PLAUSIBILITY OF A ROBOT'S EXPLANATION SHAPES PHANTOM COSTS

BENJAMIN LEBRUN*, CHRISTOPH BARTNECK, & ANDREW VONASCH

* PhD Candidate in Psychology, University of Canterbury, New Zealand, benjamin.lebrun@pg.canterbury.ac.nz

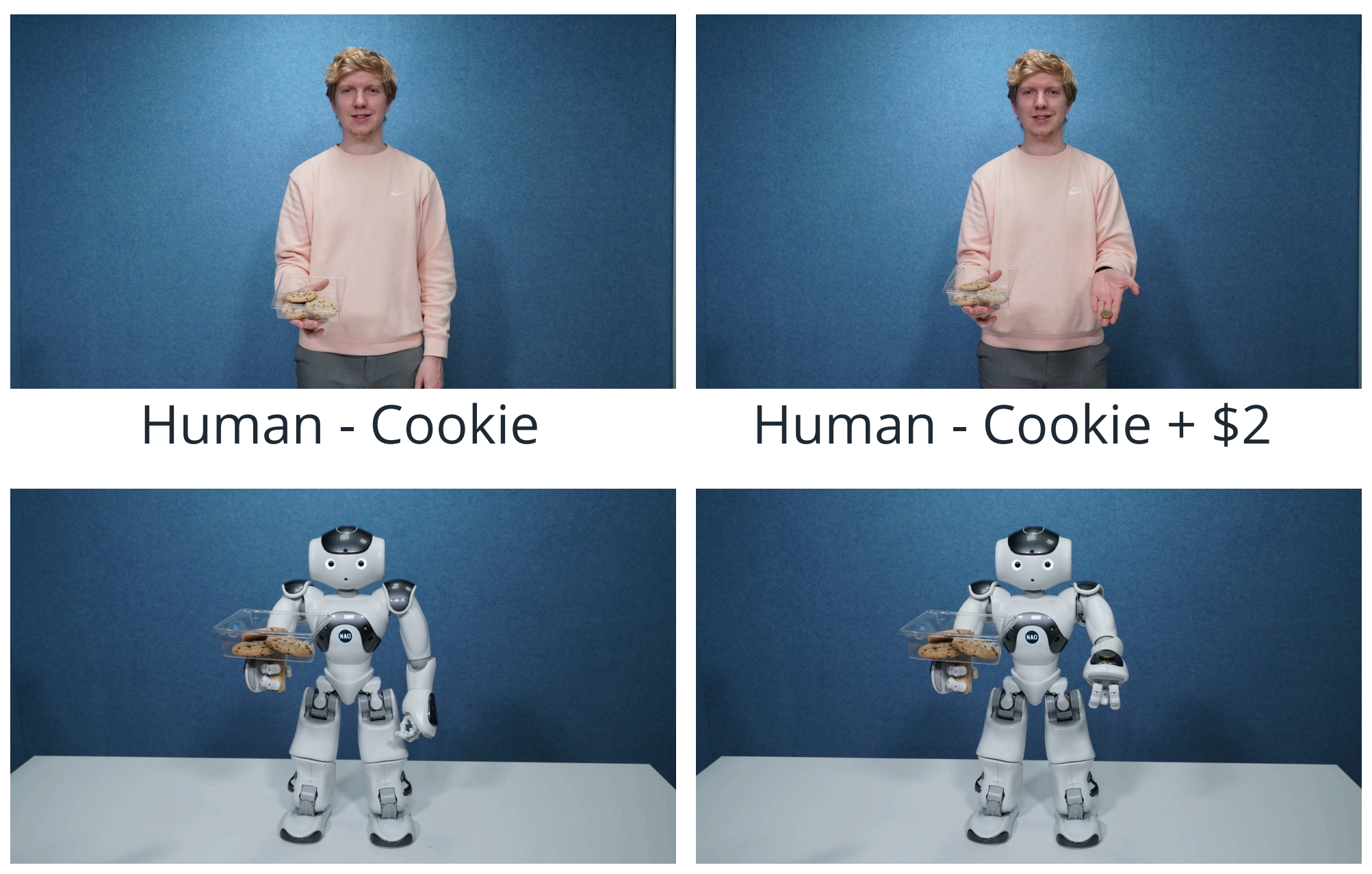
Introduction

- In HHI, overly generous offers without clear justification elicit **phantom costs** (e.g., hidden motives, risk) and **reduce acceptance** [1].
  - Cookie with or without $2, justifying they were eating cookies with friends.
- Extended to HRI [2-3]
  - However: **implausible for robots**
  - People show epistemic vigilance toward robot-provided information [4]
- **Perceived plausibility**: how believable an explanation is given the agent communicating it.

Objectives and Predictions

1. Replicate the phantom costs effect in HRI.
2. Extend the paradigm by testing how perceived plausibility shapes: Phantom costs, Offer acceptance, Trust in the agent.
   - Implausibility: ↑ phantom costs (H1), ↓ offer acceptance (H2) and ↓ trust in the agent (H3).
   - H4: Stronger effect on moral than performance trust.

Methodology *N = 292; online*



Human - Cookie

Human - Cookie + $2

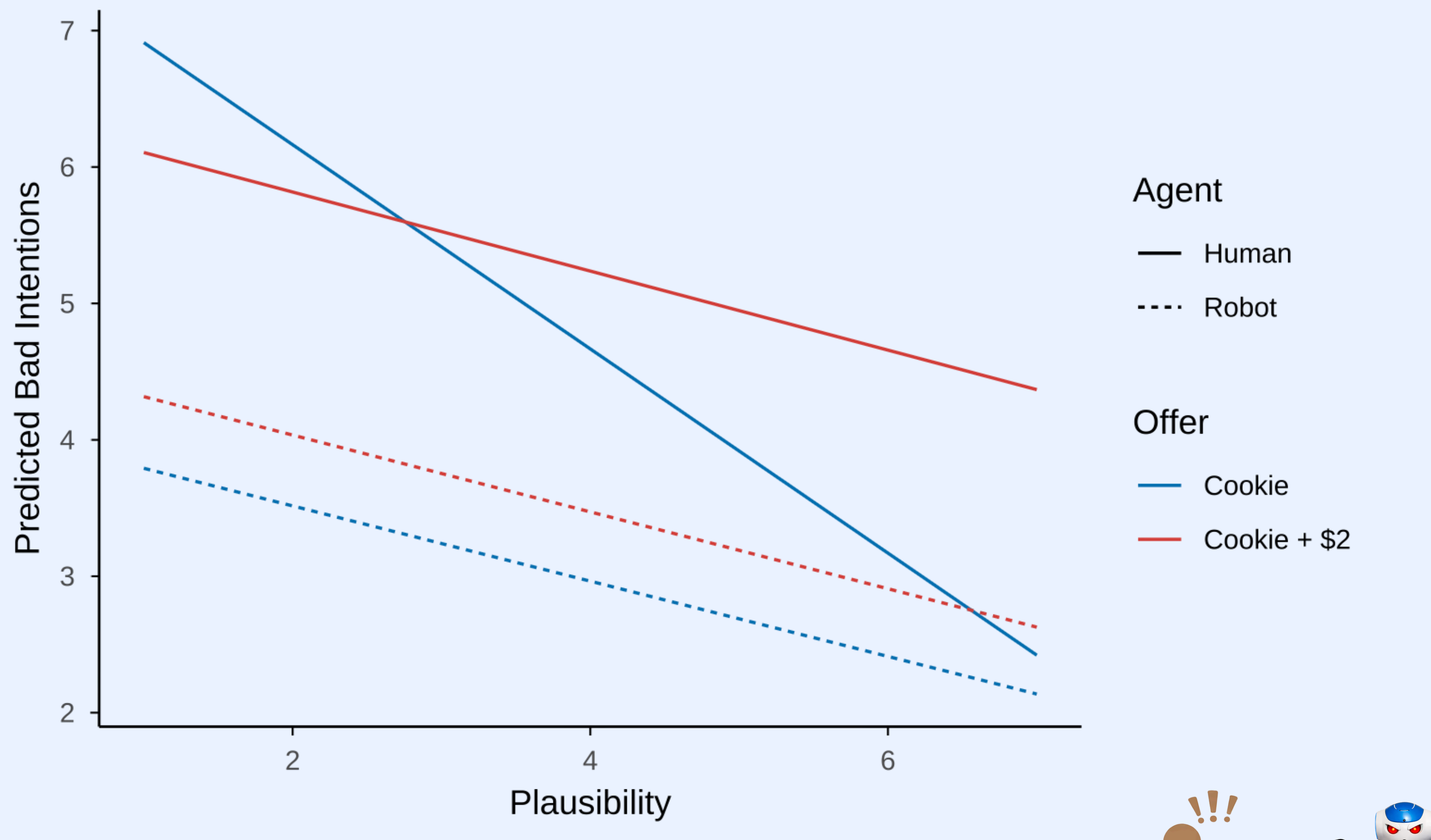
Robot - Cookie

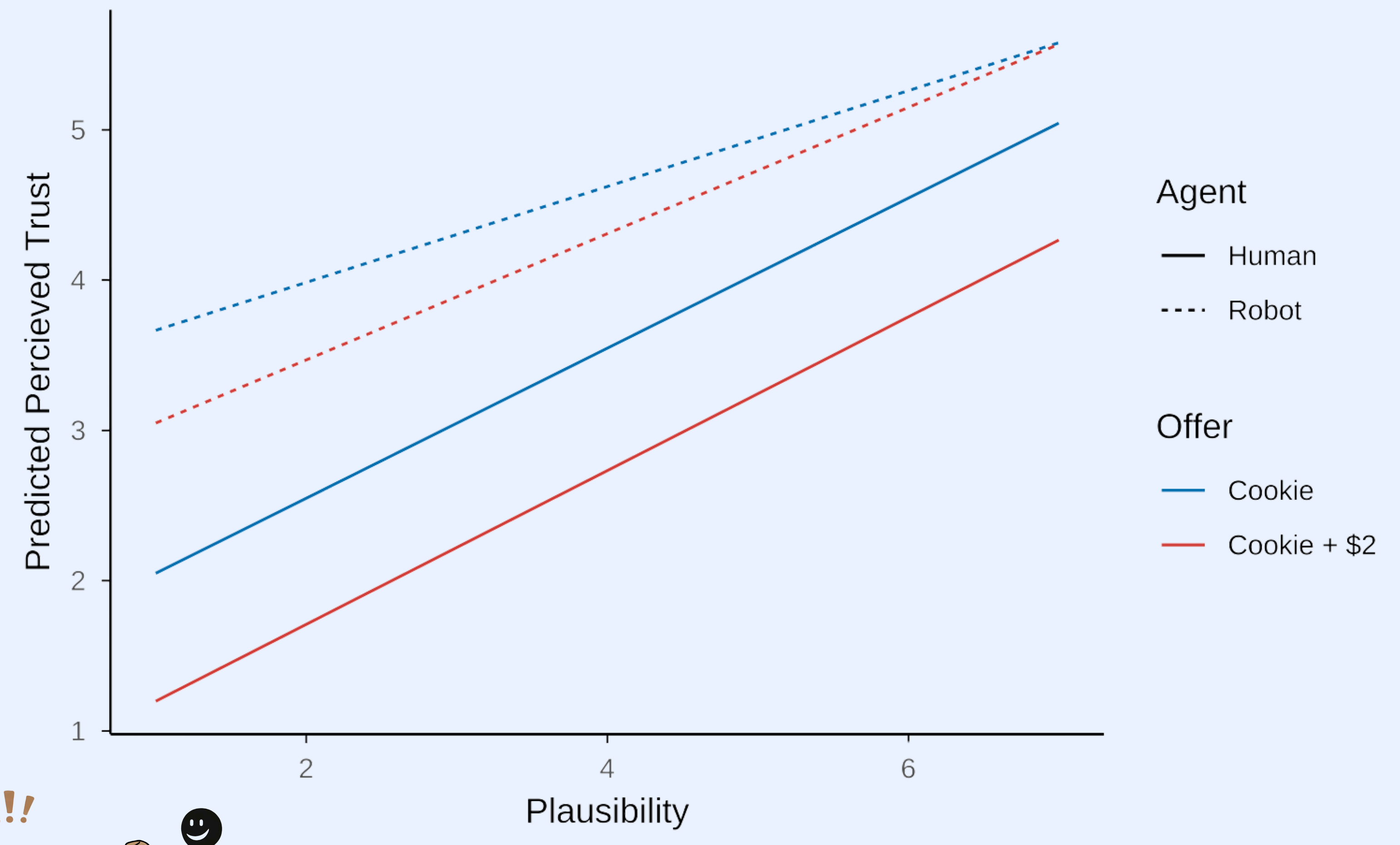
Robot - Cookie + $2

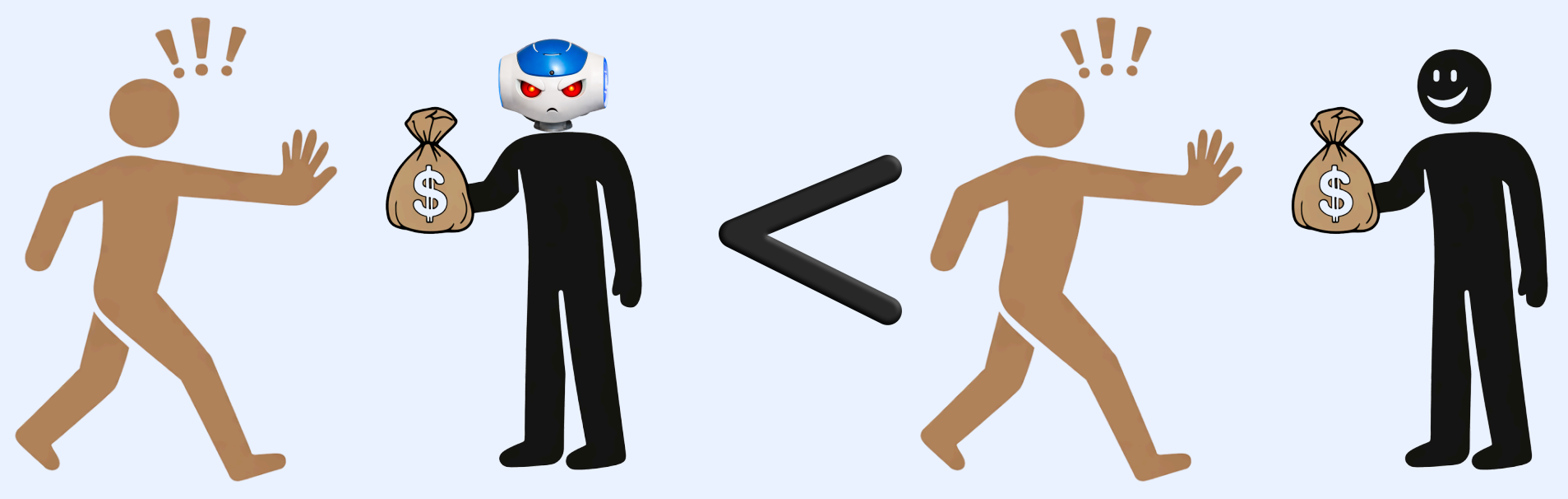
Reprinted with permission from [3], © IEEE 2025

Mesures

- **Decision-making** (accept/reject the offer)
- **Justification** (text-entry box)
- **Phantom costs** (Likert scales)
- **Perceived plausibility**: "It is plausible that a [agent] might have just eaten cookies with friends" (Likert scale)
- **Trust** (shortened MDMT [5])
  - Moral: benevolent, has goodwill; genuine, sincere; ethical, moral.
  - Performance: competent; reliable.







➔ Plausibility effects did not differ between moral and performance trust.

Conclusion

- **Replication of phantom costs in HRI.**
- **Perceived plausibility of an agent's explanation**
  - **decreases phantom costs**
  - **increases acceptance and trustworthiness.**

To reduce phantom costs and promote effective HRI, robot explanations should be:

- **sufficient (to justify the action)**
- **plausible (given the characteristics of the robot)**

[1] A. Vonasch, R. Mofradidoost, and K. 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

[2] 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

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

[4] 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

[5] 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