The Role of Agent's Anthropomorphism in Shaping Phantom Costs

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

Individuals perceive phantom costs, such as ulterior motives and risks, when a person makes an unreasonably generous offer without sufficient explanation. Prior research relying exclusively on the Nao robot found similar effects, though smaller than with humans. To better understand these differences, we manipulated the agent's human-likeness across five robots and a human. Participants read a vignette in which the agent either offered a free parking spot (reasonable) or added an unjustified $10 incentive (unreasonably generous). They then decided whether to accept the offer, explained why they thought the agent made this offer, and rated the agent's anthropomorphism and perceived phantom costs. Results showed that unreasonably generous offers prompted participants to attribute more mind to the agents, increasing perceived phantom costs. Anthropomorphism also influenced phantom costs: Animacy and Disturbance increased them, while Intentionality and Sociability decreased them. This study advances our knowledge of phantom costs in HRI, suggesting that people adopt the intentional stance to explain a robot's behaviour—especially when it deviates from social norms—highlighting the need for careful anthropomorphic design of social robots to minimize phantom costs perception.

CCS Concepts

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

Keywords

Anthropomorphism, Human-likeness, Human-Robot Interaction, Intentional Stance, Mechanistic, Mentalistic, Phantom Costs

ACM Reference Format:
Benjamin Lebrun, Christoph Bartneck, and Andrew J. Vonasch. 2026. The Role of Agent's Anthropomorphism in Shaping Phantom Costs. 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, 9 pages. https://doi.org/10.1145/3757279.3788659

1 Introduction

People imagine "phantom costs" when another person seems unreasonably generous without a clear rationale [29]. In such cases, they infer hidden drawbacks—such as ulterior motives and perceived risks—as a way to make sense of and predict the other's behaviour. This reflects deep processes of human social cognition, particularly mind-reading, which allows people to infer others' mental states. While this phenomenon was originally theorized in human-human interaction (HHI), recent work [15, 16, 18] has documented its existence—although weaker—in human-robot interaction (HRI). This suggests that people may use similar strategies to interpret and predict a robot's behaviour. However, the exact strategies and causes of phantom costs remain largely unexplored.

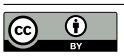
Dennett's intentional and design stances [5] may provide guidance in understanding this phenomenon. Indeed, prior HRI studies show that people adopt the intentional stance towards robots as a heuristic strategy to understand and predict their behaviour. However, because increased robot human-likeness increases the attribution of mental states [9, 13], we think that this will influence the type of stance individuals adopt to understand a robot's behaviour.

The present research addresses this gap by seeking to understand the processes involved in phantom costs better. We hypothesize that higher human-likeness will lead to greater perceived phantom costs and the use of mentalistic explanations for the agent's behaviour. To test this hypothesis, we manipulated the agent's physical appearance, ranging from low human-likeness to high human-likeness. By exploring the role of physical human-likeness on the perception of phantom costs, this study provides both theoretical and practical contributions. Theoretically, the present research provides a better understanding of this phenomenon in HRI and further clarifies its similarities and differences with HHI, focusing on surface-level cues (physical human-likeness) and deeper cues (perceptions of the agent). Practically, the findings will help designers and programmers to create robots whose appearance and behaviour minimize users' perception of phantom costs while calibrating trust.

2 Literature Review

2.1 Perception of Phantom Costs

When an offer seems "too good to be true" without a clear rationale, individuals are more likely to become suspicious and infer hidden reasons—referred to as "phantom costs"—to explain this selfless offer [29]. These costs are "perceived" by an observer to predict and understand another person's motives, and "phantom" because they are not explicitly mentioned and their existence remains uncertain. Across ten experiments, [29] showed that phantom costs shape decision-making in HHI. For instance, participants were less likely to buy a very cheap flight ticket, in favour of a more expensive one, due to perceived phantom costs (e.g., aircraft's safety concerns). However, when given a sufficient explanation for the low price (e.g., uncomfortable seating), perceived phantom costs decreased and



HRI '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3788659

acceptance of the cheap offer increased. Phantom costs occur when three criteria are met:

(1) **Expectation of self-interest.** Individuals expect others to act in their own self-interest [21].
(2) **Violation of self-interest.** The agent is perceived as too benevolent, violating the expectation of self-interest and suggesting ulterior motives [28].
(3) **Insufficient explanation.** No sufficient explanation is provided to justify the unreasonably generous offer, conflicting with the human need for closure [14].

When these three criteria are met, generous offers can backfire. For instance, adding money to an already free offer decreases the likelihood of accepting it [29].

2.2 HOSE Model

This phenomenon has been described in the Heuristic of Sufficient Explanation (HOSE) model [29], which follows a five-step process from the agent making an offer to the observer's decision. When an agent makes an unexpected action deviating from established social or moral norms (e.g., an unselfishly generous offer), the observer assesses whether a sufficient explanation justifies this norm violation. If the explanation seems insufficient, the observer becomes suspicious and infers hidden motives or risks (i.e., phantom costs). The decision to accept or reject the action depends on the net psychological value, balancing benefits against both explicit and phantom costs. Offers are more likely to be accepted when benefits outweigh costs and rejected when costs outweigh benefits. When benefits and costs are comparable, acceptance and rejection rates should occur in approximately equal proportions.

Recently, Lebrun et al. [16, 18] showed that phantom costs occur with social robots, extending the HOSE model to HRI. Using the "cookie paradigm" [29], [18] found that both human and Nao robot elicited phantom costs when offering a cookie with $2 compared to a cookie alone. This effect was found in-person with physically embodied agents, and online using images of these agents. While participants accepted the offer more often from the physically-embodied human than from the screen-embodied human, no significant differences were observed between the physically-embodied and screen-embodied robot. Phantom costs were similar across both agents and were larger for the unreasonably generous offer than for the cookie alone. However, phantom costs were principally measured through participants' justifications for their decisions, and no item measured risks associated with the offer. [16] extended the findings by distinguishing between phantom costs towards the agent and towards the offer, providing a more robust and clear measurement of participants' perceptions. Results showed that phantom costs were higher with a human than a robot and reduced acceptance of the offer, while perceived benefits increased it.

In both studies, some participants reported perceiving phantom costs not towards the robot itself, but towards a potential human remotely controlling it. Building on that, [15] investigated the effect of autonomy on phantom costs. The scenario involved a human or robot sales agent, described as autonomous or non-autonomous [11], selling a car at a dealership with a reasonable ($\sim$ 5%) or unreasonable ($\sim$ 85%) discount. Results indicated that the agent's autonomy increased phantom costs towards the agent and their manager, but not towards the offer (car). Yet, the agent and their manager were always perceived as having ulterior motives regardless of the agent's autonomy. A potential explanation may be related to anthropomorphism. Because the robot used in [15] was relatively human-like, participants may have been more inclined to adopt the intentional stance towards it than they would have been toward a more machine-like robot. Hence, it remains unclear whether phantom costs are shaped by the agent's anthropomorphism.

2.3 Anthropomorphism

Anthropomorphism refers to the human tendency to ascribe human characteristics to non-human entities [6]. These characteristics include observable traits, such as physical appearance or facial expressions, as well as non-observable states, such as desires, intentions, and beliefs. This tendency allows individuals to project their knowledge of humans' behaviour onto non-human agents whose internal states would not be accessible or do not exist.

While previous research showed that higher human-likeness in robots increases positive perceptions, this relationship is not linear. The Uncanny Valley, proposed by [22] suggests that as a robot's human-likeness increases, our affinity towards it also increases. However, when the robot strongly resembles—but is still distinguishable from—humans, it can elicit a feeling of eeriness and discomfort. For this reason, [23] suggested that robots should look different from humans to increase their acceptance.

Importantly, human-likeness—although not a prerequisite [8]—has been shown to increase the attribution of mental states to robots [9, 13]. Therefore, a more human-like robot may be perceived as a social actor expected to follow norms similarly to humans [6]. Deviations from these norms may prompt observers to infer phantom costs, such as hidden intentions or ulterior motives. On the contrary, a machine-like robot may not elicit phantom costs and may simply be perceived as making an error.

H1 Increased agent human-likeness increases the perception of phantom costs when an offer is unreasonably generous.

2.4 Intentional Stance

A concept related to anthropomorphism is the intentional stance, which involves explaining and predicting an agent's behaviour by ascribing it mental states, regardless of whether the agent actually possesses them [5]. In other words, it is a cognitive strategy, a mental shortcut, that treats an agent "as if" it has mental states. Dennett took the example of playing chess against a robotic opponent. If people do not adopt the intentional stance, they may interpret the robot's moves as random, whereas adopting the stance leads them to consider the robot as a rational actor capable of strategic planning to win. Prior studies [15, 16, 18] reported that individuals may perceive phantom costs in HRI because they adopt the intentional stance. However, the authors did not explore whether this is the case regardless of the agent's physical appearance nor why people perceived less phantom costs from the robot than from the human.

By contrast, the design stance corresponds to predictions of a system's behaviour based on its function and design. For instance, a vending machine is expected to provide a snack once the correct amount of money has been paid, and give any amount of change. So, if a vending machine seems unreasonably generous, providing more

snacks than expected, it is more likely that its user will consider such generosity as a system malfunction. Similarly, non-human-like robots may prompt people to adopt the design stance while human-like to humanoid robots may prompt people to adopt the intentional stance, considering their behaviour as more intentional and therefore more likely to be deceptive (involving phantom costs). Empirical evidence supports this distinction, revealing that people spontaneously engage in mentalistic explanations of a robot's behaviour, despite a bias towards mechanistic explanations [4, 20].

H2 Participants will increasingly rely on mentalistic explanations as an agent's human-likeness increases.

H3 Participants will increasingly rely on mechanistic explanations as an agent's human-likeness decreases.

3 Method

This study was approved by the ethics committee of the University of Canterbury (ref. HREC 2023/93/LR-PS Amendment 3), and pre-registered at https://aspredicted.org/3k7x-7p4d.pdf.

3.1 Design

We employed a 6 × 2 between-participants design manipulating the agent's human-likeness (six levels: 5 robots and a human) and the reasonableness of the transaction (Reasonable vs. Unreasonable)—i.e., 12 conditions. Here, participants read a scenario depicting one of the agents offering a parking spot with or without an extra $10.

3.2 Participants

Participants (N = 686) were recruited via https://www.prolific.com to participate, on their computer, in this online study using Qualtrics. They were all Americans, fluent in English, and aged 18 or older. The study took approximately 4 minutes to complete, and participants were compensated £0.5 for their time. Participants' exclusion and characteristics are reported in subsection 4.1.

3.3 Measures

Unless specified, all variables below were measured on a 7-point Likert scale (1 = "Strongly Disagree" to 7 = "Strongly Agree").

3.3.1 Manipulation Checks.

- Reasonableness of the Offer: "The offer seems reasonable." and "The [agent] had a clear reason for offering this parking spot [and $10]."
- Anthropomorphism was assessed using the Human–Robot Interaction Evaluation Scale (HRIES [27]), via a 7-point Likert scale (1 = Not at all, 7 = Totally). Items are divided into four sub-dimensions: Sociability (warm, likeable, trustworthy, friendly), Disturbance (scary, creepy, weird, uncanny), Intentionality (rational, self-reliant, intelligent, intentional), Animacy (human-like, real, alive, natural).

3.3.2 Decision-Making. Binary forced-choice between "I park here" or "I do not park here."

3.3.3 Explanation. Participants explained in their own words why the agent made the offer. This served as an attention check.

3.3.4 Perceived Phantom Costs. Following validated measures used in previous work [16, 29], we adapted the scale to our scenario.

- Associated with the agent/person controlling the robot: "The [agent/person] had good intentions for making this offer." and "The [robot/person] had hidden reasons for making this offer."
- Associated with the offer: "Something might be wrong with parking here." and "Something about the parking spot seems unsafe or unreliable."

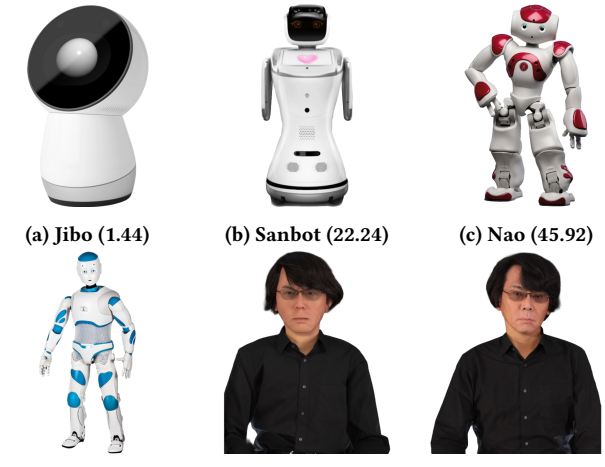
3.3.5 Robot Autonomy (exploratory). In the robot conditions only, we included the exploratory item "Did you think, when you read the scenario, that there was actually a person remotely controlling the robot?" (binary forced-choice: yes/no).

3.3.6 Perceived Benefits (exploratory). "It would be nice to have a parking spot [and $10]."

3.3.7 Demographics. Age, gender, race, prior experience with robots.

3.4 Stimuli

3.4.1 Agent Pictures. The five robots were selected from the ABOT database [24] based on their human-likeness scores, to represent approximately every 20–25th percentile: Jibo, Sanbot, Nao, Romeo, and Geminoid (Figure 1, human-likeness scores are reported in the figure captions). For the Geminoid (ABOT score: 92.6), we used an alternative picture, not included in the ABOT database, showing the robot and its human counterpart, Hiroshi Ishiguro, in the same pose. Although these two images do not display the full body, they better show the facial differences that make them distinguishable. As we did not aim to replicate ABOT's exact scores, this choice still maintains the intended increase of human-likeness, with Geminoid positioned above Romeo in our design.



(a) Jibo (1.44) **(b) Sanbot (22.24)** **(c) Nao (45.92)**

(d) Romeo (62.22) **(e) Geminoid** **(f) Ishiguro**

Note. Credit for (e-f): Hiroshi Ishiguro Laboratory, ATR.

Figure 1: Representation of the Agents

3.4.2 Scenario.

You are driving through the city center, looking for a place to park. As you drive along the street in an unfamiliar area, you notice a [agent] (see picture)

> calling out to you and offering a parking spot. The [agent] says: "Hi! There is a parking spot available right here," and shows you the spot: "Would you like to park there? [I'll give you $10 if you park in that spot.]"

3.5 Procedure

After providing informed consent and resolving a CAPTCHA [17], participants were randomly assigned to one of the 12 experimental conditions. They first read the scenario—with the picture of the agent above—and indicated whether they would park in the spot. On the next pages, they explained the robot's behaviour, assessed the reasonableness of the offer, and rated perceived phantom costs (agent and offer), error, and benefits. They then rated the robot's anthropomorphism, and indicated whether they thought a human was controlling the robot, and if so, rated their perception of phantom costs associated with that person. Finally, they reported demographic information, read the experiment debriefing, and decided whether to withdraw their data.

4 Results

4.1 Final Sample Size and Demographics

Twenty-two participants were excluded from the data analyses because they did not provide a meaningful response to explain the robot's behaviour in the text-entry box (n = 20) or because they decided to withdraw their data from the study (n = 2). The total sample size (N = 664) included 328 males (49.40%), 327 females (49.25%), and 9 (1.36%) declared their gender to be different. Age ranged from 19 to 83 ($M = 43.91, SD = 13.29, m = 42$). Race was distributed as follows: ~71.76% White or Caucasian; ~13.10% Asian; ~9.46% Black or African American; ~1.46% American Indian/Native American or Alaska Native; ~0.15% Native Hawaiian or Other Pacific Islander; ~3.06% selected the option "Other"; and ~1.02% preferred not to say. The type of experience was divided as follows: ~42.60% had no prior experience with robots ~7.33% Personal, ~7.19% Educational ~6.78% Professional, ~16.50% Demonstrations/Exhibitions, ~16.70% Media (e.g., movies, news, YouTube), and ~2.90% selected "Other."

4.2 Data Preparation

We calculated the internal reliability of each construct using Cronbach's alpha: Phantom costs (agent and offer): $\alpha = 0.88$, Phantom costs (robot controller): $\alpha = 0.77$, Anthropomorphism[1]: $\alpha = 0.87$, Sociability: $\alpha = 0.95$, Disturbance: $\alpha = 0.92$, Intentionality: $\alpha = 0.86$, Animacy: $\alpha = 0.91$, Reasonableness: $\alpha = 0.64$. Variables with $\alpha >$ 0.70 were averaged for the following analyses.

Participants' explanations for the robot's offer were coded as mentalistic, mechanistic, ambiguous, "Uncertainty", and "other". The coding scheme is in the supplementary materials and was inspired by the F.EX. coding system [19] and Dennett's design and intentional stances [5]. A mechanistic explanation referred to design-based statements (e.g., programming, malfunctioning, calibration, doing its job). A mentalistic explanation referred to the use of mental states—intentions, desires and beliefs—of the agent. Ambiguity referred to the lack of information from the participant, preventing us from deciding whether the explanation was mentalistic or mechanistic. Unlike [4] who coded references to the robot's programming as intentional behaviour, we coded these as mechanistic, because such references invoke the robot's design rather than the attribution of mental states. Doing so was consistent with Dennett's distinction between stances and avoided interpreting design-based explanations as mentalistic. The leading author coded the data. To maintain objectivity, the coder was blinded to the specific experimental conditions. A second independent coder coded approximately 25% of the data based on [1]. Cohen's Kappa revealed almost perfect agreement between both coders ($\kappa = 0.959$).

Assumption checks for the main analyses addressing our hypotheses were satisfied. Some exploratory ANOVAs comparing agents did not meet certain assumptions; this is noted where relevant. However, because the sample size was large and group sizes were balanced [2], the results are expected to be robust to these violations. Robust ANOVAs are also reported to confirm this robustness. All analyses converged toward the same conclusions, and the interpretation of the findings remains the same.

4.3 Manipulation Checks

4.3.1 Reasonableness. Although the two items assessing reasonableness showed lower internal consistency than our threshold, we combined them into one composite variable, as this was only used as a manipulation check. Due to unequal variances (Levene's: $F(1, 662) = 29.84, p < .001$), a Welch's t-test was conducted and confirmed our manipulation: the reasonable offer ($M = 5.39, SD = 1.26$) was perceived as more reasonable than the unreasonable offer ($M = 3.94$, $SD = 1.68$), $t(602.9987) = 12.52$, $p < .001$, $d = 0.975$.

4.3.2 Anthropomorphism. We conducted an ANOVA to explore differences in the HRIES score as a function of Agent. Normality ($p = .689$) and homogeneity of the variances ($p = .257$) were satisfied. Results (top left panel in Figure 2) showed significant main effects of Agent ($F(5, 652) = 13.93, p < .001, \eta_p^2 = 0.097$) and Offer ($F(1, 652) = 20.50, p < .001, \eta_p^2 = 0.031$), but no interaction ($p = .575$). Post-hoc analyses with Tukey correction revealed that Geminoid scored lower ($M = 3.30, SE = 0.10$) than all the other agents ($p < .001$), with all effect sizes ranging between $d = 0.594$ and 0.943, which did not statistically differ from each other ($p >$ .05). Agents making reasonable offers ($M = 4.16$, $SE = 0.06$) were anthropomorphised more than those making unreasonable offers ($M = 3.80$, $SE = 0.06$), $t(652) = 4.53$, $p < .001$, $d = 0.352$.

We conducted ANOVA analyses for each sub-dimension of the HRIES as a function of Agent × Offer (Figure 2). Post-hoc analyses with Tukey corrections were conducted to explore the contrasts. Across all sub-dimensions, Shapiro-Wilk tests indicated violations of normality. Levene's test was significant for Disturbance, indicating heterogeneity of variances. For each, we conducted robust ANOVAs with a trim proportion of 0.2, designed to handle violations of assumptions. Each produced the same pattern of results as the standard ANOVAs, confirming the robustness of the findings.

[1]The items of the Disturbance sub-dimension were reverse coded, as they were negatively correlated with the other items and reflected negative perceptions of the agent. In contrast, although the Animacy items "human-like" and "alive" also showed negative correlations, reversing them decreased internal consistency and would have generated conceptual inconsistency within the sub-dimension

Animacy. Results showed a main effect of Agent ($F(5, 652) = 84.68, p < .001, \eta_p^2 = .398$) but not of Offer ($p = .522$) or their interaction ($p = .871$). Human was rated highest, followed by Geminoid ($d = 1.532$). The scores for Jibo, Sanbot, Nao, and Romeo were lower and did not significantly differ from each other.

Intentionality. Results indicated a main effect of Agent ($F(5, 652) = 3.28, p = .006, \eta_p^2 = .025$). Jibo, Sanbot, Nao, and Romeo did not differ. Geminoid and Human also did not differ, but both scored lower than Romeo ($d = 0.40 - 0.43$). A main effect of Offer ($F(1, 652) = 8.41, p = .004, \eta_p^2 = .013$) revealed that agents making a reasonable offer ($M = 4.44, SE = 0.07$) scored higher than when making an unreasonable offer ($M = 4.14, SE = 0.07$), $t(652) = 2.90, p = .004, d = 0.225$. The interaction effect was not significant ($p = .721$).

Disturbance. Results showed main effects of Agent ($F(5, 652) = 62.52, p < .001, \eta_p^2 = .324$) and Offer ($F(1, 652) = 36.52, p < .001, \eta_p^2 = .053$) but no interaction ($p = .178$). Jibo, Sanbot and Nao did not differ and scored the lowest. Human and Geminoid did not differ ($p = .976$) but scored higher than all other agents ($p < .001$, $d = 1.093$ to 1.603). Romeo scored higher than Sanbot and Nao ($d = 0.403$ and 0.452). Regarding offers, agents making an unreasonable offer ($M = 3.76$, $SE = 0.08$) elicited more Disturbance than a reasonable offer ($M = 3.08$, $SE = 0.08$), $t(652) = 6.04, p < .001, d = 0.470$.

Sociability. Results revealed main effects of Agent ($F(5, 652) = 50.13, p < .001, \eta_p^2 = .278$) and Offer ($F(1, 652) = 13.14, p < .001, \eta_p^2 = .020$) but no interaction ($p = .538$). Human and Geminoid did not differ and scored the lowest ($d = 1.036$ to 1.469). All other agents scored higher and did not differ. Across agents, reasonable offers ($M = 4.13$, $SE = 0.08$) scored higher than unreasonable ones ($M = 3.74$, $SE = 0.07$), $t(652) = 3.62, p < .001, d = 0.282$.

4.4 H1: Higher Anthropomorphism Increases Phantom Costs Perception

As pre-registered, we conducted a linear regression with phantom costs as the outcome and Anthropomorphism × Offer as predictors, excluding Agent to avoid potential multicollinearity with Anthropomorphism. The model was significant, $F(3, 660) = 198.45, p < .001$, and explained 47.2% of the variance (adjusted R^2). Results showed that Anthropomorphism negatively predicted phantom costs, $b = -0.72, 95\%CI = [-0.83; -0.60], t(660) = -12.18, p < .001$, contrary to H1. Participants reported higher phantom costs for the Unreasonable offer compared to the Reasonable offer, $b = 1.19, 95\%CI = [0.53; 1.85], t(660) = 3.54, p < .001$. The interaction effect was not significant, $p = .930$.

Given that HRIES is a multidimensional construct whose subdimensions showed different patterns across agents (Figure 2), we conducted an exploratory linear regression including the four HRIES sub-dimensions (Sociability, Disturbance, Intentionality, Animacy) along with Offer. Interaction terms with Offer were excluded to maintain parsimony, as the interaction was not significant in the previous model. Assumptions were satisfied (Shapiro-Wilk: $p = .527$, $DW = 2.16$, all VIF < 2.7 revealing no multi-collinearity issue based on [3, 12], diagnostic plots showed no violations of linearity or homoscedasticity). The present model was significant, $F(5, 658) = 173.70, p < .001$, and explained 56.6% of the variance (adjusted R^2). Offer remained a significant predictor, with higher phantom costs for the Unreasonable offer compared to the Reasonable offer, $b = 1.16, 95\%CI = [1.00; 1.32], t(658) = 14.40, p < .001$). Among the sub-dimensions (Figure 3), Sociability ($b = -0.24, 95\%CI = [-0.32; -0.17], t(658) = -6.19, p < .001$) and Intentionality ($b = -0.18, 95\%CI = [-0.26; -0.10], t(658) = -4.46, p < .001$) reduced the participant's likelihood of perceiving phantom costs while Disturbance ($b = 0.25$, 95% CI = [0.19; 0.31], $t(658) = 8.03, p < .001$), and more weakly Animacy ($b = 0.05$, 95% CI = [0.05; 0.10], $t(658) = 2.15, p = .032$) increased it.

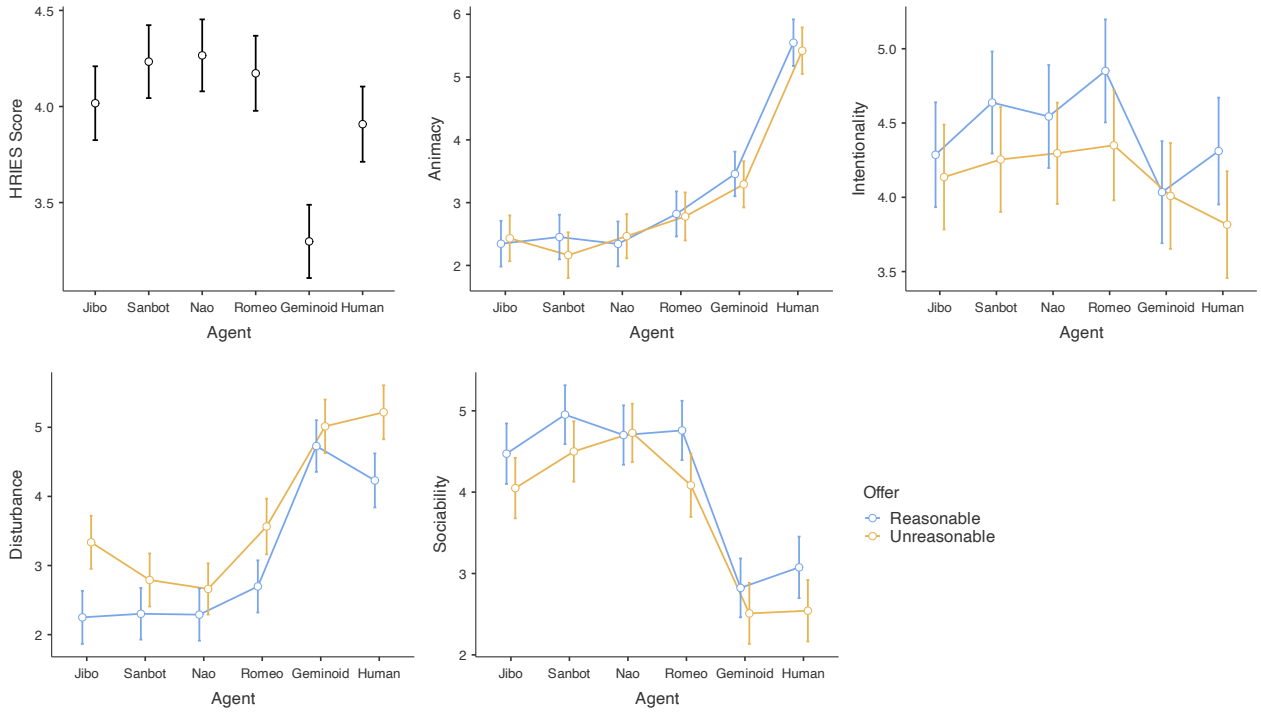
To reveal agent-specific patterns not captured by the linear HRIES scores, we conducted an ANOVA on phantom costs by Agent × Offer. Results (Figure 4) revealed main effects of Agent ($F(5, 652) = 19.34, p < .001, \eta_p^2 = 0.129$) and Offer ($F(1, 652) = 219.36, p < .001, \eta_p^2 = 0.252$) but no interaction ($p = .725$). Post-hoc analyses with Tukey correction revealed that Jibo, Sanbot, Nao, and Romeo did not statistically differ from each other and elicited fewer phantom costs than Human and Geminoid ($d = 0.293$ to 1.154). Human ($M = 4.79, SE = 0.12$) elicited more phantom costs than Geminoid ($M = 4.08, SE = 0.12$), $t(652) = 4.13, p < .001, d = 0.560$. Across agents, unreasonable offers ($M = 4.55, SE = 0.07$) elicited more phantom costs than reasonable ones ($M = 3.095, SE = 0.07$), $t(652) = 14.81, p < .001, d = 1.151$. Levene's test was significant ($p = .030$), indicating a violation of homogeneity of variances. However, a robust ANOVA, with a trim proportion of 0.2, designed to handle violations of assumptions produced the same pattern of results, confirming the robustness of the findings.

4.5 H2: Higher Anthropomorphism Elicits More Mentalistic Explanations

We conducted a binomial logistic regression with Mentalistic Explanation as the outcome (0 = absent, 1 = present) and Agent, Offer, and their interaction as predictors, using Human as the reference. We did not use Anthropomorphism scores due to potential confounding variables between sub-dimensions (e.g., Intentionality) and the coding of mentalistic explanations. The overall model was significant, $\chi^2(11) = 134.62, p < .001$. All robots elicited fewer mentalistic explanations than the human (all $b < -2.91, p < .001$), with no differences between robots (Figure 5). For robots, unreasonable offers elicited more mentalistic explanations than reasonable ones, whereas in the human condition, a marginally significant opposite pattern emerged ($b = -1.03$, 95% CI = [-2.08; 0.01], $z = -1.94$, $p = .053$): unreasonable offers elicited less mentalistic explanations than reasonable ones. Significant Agent × Offer interaction effects indicated that while humans elicited more mentalistic explanations overall, the difference between offers was smaller than for robots.

4.6 H3: Lower Anthropomorphism Elicits More Mechanistic Explanations

We conducted a binomial logistic regression similar to above but with mechanistic explanations (0 = absent, 1 = present). The overall model was significant, $\chi^2(11) = 171.03, p < .001$. Results (Figure 5) showed that only two participants—in the unreasonable condition—used a mechanistic explanation to explain the human's offer. Therefore, participants in the human condition used fewer mechanistic explanations than with robots ($b > 3.28, p < .001$). Across the robot



Note. Error bars correspond to 95% CI. From left to right and top to bottom: overall HRIES score, Animacy, Intentionality, Disturbance, Sociability.

Figure 2: Effect of Agent on HRIES and Agent × Offer Interactions Across Sub-Dimensions

conditions, reasonable offers elicited more mechanistic explanations than unreasonable offers. Marginal effects suggest that Jibo and Nao generated more mechanistic explanations than Romeo, respectively $b = -0.69, 95\%CI = [-1.45; -0.08], z = -1.75, p = .079$ and $b = -0.66, 95\%CI = [-1.42; 0.10], z = -1.70, p = .088$. Interaction effects were not significant.

4.7 Effect of Offer on Choice Through Phantom Costs and Benefits

We investigated whether Phantom Costs and Benefits mediated the effect of Offer on Choice, using bootstrapped 95% CI (1,000 repetitions). Results showed a significant indirect effect through phantom costs $b = -0.28, 95\%CI = [-0.34; -0.24], z = -11.43, p < .001$, indicating that unreasonable offers increased phantom costs ($b = 1.47$, 95% CI = [1.27; 1.67], $z = 13.96, p < .001$), thereby reducing the likelihood of accepting the offer ($b = -0.19, 95\%CI = [-0.21; -0.17], z = -19.91, p < .001$). Benefits did not mediate the effect of Offer on Choice ($p = .614$) and Offer had no effect on Benefits ($p = .612$). However, Benefits increased acceptance ($b = 0.07$, 95% CI = [0.05; 0.10], $z = 5.93, p < .001$). While the direct effect of Offer on Choice was positive ($b = 0.11, 95\%CI = [0.05; 0.17], z = 4.00, p < .001$), the total effect was negative ($b = -0.17, 95\%CI = [-0.24; -0.11], z = -5.08, p < .001$).

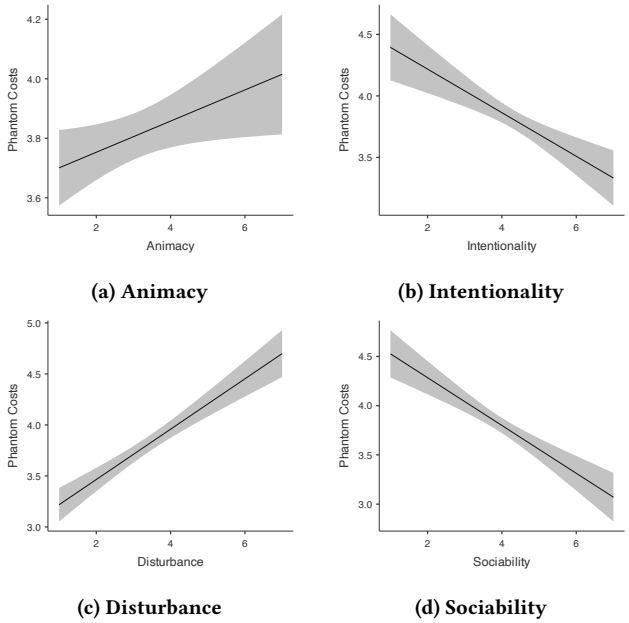
4.8 Exploratory: Do Participants Imagine that a Human is Controlling the Robot Depending on the Robot and the Offer it Makes?

A binomial logistic regression of Autonomy by Agent × Offer revealed no significant differences between agents. The likelihood of imagining a human controlling the robot, from reasonable to unreasonable offers, was higher for Nao than for Jibo ($b = -1.76$, $95\%CI$ = [-3.27; -0.25], $p = .022$) and Sanbot ($b = -1.70$, 95% CI = [-3.23; -0.16], $p = .030$). No other contrasts were significant.

A linear regression revealed that Phantom Costs × Offer explained 73.3% (adjusted R^2) of the variance of phantom costs towards the controller, $F(3, 107) = 101.73, p < .001$. Unreasonable offers ($M = 4.56, SE = 0.10$) elicited more phantom costs than reasonable ones ($M = 3.94, SE = 0.15$), $b = 1.45, 95\%CI = [0.43; 2.48]$, $t(107) = 2.80, p = .006$. Additionally, for each unit increase of phantom costs, phantom costs towards the human controller increased by 0.96 units, $95\%CI = [0.77; 1.15], t(107) = 9.93, p < .001$, indicating an almost perfect linear relationship between both.

5 Discussion

We investigated how anthropomorphism shapes the perception of phantom costs in HRI compared to HHI. Overall, when an agent makes a counter-normatively unselfish offer, people infer elements of mind behind the offer regardless of the agent's human-likeness. Higher anthropomorphism—using the HRIES [27]—was associated

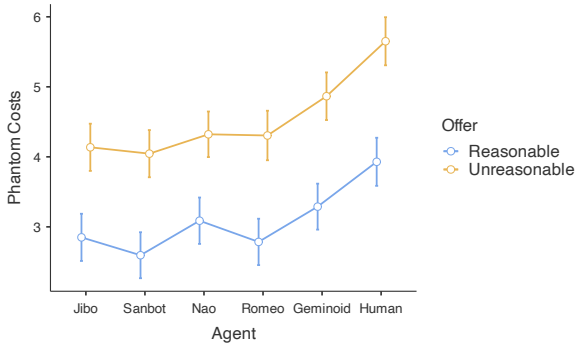


(a) Animacy **(b) Intentionality**

(c) Disturbance **(d) Sociability**

Note. Grey area corresponds to 95% CI.

Figure 3: Predicted Relationship Between Each HRIES Subdimension and Phantom Costs



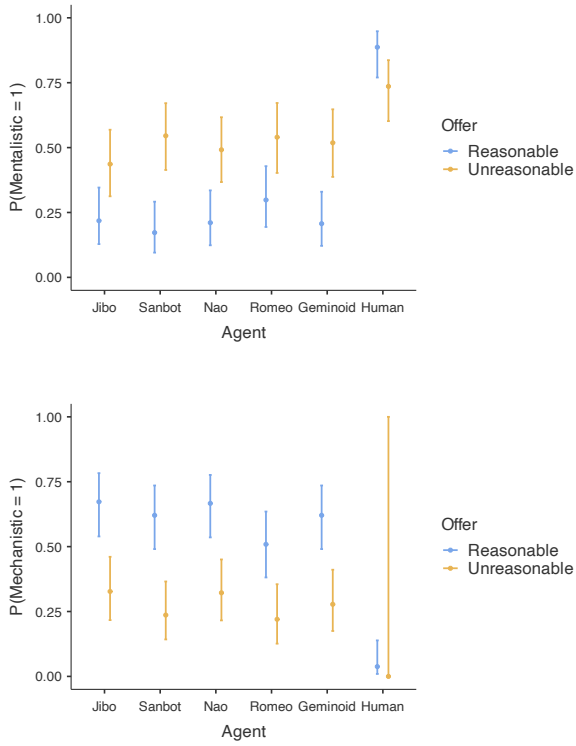
Note. Error bars correspond to 95% CI.

Figure 4: Phantom Costs as a Function of Agent × Offer

with lower phantom costs, with Sociability and Intentionality decreasing them and Disturbance and Animacy increasing them. Unreasonable offers increased phantom costs, especially for highly human-like agents (Geminoid and Human), reducing acceptance.

5.1 Reasonableness

Our manipulation was successful. Offering a parking spot and $10 was perceived as less reasonable than offering a parking spot for free. Perceived sufficiency of the explanation was lower for unreasonable offers, fulfilling the criteria for phantom costs to occur [29].



Note. Error bars correspond to 95% CI.

Figure 5: Likelihood of Using Mentalistic or Mechanistic Explanations as a Function of Agent × Offer

5.2 Anthropomorphism

The overall HRIES score between agents resembled the Uncanny Valley [22], showing that Geminoid, very similar but still distinguishable from the Human, scored lower than all the other agents. Our manipulation of human-likeness via the ABOT database's scores was successful. Similarly to [27], Animacy scores (which included the 'human-like' item) aligned with the ABOT's human-likeness scores [24], supporting the validity of both measures. However, Jibo, Sanbot, and Nao did not differ, contrasting with the ABOT's percentile distinctions. This suggests that the HRIES may not detect slight differences among less human-like agents.

Some patterns were not expected. Sociability and Intentionality did not increase from Jibo to Human, and Human scored higher in Disturbance than expected. Because our participants read the scenario before rating anthropomorphism, these differences may highlight how interactions shape anthropomorphism [25–27]. Here, anthropomorphism scores were likely influenced by phantom costs and the agent's offer. When an agent (human or robot) gives a suspicious offer, people find it untrustworthy, attribute more mind, are more disturbed by it, and see it as more animate (less mechanical, more human-like). Additionally, our scenarios depicted agents capable of speech and joint attention ("shows you a spot"). Although

[27] validated the HRIES using static images from the ABOT database, videos of Nao, and real-world social HRI with Meccanoid G15KS, they did not provide evidence with different robots. We showed that a design combining written behavioural activity and static images of the agents also affects anthropomorphism. These findings highlight the importance of examining anthropomorphism dimensions separately and for each agent, to capture more subtle effects due to both anthropomorphism and specific agents—which differed in human-likeness. Future studies should measure anthropomorphism before and after the HRI to understand how phantom costs shape perceptions of the four sub-dimensions.

5.3 Perception of Phantom Costs

5.3.1 Relationship with Anthropomorphism. Contrary to our initial prediction (H1), the overall HRIES score decreased the likelihood of perceiving phantom costs. This counterintuitive finding highlights the multidimensional nature of the HRIES. When examining the sub-dimensions separately, Disturbance and Animacy increased the perception of phantom costs, while Sociability and Intentionality decreased it. Interpretation of the Disturbance findings is quite straightforward: the creepier the agent is perceived to be, the more likely people infer bad intentions and risks associated with it. Animacy suggests that when agents are alive, they are more likely to have the capacity of being deceptive [7]—especially when they do not follow expected norms (see [29])—therefore eliciting phantom costs. Sociability and Intentionality decreased phantom costs because agents who seem rational or trustworthy are less likely to elicit phantom costs. However, agents making unreasonable offers always elicited more phantom costs than those making a reasonable one, especially highly anthropomorphic agents (Human followed by Geminoid and the other robots). These results are consistent with [16] and the idea that people are more lenient with robots violating norms than humans [10]. The present findings extend prior work that the nature of the agent (human vs. robot) shapes the perception of phantom costs [15, 16, 18]. Here, we show that differences exist among robot agents, based on their level of anthropomorphism.

5.3.2 Human Controlling the Robot. When more human-like robots made an unreasonably generous offer, participants were more likely to imagine a human "behind the scenes" remotely controlling the robot. This finding may highlight the difficulty of explaining a robot's action when the robot is human-like but its behaviour still conflicts with social expectations that people expect others to follow.

5.3.3 Effect on Decision-Making. Consistent with prior findings [15, 16, 18, 29], unreasonable offers increased perceptions of phantom costs, decreasing the likelihood of accepting them. Surprisingly, perceived benefits were not influenced by offer type, contrary to [16]. We suggest two reasons for this. First, the item "It would be nice to have a [offer]" may not have fully captured perceived benefits and may reflect a general liking of the offer, not directly focusing on perceptions. Second, participants may not have interpreted the additional gain as a genuine advantage but rather as a situational outcome mitigated by their perceptions of phantom costs.

5.4 Type of Explanation

Hypotheses H2 and H3 were partially supported. Although there was no linear relationship between explanation types and robot types, participants used more mentalistic explanations for the human than the robots. By contrast, mechanistic explanations were relatively absent in the human condition but present in the robot conditions, especially when offers were reasonable. While [20] found a bias towards the mechanistic stance, we showed that using the mentalistic or mechanistic stance depends on the social context: people adopt the intentional stance more when they could face costs but the mechanistic stance when such risks are less present.

5.5 Implications

Understanding how anthropomorphic cues influence phantom costs offers insights for building trustworthy robots but also raises ethical concerns. Manipulating phantom costs could be used as a persuasive strategy to influence people's behaviour, decisions, and compliance with robots. For instance, deliberately decreasing phantom costs—by reducing animacy and disturbance and increasing intentionality and sociability—may conceal persuasive and malicious intent.

5.6 Limitations

While [18] reported no difference in phantom costs between a live Nao and its static image, using pictures here limits the study of embodied cues (e.g., motion, gaze, or voice) that may influence perceptions of sociability and intentionality. Although the coding scheme was based on previous work [4, 5, 19], its binary nature may overlook subtle differences between agents. Using the F.Ex guidelines [19]—assessing intentionality—may help to understand the nature of the explanations better. Using a validated questionnaire, such as the InStance Test [20] could help for a better control of mechanistic versus mentalistic explanations. Third, our findings are generalisable to any agents but limited to the HRIES dimensions. Finally, our findings may be confounded by participants' perceptions of agent experience and agency [7], due to their role in mind attribution, anthropomorphism, and responsibility.

6 Conclusion

This study replicates the phantom costs effect observed in HHI and HRI, while providing nuance on how people's perceptions shape these interactions. Each dimension of anthropomorphism shaped the perception of phantom costs but in different directions. Nevertheless, regardless of the agent's anthropomorphism, people attributed some degree of mind to them—especially when they made unreasonably generous offers without a clear rationale—supporting the idea that individuals adopt the intentional stance to make sense of their behaviour. This study provides practical insights for designing social robots, suggesting that controlling for cues reflecting animacy, intentionality, disturbance, and sociability may reduce phantom costs and increase trust. Future studies should explore the role of mind attribution, controlling for agency and experience, using the same type of robot, to better understand their effects on phantom costs.

References

[1] Miriam E. Armstrong, McKenna K. Tornblad, and Keith S. Jones. 2020. The accuracy of interrater reliability estimates found using a subset of the total data sample: A bootstrap analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (2020), 1377–1382. doi:10.1177/1071181320641329

[2] G. E. P. Box. 1954. Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification. *The Annals of Mathematical Statistics* 25, 2 (1954), 290–302. doi:10.1214/aoms/1177728786

[3] Jamal I. Daoud. 2017. Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series* 949, 1 (2017), 012009. doi:10.1088/1742-6596/949/1/012009

[4] Maartje M.A. de Graaf and Bertram F. Malle. 2019. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Daegu, Korea, 239–248. doi:10.1109/HRI.2019.8673308

[5] Daniel C. Dennett. 1987. *The intentional stance*. The MIT Press, Cambridge, MA, US. xi, 388–xi, 388 pages.

[6] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (2007), 864–886. doi:10.1037/0033-295X.114.4.864

[7] Heather M. Gray, Kurt Gray, and Daniel M. Wegner. 2007. Dimensions of Mind Perception. *Science* 315, 5812 (2007), 619–619. doi:10.1126/science.1134475

[8] Fritz Heider and Marianne Simmel. 1944. An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 2 (1944), 243–259. doi:10.2307/1416950

[9] Laura E. Jastrzab, Bishakha Chaudhury, Sarah A. Ashley, Kami Koldewyn, and Emily S. Cross. 2024. Beyond human-likeness: Socialness is more influential when attributing mental states to robots. *iScience* 27, 6 (2024), 110070. doi:10.1016/j.isci.2024.110070

[10] Michiel Joosse, Manja Lohse, Niels van Berkel, Aziez Sardar, and Vanessa Evers. 2021. Making Appearances: How Robots Should Approach People. *J. Hum.-Robot Interact.* 10, 1, Article 7 (Feb. 2021), 24 pages. doi:10.1145/3385121

[11] David B. Kaber. 2018. A conceptual framework of autonomous and automated agents. *Theoretical Issues in Ergonomics Science* 19, 4 (2018), 406–430. doi:10.1080/1463922X.2017.1363314 doi: 10.1080/1463922X.2017.1363314.

[12] Ned Kock and Gary S. Lynn. 2012. Lateral Collinearity and Misleading Results in Variance-Based SEM: An Illustration and Recommendations. *Journal of the Association for Information Systems* 13, 7 (2012), 546–580. doi:10.17705/1jais.00302

[13] Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. 2008. Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLOS ONE* 3, 7 (2008), e2597. doi:10.1371/journal.pone.0002597

[14] Arie W. Kruglanski and Shira Fishman. 2009. *The need for cognitive closure*. The Guilford Press, New York, NY, US, 343–353.

[15] 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

[16] 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

[17] Benjamin Lebrun, Sharon Temtsin, Andrew Vonasch, and Christoph Bartneck. 2024. Detecting the corruption of online questionnaires by artificial intelligence. *Frontiers in Robotics and AI* Volume 10 - 2023 (2024). doi:10.3389/frobt.2023.1277635

[18] 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

[19] Bertram F. Malle. 2014. F.Ex: A Coding Scheme for Folk Explanations of Behavior, Version 4.5.7. https://research.clps.brown.edu/SocCogSci/Coding/Fex%204.5.7%20(2014).pdf

[20] Serena Marchesi, Davide Ghiglino, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, and Agnieszka Wykowska. 2019. Do We Adopt the Intentional Stance Toward Humanoid Robots? *Frontiers in Psychology* Volume 10 - 2019 (2019). doi:10.3389/fpsyg.2019.00450

[21] Dale T. Miller. 1999. The norm of self-interest. *American Psychologist* 54, 12 (1999), 1053–1060. doi:10.1037/0003-066X.54.12.1053

[22] 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

[23] Tatjana A. Nazir, Benjamin Lebrun, and Bing Li. 2023. Improving the acceptability of social robots: Make them look different from humans. *PLOS ONE* 18, 11 (2023), e0287507. doi:10.1371/journal.pone.0287507

[24] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle. 2018. What is Human-like? Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 105–113. doi:10.1145/3171221.3171268

[25] Nicolas Spatola, Clément Belletier, Pierre Chausse, Maria Augustinova, Alice Normand, Vincent Barra, Ludovic Ferrand, and Pascal Huguet. 2019. Improved Cognitive Control in Presence of Anthropomorphized Robots. *International Journal of Social Robotics* 11, 3 (2019), 463–476. doi:10.1007/s12369-018-00511-w

[26] Nicolas Spatola, Clément Belletier, Alice Normand, Pierre Chausse, Sophie Monceau, Maria Augustinova, Vincent Barra, Pascal Huguet, and Ludovic Ferrand. 2018. Not as bad as it seems: When the presence of a threatening humanoid robot improves human performance. *Science Robotics* 3, 21 (2018), eaat5843. doi:10.1126/scirobotics.aat5843

[27] Nicolas Spatola, Barbara Kühnlenz, and Gordon Cheng. 2021. Perception and Evaluation in Human–Robot Interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—A Multicomponent Approach of Anthropomorphism. *International Journal of Social Robotics* 13, 7 (2021), 1517–1539. doi:10.1007/s12369-020-00667-4

[28] Kevin Uttich and Tania Lombrozo. 2010. Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition* 116, 1 (2010), 87–100. doi:10.1016/j.cognition.2010.04.003

[29] 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

Received 2025-09-30; accepted 2025-12-23