of (dis)honest behaviors towards robots, and of how *mutual* trust can be established between humans and robots is a significant challenge that must be addressed.

Deciding to act honestly or not is always a dilemma between what is good for oneself and what is beneficial for the other individual. This can be influenced by cognitive processes – past experiences, cultural norms, and cost-benefit analyses – but also by more intuitive unconscious processes that often override cognition-driven evaluations [3], especially in absence of previous information. These rapid, automatic judgments about others are based on a complex interplay between a variety of conscious behavioral signals (e.g., facial expressions, posture) [4,5], unconscious autonomic cues (e.g., pupil size, blushing) [6] and bodily responses (e.g., heart rate, skin conductance, brain activity) [7], which facilitate the intuitive prediction of each other behavior and shape subsequent interactions [8]. During social interaction, these signals, cues, and responses interplay in a continuous mutual exchange that complements our cognitive processes. For instance, pupil size seems to have a substantial effect on the way we perceive others [9], [10]: human faces with large pupils may be perceived as more attractive [11] and trustworthy [12]. Moreover, people seem to be more honest when interacting with a partner with large pupil size [10]. Like language, non-verbal communication varies significantly across cultures and contexts, adding layers of complexity to human interaction.

When interacting through or with technological mediums, humans may (or may not!) apply the same heuristics they would use with other living beings [13]. For example, in previous work [14], we demonstrated that people mimic non-verbal cues to a similar extent in video calls as in face-to-face interactions. Although this study focused on select cues, it suggests that people might employ the same intuitive judgments when interacting via technology as they do in real life. Of course, while mimicking cues from a person on the other side of a video call is one thing, it's another when the interaction partner is an artificial agent. This raises a fundamental question: how do humans interpret these cues when they come from a non-human? Artificial agents are designed with specific visual elements—eyes, faces—that convey certain impressions [15, 16], yet we understand little about how these design choices impact human behavior unconsciously, especially given the unfamiliarity of such interactions. Further, since non-verbal behaviors vary across cultures, it's essential to investigate whether these cues, when applied to artificial agents, elicit culturally distinct responses.

Building on these insights, our current studies aim to address three central research questions: **RQ1.** Are people more dishonest when interacting with artificial agents than with humans? **RQ2.** To what extent do unconscious cues (e.g., pupil size) shape interactions and perceptions of artificial agents as they do in human interactions? **RQ3.** How do these dynamics vary across cultures, revealing potential cultural sensitivities in human-artificial agent interactions?

II. CURRENT WORK

Our research sought to investigate whether individuals would exhibit greater dishonesty when interacting with artificial agents compared to humans, and if unconscious cues would influence this behavior. To this end, our current work employs a modified coin-toss game adapted from [10], which is well-known to elicit dishonest behaviors. Participants engage with three types of agents: a humanoid avatar on a screen, a machine-like robot, and a human confederate acting as a genuine participant. Across all agent types, we manipulate pupil size to be either dilated or constricted. For human confederates, this is achieved using customized contact lenses. Assessing pupil size over other signals has two main advantages: it is an unconscious cue beyond the participant's control, reducing potential bias; it can be consistently manipulated across diverse agent types while minimizing uncanny valley effects, which is crucial given the significant differences between our interaction agents. Besides, we continuously measured pupil size changes in the participants' eyes. To ensure a diverse sample and environment, we recruited European participants both from a controlled laboratory (N=97) setting and in the wild from the general public (N=103). This approach allows us to gather data from a relatively heterogeneous sample, enhancing the validity and generalizability of our findings. Our preliminary results suggest that our participants show a clear difference in dishonesty between human and artificial agents, being more honest toward the human. Large pupils seem to amplify these differences, whereas small pupils tend to equalize the levels of dishonesty across agent types.

Given that social cues like eye gaze and pupil size are culturally contextual, we replicated this study in Japan to examine potential cross-cultural differences. Previous literature shows that European and Japanese participants hold distinct explicit attitudes toward robots and avatars – with Japanese being more favorable to artificial agents compared to Europeans – but no difference was found at the implicit level [17]. These cultural nuances prompted us to replicate the study with a Japanese sample (N=75), who completed the same experiment as their European counterparts. Results, albeit preliminary, show that Japanese participants seem to behave equally dishonestly towards the human and artificial agents, with one notable exception: when playing with agents displaying constricted pupils, a slight difference emerged between avatars and human agents. Collectively, our findings suggest that subtle physiological cues like pupil size can influence dishonesty, with different effects across cultures and agent types.

III. FUTURE WORK

A. *Network Analysis of Neuropsychophysiological Signals for Human-Centered AI*

Our current studies on the influence of cues on dishonesty highlighted a critical asymmetry in HRI research: while humans leverage social and emotional intelligence when assessing the trustworthiness of robot systems, current artificial agents do not (yet) possess the same sophisticated capabilities to fully interpret and respond to these non-verbal and often unconscious (dis)honesty signals. A key research question, then, is what honesty signals can be identified in human-robot interaction that would help AI systems respond meaningfully, and how do they interplay together (**RQ4**)? This question is central to our future work, where we will integrate methods from developmental psychology, social neuroscience, and human-robot interaction to perform a multi-modal network-based analysis. Sticking to our current approach, participants will engage in a real-life honesty game with a robot and with another human. We will continuously record facial expressions, heart rate, skin conductance, and pupil dilation, in combination with brain activity – measured through mobile functional near-infrared spectroscopy (fNIRS) over the temporoparietal junction. These signals, crucial for assessing emotional arousal [18], decision-making [19], and social cognition [20], have mostly been collected in isolation [21], with only limited insights gathered during real interaction with artificial agents. To better contextualize the observed behavior and physiological responses, we will also collect multiple self-report measures from the participants. These will include attitudes and previous experience with AI, baseline honesty traits, reward sensitivity, and the level of anthropomorphism they ascribe to the agents. Our comparative network analysis approach will identify specific patterns and dynamics between the bio-behavioral data (facial expressions, physiology, neural activity, self-report), showing similarities and differences in the interaction with robots and humans.

B. *Generational AI*

Our current cross-cultural research highlights the importance of individual differences to understand honesty towards artificial agents. While culture plays a role, generational differences are also a key. Dishonest behavior may vary across generations [22], as well as attitudes toward AI [10]: baby boomers may view robots as distant or fictional, millennials as an emerging reality, and Gen Z as part of daily life. We have no guarantee that an artificial agent developed and optimized for one group of users will work for users of different ages as well. How do these generational differences influence honesty towards artificial agents (**RQ5**)? To address this, we plan to invite individuals from different generations (20-, 40-, and 60-year-olds) to play with the robots or other humans. We will build network models for each age group offering similarities and differences in bio-behavioral signatures of dishonesty towards artificial agents and providing a direct and mechanistic input for human-robot interactions. Our project will explore how these cues vary across ages, revealing generational similarities and barriers, and facilitating artificial agent designs that accommodate a range of users, especially older adults at risk of digital exclusion.

IV. TOWARDS REAL SOCIAL ROBOTS

The interdisciplinary journey from understanding honesty in human-human interaction to decoding patterns of biobehavioral

signals in human-robot interaction put our research at the cutting edge of HRI. By merging social neuroscience, neuropsychophysiology, and comparative robotics, we will shed light on how honesty is perceived and reciprocated in human-robot interactions. In addition, albeit limited, we have the opportunity to understand what circumstances trigger dishonest behavior in people. Integrating these insights into artificial systems could substantially improve human-robot collaboration, making future artificial agents more intuitive and human-centered.

Our holistic approach aligns with the broader conceptual framework proposed in our recent paper [23], where we call for a shift toward a "bottom-up" approach in social robotics that emphasizes spontaneous sociality emerging from robot-robot interactions. By exploring the fundamental, spontaneous mechanisms of social behaviors in robots, we can gain a more comprehensive understanding of the core principles that drive sociality across species and environments. This perspective complements our research focus by suggesting that understanding how fundamental behaviors emerge in robots could spotlight why certain low-level cues (e.g., pupil size) are used in complex social decisions. Through evolutionary robotics, we can experiment with and accelerate the conditions that foster these cues, potentially revealing how honesty indicators develop in both humans and artificial agents.

ACKNOWLEDGMENT

Past and present work is supported by ERC grant (804582 to M.E.K. and 101117045 to R.H.) and by the JSPS Short Term Fellowship awarded to F.D (PE22041).

REFERENCES

[1] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Acad. Manag. Ann.*, vol. 14, no. 2, pp. 627–660, 2020, doi: 10.5465/annals.2018.0057.

[2] C. Jacobsen, T. R. Fosgaard, and D. Pascual-Ezama, "Why Do We Lie? a Practical Guide To the Dishonesty Literature," *J. Econ. Surv.*, vol. 32, no. 2, pp. 357–387, 2018, doi: 10.1111/joes.12204.

[3] E. Prochazkova and M. E. Kret, "Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion," *Neurosci. Biobehav. Rev.*, vol. 80, no. October 2016, pp. 99–114, 2017, doi: 10.1016/j.neubiorev.2017.05.013.

[4] K. U. Likowski, A. Mühlberger, A. B. M. Gerdes, M. J. Wieser, P. Pauli, and P. Weyers, "Facial mimicry and the mirror neuron system: simultaneous acquisition of facial electromyography and functional magnetic resonance imaging," *Front. Hum. Neurosci.*, vol. 6, no. July, pp. 1–10, 2012, doi: 10.3389/fnhum.2012.00214.

[5] B. Tia, A. Saimpont, C. Paizis, F. Mourey, L. Fadiga, and T. Pozzo, "Does observation of postural imbalance induce a postural reaction?," *PLoS One*, vol. 6, no. 3, 2011, doi: 10.1371/journal.pone.0017799.

[6] M. E. Kret, "Emotional expressions beyond facial muscle actions. A call for studying autonomic signals and their impact on social perception," *Front. Psychol.*, vol. 6, no. May, pp. 1–10, 2015, doi: 10.3389/fpsyg.2015.00711.

[7] R. V. Palumbo *et al.*, "Interpersonal Autonomic Physiology: A Systematic Review of the Literature," *Personal. Soc. Psychol. Rev.*, vol. 21, no. 2, pp. 99–141, 2017, doi: 10.1177/1088868316628405.

[8] R. T. Boone and R. Buck, "Emotional Expressivity and Trustworthiness: The Role of Nonverbal Behavior in the Evolution of Cooperation," *J. Nonverbal Behav.*, vol. 27, no. 3, pp. 163–182, 2003.

[9] M. E. Kret, "The role of pupil size in communication. Is there room for learning?," *Cogn. Emot.*, vol. 32, no. 5, pp. 1139–1145, 2018, doi: 10.1080/02699931.2017.1370417.

[10] J. A. van Breen, C. K. W. De Dreu, and M. E. Kret, "Pupil to pupil: The effect of a partner's pupil size on (dis)honest behavior," *J. Exp. Soc. Psychol.*, vol. 74, no. November 2017, pp. 231–245, 2018, doi: 10.1016/j.jesp.2017.09.009.

[11] S. Tombs and I. Silverman, "Pupillometry - A sexual selection approach," *Evol. Hum. Behav.*, vol. 25, no. 4, pp. 221–228, 2004, doi: 10.1016/j.evolhumbehav.2004.05.001.

[12] M. E. Kret and C. K. W. De Dreu, "Pupil-mimicry conditions trust in partners: Moderation by oxytocin and group membership," *Proc. R. Soc. B Biol. Sci.*, vol. 284, no. 1850, pp. 1–10, 2017, doi: 10.1098/rspb.2016.2554.

[13] K. J. Kim, "Heuristics in digital communication media: theoretical explications and empirical observations," *Qual. Quant.*, vol. 49, no. 5, pp. 2187–2201, 2015, doi: 10.1007/s11135-014-0103-y.

[14] F. Diana, O. E. Juárez-mora, W. Boekel, R. Hortensius, and M. E. Kret, "How video calls affect mimicry and trust during interactions," *Phil. Trans. R. Soc. B*, vol. 378, 2023.

[15] S. Woods, "Exploring the design space of robots: Children's perspectives," *Interact. Stud.*, vol. 18, pp. 1390–1418, 2006, doi: 10.1016/j.intcom.2006.05.001.

[16] E. Phillips, D. Ullman, M. M. A. De Graaf, and B. F. Malle, "What does a robot look like?: A multi-site examination of user expectations about robot appearance," *Proc. Hum. Factors Ergon. Soc.*, vol. 2017-Octob, pp. 1215–1219, 2017, doi: 10.1177/1541931213601786.

[17] F. Diana, M. Kawahara, I. Saccardi, R. Hortensius, A. Tanaka, and M. E. Kret, "A Cross-Cultural Comparison on Implicit and Explicit Attitudes Towards Artificial Agents," *Int. J. Soc. Robot.*, vol. 15, no. 8, pp. 1439–1455, 2023, doi: 10.1007/s12369-022-00917-7.

[18] H. A. Shehu, M. Oxner, W. N. Browne, and H. Eisenbarth, "Prediction of moment-by-moment heart rate and skin conductance changes in the context of varying emotional arousal," *Psychophysiology*, vol. 60, no. 9, pp. 1–15, 2023, doi: 10.1111/psyp.14303.

[19] G. Forte, M. Morelli, B. Grässler, and M. Casagrande, "Decision making and heart rate variability: A systematic review," *Appl. Cogn. Psychol.*, vol. 36, no. 1, pp. 100–110, 2022, doi: 10.1002/acp.3901.

[20] M. Ahmad and A. Alzahrani, "Crucial Clues: Investigating Psychophysiological Behaviors for Measuring Trust in Human-Robot Interaction," *ACM Int. Conf. Proceeding Ser.*, pp. 135–143, 2023, doi: 10.1145/3577190.3614148.

[21] I. Ben Ajenaghughrure, S. Da, C. Sousa, and D. Lamas, "Measuring Trust with Psychophysiological Signals : A Systematic Mapping Study of Approaches Used," *Multimodal Technol. Interact.*, vol. 4, no. 63, pp. 1–29, 2020, doi: 10.3390/mti4030063.

[22] A. M. O. Connor, R. A. Judges, K. Lee, and A. D. Evans, "Examining honesty – humility and cheating behaviors across younger and older adults," *Int. J. Behav. Dev.*, vol. 46, no. 2, pp. 112–117, 2022, doi: 10.1177/01650254211039022.

[23] F. Diana, L. Cañamero, R. Hortensius, and M. E. Kret, "Merging sociality and robotics through an evolutionary perspective," *Sci. Robot.*, vol. 9, no. eadk6664, 2024, doi: 10.1126/scirobotics.adk6664.