Plausible Explanations Reduce Phantom Cost Perception in HRI

Benjamin Lebrun
University of Canterbury
Christchurch, New Zealand
benjamin.lebrun@pg.canterbury.ac.nz

Christoph Bartneck
University of Canterbury
Christchurch, New Zealand
christoph.bartneck@canterbury.ac.nz

Andrew J. Vonasch
University of Canterbury
Christchurch, New Zealand
andrew.vonasch@canterbury.ac.nz

Abstract

Recent studies found that people imagine phantom costs—bad intentions and risks—when a human or a robot makes an overly generous offer without sufficient explanations. However, these studies used a paradigm in which the agent justified a cookie with $2 offer by saying they had eaten cookies with friends, a scenario we think implausible for robots. The present study replicated this paradigm while measuring perceived plausibility of the agent's justification. Results indicated that, unlike humans, the justification of eating cookies with friends was perceived as implausible when said by a robot. This perceived implausibility increased perceived phantom costs, reduced trust in the robot, and decreased offer acceptance. This study suggests that phantom costs occur when explanations are both insufficient and implausible, highlighting the need for sufficient and plausible explanations to promote effective HRI.

CCS Concepts

• **Human-centered computing** → **User models**; **Heuristic evaluations**; **User studies**; **HCI theory, concepts and models**; **Empirical studies in HCI**.

Keywords

Decision-making, HRI, Perceived Plausibility, Phantom Costs, Trust

ACM Reference Format:
Benjamin Lebrun, Christoph Bartneck, and Andrew J. Vonasch. 2026. Plausible Explanations Reduce Phantom Cost Perception in HRI. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3757279.3788804

1 Introduction

As robots become more integrated into daily life, understanding how people respond to them and interpret their behaviors is essential. Recent studies showed that people perceive phantom costs—such as hidden motives and risks—when robots make overly generous offers without a clear rationale [7, 8]. These findings align with prior work in human-human interactions (HHI) [18, 19]. However, extending this framework to human-robot interaction (HRI) raises additional questions about how explanations are evaluated. Specifically, Lebrun et al. [7, 8] adapted the cookie paradigm—originally developed for HHI [19]—by introducing a robot agent. In their study, either a human or robot agent offered a cookie with or without $2 and justified the offer by saying they had cookies left over after eating with friends. While this justification seems acceptable in HHI, it may be less appropriate in HRI, as people may not expect robots to eat or have friends. Thus, explanations that seem sufficient to justify an offer in HHI may not be perceived as plausible when communicated by a robot.

We propose that the "perceived plausibility" of information—i.e., how believable the information is given the agent producing it—shapes people's responses. In the present study, agents differed in nature, being either a human or a robot. However, our primary interest lay in the robot agent, as perceptions of plausibility for a human saying that they ate cookies with friends are less likely to vary across participants. To our knowledge, no prior work has directly explored this in HRI, although some suggest its importance: people are epistemically vigilant toward information communicated by robots [2, 3, 15] and inconsistencies between robot features and behaviors can reduce trust [5, 16].

2 Objectives and Implications in HRI

This study has two objectives. First, it aims to replicate prior work on phantom costs using the same experimental cookie paradigm [7, 8]. Second, it extends this paradigm by incorporating measures of perceived plausibility and trust to examine how perceived plausibility shapes phantom costs, decision-making, and trust in HRI. Integrating insights from phantom costs and epistemic vigilance, we predict that perceived implausibility of information communicated by a human or robot agent elicits suspicion—and by extension, phantom costs—and reduces perceived trustworthiness of this agent, thereby decreasing acceptance of the offer. We further hypothesize that "moral trust"—reflecting benevolence and good intentions [10]—will be more affected than "performance trust"—reflecting competence and accurate information [1]—because phantom costs relate to concerns about an agent's ulterior motives.

Theoretically, this study extends current knowledge on phantom costs by introducing "perceived plausibility," offering a more comprehensive framework for understanding how humans perceive and respond to robots. Practically, the findings inform the design of robot communication strategies that increase transparency [21] and trust, thereby promoting effective HRI.

3 Methods

We used a 2 (Agent: Human vs. Robot) × 2 (Offer: Cookie vs. Cookie + $2) between-participants design, similarly to Lebrun et al. [7, 8]. In addition, we included questions relevant to the purposes of this study on the last pages of the experiment. These questions were preregistered at https://aspredicted.org/SMT_QQ2 and approved by the Human Research Ethics Committee (ref HREC 2023/93/LR-PS).

HRI '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3788804

3.1 Participants

We recruited 321 American participants via Prolific to complete a study hosted on Qualtrics, using a representative sampling of the US population in terms of age, gender, and race. They were all fluent in English and at least 18 years old. They were compensated £0.4 for completing the study (about 3 minutes).

3.2 Stimuli

Pictures. The pictures (Figure 1) depicted a human or Nao robot (Aldebaran Robotics) agent, standing upright and holding a transparent container of cookies in the right hand. Depending on the condition, the left hand was either empty or held a $2 coin.

(a) Human - Cookie

(b) Human - Cookie + $2

(c) Robot - Cookie

(d) Robot - Cookie + $2

Figure 1: Pictures Showing the Agents holding the Offers. Reprinted with permission from [7], © IEEE 2025.

Scenario. Four scenarios, each corresponding to one experimental condition illustrated in Figure 1, described the situation and offer to participants. The scenarios were written as follows:

> During a break, you go for a walk. A [person/robot] carrying a clear container with several cookies starts talking to you:
> "Hi, excuse me, I've just been having food with my friends and we had some cookies left over. Would you like one? [I'll pay you 2 dollars if you eat it.]"

3.3 Measures

Decision-making. Participants made a forced binary choice indicating whether they would eat the cookie offered by the agent.

Justification. Participants justified their decision in an open-ended text response. Those who provided irrelevant justifications or mentioned dietary restrictions were excluded from data analyses.

Phantom costs. Participants assessed the intentions of the agents using 7-point Likert scales (1 = "Strongly disagree" to 7 = "Strongly agree"). These statements were "The [agent] had good intentions for offering me the cookie" and "The [agent] had hidden reasons for offering me the cookie."

Plausibility. Plausibility of the agent's statement was measured using a 7-point Likert scale (1 = "Strongly disagree" to 7 = "Strongly agree") via the statement "It is plausible that a [agent] might have just eaten cookies with friends."

Trust. Trust was assessed using a shortened version of the MDMT v2 [10, 11]. We selected the most relevant items: Benevolence ("benevolent" and "has goodwill"), Transparent ("genuine" and "sincere"), Ethical ("ethical" and "moral"), Competent ("competent"), and Reliable ("reliable").

3.4 Procedure

After providing informed consent, participants were randomly assigned to one of the four scenarios. The study was organized across multiple pages. On page 1, participants read the scenario, viewed the associated picture, and indicated whether they would eat the cookie. On page 2, they justified their decision. On page 3, they completed the items assessing phantom costs. On page 4, they evaluated the perceived plausibility of the agent's explanation. On page 5, they completed the MDMT. Finally, participants were presented with a debriefing page and were given the opportunity to withdraw their data before being redirected to Prolific for compensation.

4 Results

4.1 Participants

Twenty-nine participants were excluded due to dietary restrictions (n = 18), inattention (n = 1), misunderstanding the scenario (n = 1), or extreme-value outliers that strongly influenced the assumptions of analyses (n = 9). The final sample size comprised 292 participants and included 141 males, 144 females, and 7 participants who declared their gender to be different. Ages ranged from 18 to 81 years old ($M = 45.22, SD = 15.75$). Race was distributed as follows: White or Caucasian (70.65%), Black or African American participants (13.87%), Asian (6.77%), and ~15% reported Hispanic origins. Conducting analyses on the sample before outliers exclusion (N = 301) led to the same pattern of results and identical conclusions.

4.2 Manipulation Check: Plausibility

We conducted an ANOVA to investigate differences in perceived plausibility of the agent's statement. Results indicated a main effect of Agent, $F(1, 288) = 239.58, p < .001, \eta_p^2 = .454$. Post-hoc comparisons using Tukey corrections revealed that "eating cookies with friends" was judged as more plausible when said by a human ($M = 5.45, SD = 0.12$) than by a robot ($M = 2.75, SD = 0.12$), $t(288) = 15.48, p < .001, d = 1.814$. There was no effect of Offer ($p = .745$) nor Agent × Offer interaction effect ($p = .052$).

4.3 Phantom Costs

We conducted a multiple linear regression to examine the effects of Agent, Offer, Perceived Plausibility, and their interactions, on phantom costs (bad intentions). The model was significant, $F(7, 284) = 18.80, p < .001$, and explained 29.98% of the variance (adjusted R^2). Results (Figure 2) revealed that the robot was perceived as having less bad intentions than the human ($\beta = -3.59, p = .001$). Additionally, higher perceived plausibility was associated with lower attributions of bad intentions ($\beta = -0.75, p < .001$). The Agent ×

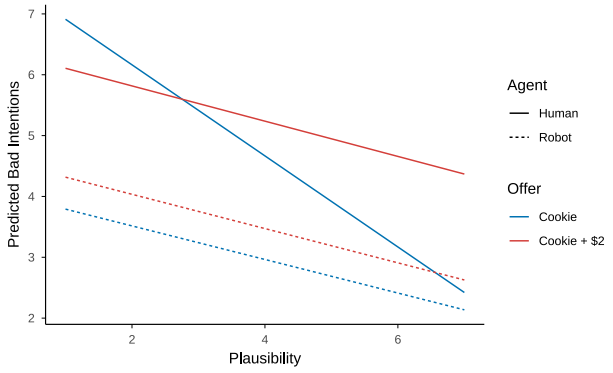

Figure 2: Perceived Bad Intentions as a Function of Perceived Plausibility by Agent and Offer.

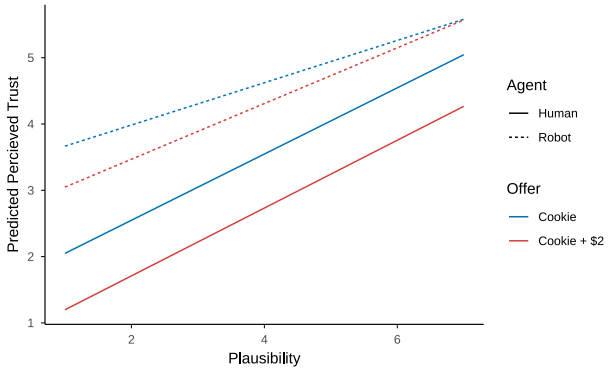

Figure 3: Perceived Trust in the Agent as a Function of Perceived Plausibility by Agent and Offer.

Plausibility interaction was significant ($\beta = 0.47, p = .022$), suggesting that the effect of Plausibility was higher for the robot than for the human. The Offer × Plausibility interaction was also significant ($\beta = 0.46, p = .040$), indicating that Plausibility had a stronger effect when the offer was a cookie plus $2 rather than a cookie.

4.4 Decision-Making

We conducted a binomial logistic regression to explore how Agent, Offer, Perceived Plausibility, and their interactions predicted participants' likelihood of eating the cookie. Results revealed an effect of Agent: participants were more likely to eat the cookie when offered by a robot than by a human, $\beta = 11.75, z = 2.95, p = .003$. Higher plausibility also increased acceptance: for each one-unit increase in perceived plausibility, participants were 7.68 times more likely to eat the cookie, $z = 3.14, p = .002$. The Agent × Plausibility interaction showed that the effect of Plausibility differed between both agents $\beta = -1.64, z = -2.46, p = .014$. The other effects were not significant, with p-values ranging between .077 and .115. The Agent × Plausibility that was no longer significant ($\beta = -0.77, p = .116$) when analyses were conducted on the sample before outliers exclusion (N = 301).

4.5 Trust

4.5.1 Reliability analyses. Reliability analyses using Cronbach's alpha showed high internal consistency across trust sub-scales: benevolence ($\alpha = .841$), transparency ($\alpha = .954$), and ethical ($\alpha = .948$). All items within the moral-trust dimension were highly consistent ($\alpha = .953$). The two items assessing performance-trust were highly consistent ($\alpha = .892$). Overall, the eight items showed very high consistency ($\alpha = .955$). Therefore, we averaged items within each sub-scale and across all to create their composite variables.

4.5.2 Predictors of trust. Multiple linear regression analyses were conducted to examine the effects of Agent, Offer, Perceived Plausibility, and their interactions, on overall trust. The model was significant, $F(7, 284) = 17.14, p < .001$, and explained 27.97% of the variance (adjusted R^2). Results (Figure 3) showed that the higher the perceived plausibility, the higher the trust ($\beta = 0.50, p = .006$). The other effects were not significant ($p > .05$).

4.5.3 Predictors of Moral and Performance Trusts. To investigate whether the effect of plausibility differed between moral and performance trust, we compared the two Plausibility slopes from regression models predicting either moral or performance trust, and Plausibility, Agent, and Offer as the predictors. The difference between the two slopes was not statistically different, $z = -0.08, p = .934$.

4.6 Power Analyses

Our main predictions concerned the effects of perceived plausibility on phantom costs and decision-making. For phantom costs, G*Power 3.1. indicated that the study was well-powered (power = 0.99), based on: $f^2 = 0.1605$, $\alpha = .05$, N = 292, one tested predictor, and seven total predictors. Under these assumptions, a minimum total sample size of 52 participants was required to have a power of 0.8. For decision-making, analyses revealed a similar power, and a minimum total sample size of 74 participants for a power of 0.8.

5 Discussion

This replication study aimed to deepen our understanding of prior findings on phantom costs [7, 8] by exploring how the perceived plausibility of an agent's justification for an overly generous offer influences phantom costs, trust, and decision-making. Overall, the more plausible participants perceived the agent's explanation to be, the more likely they were to accept the offer, the less phantom costs they perceived, and the more they trusted the agent.

5.1 Perceived Plausibility

Although some participants considered it plausible that a robot eats cookies with friends, it was generally perceived as less plausible than when expressed by a human. Additionally, we showed that perceived plausibility differs from "insufficiency" of the information described in Vonasch et al. [19], that is, the extent to which an explanation was considered good enough to justify the agent's offer [8]. Sufficiency hence depended on the offer rather than the agent. In contrast, plausibility was not related to the offer but rather to whether the explanation was believable given the agent communicating it.

These findings align with research showing that people are attentive to information communicated by robots [2, 3, 15]. Our findings also add nuance to the concept of epistemic vigilance [15], suggesting that content and source are not necessarily judged independently. While eating cookies with friends is a plausible justification, it becomes implausible when communicated by a robot. Such an explanation may be interpreted as an attempt to make the robot appear more human-like. However, because the justification conflicts with people's expectations about robots, it may increase uncanniness and negative perceptions of the robot [12].

5.2 Phantom Costs and Decision-Making

Consistent with our predictions, decision-making and phantom costs were shaped by perceived plausibility. Specifically, perceived plausibility decreased phantom costs and increased acceptance of the offer; and the lower the perceived plausibility, the more phantom costs and the lower the acceptance rate. These findings may help reconcile differences across prior studies where Lebrun et al. [8] observed lower acceptance rates for the cookie + $2 offer whereas Lebrun et al. [7] did not find such differences. These discrepancies may reflect factors beyond beyond perceived benefits [7] but also whether the agent's explanation is perceived as plausible. When explanations are implausible, people are more likely to perceive phantom costs.

Although [3] showed that people perceive the inconsistency between a robot's statement and its physical appearance, they also found that some participants tend to infer "concealed capacities" to the robot. For this reason, we think that even when a robot justifies its behavior with an implausible statement, some people will simply rationalize the explanation, as a coping strategy to prevent them from perceiving suspicion that might be more cognitively costly.

5.3 Trust

We used the second version of the MDMT [11] to investigate whether higher perceived plausibility increased trust in the agent and whether plausibility affected moral and performance trust dimensions differently. Despite using a reduced set of the MDMT items, reliability scores supported the use of these measures in both HRI and HHI contexts. As predicted, higher perceived plausibility was associated with greater trust, likely because plausible explanations reduced phantom costs, thereby increasing trust, consistent with [6]. Additionally, moral trust did not differ between both agents but was lower when the agent offered money. This is consistent with findings showing that this offer violates the norm of self-interest [19] and may be perceived as morally questionable, and the agents as more likely to be deceptive. These results suggest that people expect robots to adhere to moral norms even if they differ from those applying to humans [9, 17, 20]. In contrast, robots were perceived as more trustworthy on the performance trust dimension—which included "reliable" and "competent". This may be because people expect robots to know what they do and tend to over-estimate their capabilities and reliability [14]. These findings highlight the need for calibrated trust in HRI [4]. Providing appropriate explanation for the robot's actions [21] might help people develop trust levels adequate for the situation, thereby reducing unwanted phantom costs and facilitating interactions.

Contrary to our predictions, perceived plausibility did not differently affect moral and performance trust. This may seem counterintuitive at first as phantom costs are not expected to be related to performance trust but to moral trust, such as benevolence and goodwill [11]. One possible explanation may be that performance-trust items such as "reliable" overlap with moral-trust [13], increasing the perceived impact on performance trust. Additionally, participants who considered that a robot could eat cookies with friends had no reason to doubt the robot's performance, especially if they over-trust the robot's capabilities [14].

5.4 Limitations

First, perceived plausibility was measured using a single item. Future work should develop and validate a multi-item questionnaire to obtain more reliable evidence of this effect. Similarly, perceived plausibility was not directly manipulated in the present study. However, because our goal was to closely replicate the original paradigm, we deliberately retained the original explanation. Second, the visual depiction of Nao may imply functional limitations that could have influenced how plausible the explanation was perceived to be. Third, although eating cookies with friends is generally plausible for humans, a few participants considered it implausible. This may reflect that some participants did not expect such a justification to be provided when making an offer and may have interpreted it as assessing the sufficiency of the explanation rather than its plausibility. For this reason, conclusions regarding the human conditions should be treated with caution but findings concerning the robot remain robust. Fourth, although trust was assessed using the MDMT, potential overlap between moral and performance trust dimensions—particularly the item "reliable"—may have reduced sensitivity to detect differences between the two slopes. Finally, we did not measure baseline trust prior to the interaction, preventing direct conclusions about how the scenario changed trust levels.

6 Conclusion and Future Research

This study replicated prior findings on phantom costs in HRI, confirming that overly generous offers can elicit suspicion and affect decision-making. Extending this, we introduced and examined "perceived plausibility" as a novel factor shaping these responses. We found that when an agent's justification for an offer was perceived as implausible, people imagined more phantom costs, trusted the agent less, and were less likely to accept the offer. Conversely, higher plausibility reduced phantom costs and increased acceptance.

Theoretically, these findings refine our understanding of phantom costs by demonstrating that explanations are evaluated in terms of their sufficiency (relative to the offer) and plausibility (relative to the agent). By linking phantom costs to plausibility, trust, and decision-making, this study connects phantom costs to broader concepts such as epistemic vigilance, trust calibration, and explainable robotics. Practically, the findings suggest that robot explanations should be both sufficient and plausible—aligning them with the robot's perceived capabilities—to reduce phantom costs and promote effective HRI. Future research should directly manipulate perceived plausibility of human and robot explanations (e.g., Low vs. High) to better understand its role in phantom costs and decision-making.

References

[1] Naeimeh Anzabi and Hiroyuki Umemuro. 2023. Influence of Social Robots' Benevolence and Competence on Perceived Trust in Human-Robot Interactions. *The Japanese Journal of Ergonomics* 59, 6 (2023), 258–273. doi:10.5100/jje.59.258

[2] Robin Gigandet, Maria C. Diana, Kenza Ouadada, and Tatjana A. Nazir. 2024. Beyond Explicit Acknowledgment: Brain Response Evidence of Human Skepticism towards Robotic Emotions. doi:10.3390/robotics13050067

[3] Robin Gigandet, Xénia Dutoit, Bing Li, Maria C. Diana, and Tatjana A. Nazir. 2023. The "Eve effect bias": Epistemic Vigilance and Human Belief in Concealed Capacities of Social Robots. In *2023 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, Piscataway, NJ, USA, 15–20. doi:10.1109/ARSO56563.2023.10187469

[4] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. de Visser, and R. Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Hum Factors* 53, 5 (2011), 517–27. doi:10.1177/0018720811417254

[5] Kim Klüber and Linda Onnasch. 2022. Appearance is not everything - Preferred feature combinations for care robots. *Computers in Human Behavior* 128 (2022), 107128. doi:10.1016/j.chb.2021.107128

[6] Benjamin Lebrun, Christoph Bartneck, David Kaber, and Andrew Vonasch. 2025. People Perceive More Phantom Costs From Autonomous Agents When They Make Unreasonably Generous Offers. arXiv:2511.07401 [cs.HC] doi:10.48550/arXiv.2511.07401

[7] Benjamin Lebrun, Christoph Bartneck, and Andrew Vonasch. 2025. Phantom Costs in Human-Robot Interaction: A Replication Study. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE Press, Piscataway, NJ, USA, 1037–1041. doi:10.1109/HRI61500.2025.10974228

[8] Benjamin Lebrun, Andrew Vonasch, and Christoph Bartneck. 2024. Too good to be true: People reject free gifts from robots because they infer bad intentions. arXiv:2404.07409 [cs.HC] doi:10.48550/arXiv.2404.07409

[9] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 117–124. doi:10.1145/2696454.2696458

[10] Bertram F. Malle and Daniel Ullman. 2021. *Chapter 1 - A multidimensional conception and measure of human-robot trust*. Academic Press, Cambridge, MA, USA, 3–25. doi:10.1016/B978-0-12-819472-0.00001-0

[11] Bertram F. Malle and Daniel Ullman. 2023. Measuring Human-Robot Trust with the MDMT (Multi-Dimensional Measure of Trust). arXiv:2311.14887 [cs.RO] https://arxiv.org/abs/2311.14887

[12] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. doi:10.1109/MRA.2012.2192811

[13] Zahra Rezaei Khavas, Monish Reddy Kotturu, S. Reza Ahmadzadeh, and Paul Robinette. 2024. Do Humans Trust Robots that Violate Moral Trust? *J. Hum.-Robot Interact.* 13, 2, Article 25 (jun 2024), 30 pages. doi:10.1145/3651992

[14] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Piscataway, NJ, USA, 101–108. doi:10.1109/HRI.2016.7451740

[15] Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. Epistemic vigilance. *Mind & Language* 25, 4 (2010), 359–393. doi:10.1111/j.1468-0017.2010.01394.x

[16] Angela Tinwell, Mark Grimshaw, and Andrew Williams. 2010. Uncanny behaviour in survival horror games. *Journal of Gaming and Virtual Worlds* 2, 1 (2010), 3–25. doi:10.1386/jgvw.2.1.3_1

[17] John Voiklis, Boyoung Kim, Corey Cusimano, and Bertram F. Malle. 2016. Moral judgments of human vs. robot agents. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Piscataway, NJ, USA, 775–780. doi:10.1109/ROMAN.2016.7745207

[18] Andrew Vonasch, Natasha Dore, and Joelle Felicite. 2022. The Psychological Origins of Conspiracy Theory Beliefs: Big Events with Small Causes Amplify Conspiratorial Thinking. doi:10.31234/osf.io/3j9xg

[19] Andrew J. Vonasch, Reyhane Mofradidoost, and Kurt Gray. 2024. People Reject Free Money and Cheap Deals Because They Infer Phantom Costs. *Personality and Social Psychology Bulletin* 0, 0 (2024), 01461672241235687. doi:10.1177/01461672241235687

[20] Aleksandra Wasielewska. 2021. Expectations towards the Morality of Robots: An Overview of Empirical Studies. *Ethics in Progress* 12, 1 (2021), 134–151. doi:10.14746/eip.2021.1.10

[21] Robert H. Wortham, Andreas Theodorou, and Joanna J. Bryson. 2017. Robot Transparency: Improving Understanding of Intelligent Behaviour for Designers and Users. In *Towards Autonomous Robotic Systems*, Yang Gao, Saber Fallah, Yaochu Jin, and Constantina Lekakou (Eds.). Springer International Publishing, Cham, Switzerland, 274–289. doi:10.1007/978-3-319-64107-2_22

Received 2025-10-07; accepted 2025-12-09