Differences in Social Expectations About Robot Signals and Human Signals

Lorenzo Parenti, Marwen Belkaid, Agnieszka Wykowska

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13393

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13393

Differences in Social Expectations About Robot Signals and Human Signals

Lorenzo Parenti,[a,b] Marwen Belkaid,[a,c] Agnieszka Wykowska[a]

[a]*Social Cognition in Human-Robot Interaction, Istituto Italiano di Tecnologia (IIT)*
[b]*Department of Psychology, University of Turin*
[c]*ETIS UMR 8051, CY Cergy Paris Université, ENSEA, CNRS*

Received 11 May 2023; received in revised form 22 November 2023; accepted 27 November 2023

Abstract

In our daily lives, we are continually involved in decision-making situations, many of which take place in the context of social interaction. Despite the ubiquity of such situations, there remains a gap in our understanding of how decision-making unfolds in social contexts, and how communicative signals, such as social cues and feedback, impact the choices we make. Interestingly, there is a new social context to which humans are recently increasingly more frequently exposed—social interaction with not only other humans but also artificial agents, such as robots or avatars. Given these new technological developments, it is of great interest to address the question of whether—and in what way—social signals exhibited by non-human agents influence decision-making. The present study aimed to examine whether robot non-verbal communicative behavior has an effect on human decision-making. To this end, we implemented a two-alternative-choice task where participants were to guess which of two presented cups was covering a ball. This game was an adaptation of a "Shell Game." A robot avatar acted as a game partner producing social cues and feedback. We manipulated robot's cues (pointing toward one of the cups) before the participant's decision and the robot's feedback ("thumb up" or no feedback) after the decision. We found that participants were slower (compared to other conditions) when cues were mostly invalid and the robot reacted positively to wins. We argue that this was due to the incongruence of the signals (cue vs. feedback), and thus violation of expectations. In sum, our findings show that incongruence in pre- and post-decision social signals from a robot significantly

Correspondence should be sent to Agnieszka Wykowska, Istituto Italiano di Tecnologia, Center for Human Technologies, Via Enrico Melen, 83, 16152 Genoa, Italy. E-mail: agnieszka.wykowska@iit.it

influences task performance, highlighting the importance of understanding expectations toward social robots for effective human–robot interactions.

Keywords: Social decision-making; Human–robot interaction; Non-verbal communication; Social cues; Response time

1. Introduction

Humans make most of their decisions in social contexts, where those decisions can be influenced by others, and affect others' choices and actions (Sanfey, 2007). In this context, the presence and interpretation of non-verbal signals are critical components of decision-making in social contexts (Burgoon, Guerrero, & Floyd, 2016; Diederich, Brendel, Morana, & Kolbe, 2022; Feine, Gnewuch, Morana, & Maedche, 2019; Knapp & Harrison, 1972). Interestingly, a substantial body of literature has highlighted the significance of non-verbal cues not only in human–human interaction but also in interactions between humans and robots (Abubshait & Wiese, 2017; Admoni & Scassellati, 2017; Burgoon et al., 2010; Knapp, Hall, & Horgan, 2013; Leathers, 1976; Eaves & Leathers, 2017; Palinko, Rea, Sandini, & Sciutti, 2016). In the contemporary era, we interact socially not only with other humans but also with artificial agents such as robots. Therefore, the question of how these agents, through their non-verbal signals, influences human decision-making remains a critical area that calls for investigation.

People naturally tend to ascribe anthropomorphic attributes to robots, especially those with human-like shapes, and expect them to behave in a socially intelligent way (Hortsmann & Kramer, 2020; Perez-Osorio, Marchesi, Ghiglino, Ince, & Wykowska, 2019; Spatola, Marchesi, & Wykowska, 2022). Therefore, a significant effort in social robotics has been dedicated to the design of social non-verbal behaviors that can facilitate human–robot interaction (HRI). For instance, the robot gaze has been shown to play an important role in handover (Abubshait et al., 2023; Moon et al., 2014) and joint attention tasks (Abubshait, Momen, & Wiese, 2020; Boucher et al., 2012; Wiese, Wykowska, Zwickel, & Müller, 2012). Studies have found that eye contact increases user engagement (Ito, Hayakawa, & Terada, 2004; Kompatsiari, Bossi, & Wykowska, 2021; Kompatsiari, Ciardo, Tikhanoff, Metta, & Wykowska, 2019; Kompatsiari, Ciardo, & Wykowska, 2022; Szafir & Mutlu, 2012) and the attribution of intentionality to the robot (Ciardo, De Tommaso, & Wykowska, 2022; Ito et al., 2004; Lombardi et al., 2023). Moreover, robots' behavior regarding proxemics (i.e., the study of space and distance in social situations) has an effect on how humans perceive their social presence (Fiore et al., 2013). These studies illustrate how non-verbal communication plays a role in HRI and suggest that people may in some cases perceive robots' communicative behaviors as social signals (Wiese, Metta, & Wykowska, 2017; Wykowska, 2021).

In the context of decision-making, it is important to consider that social signals may be received either before or after a decision is made (Parenti et al., 2021a). In the former case, a non-verbal signal can serve as a cue or advice and inform predictions about possible decision outcomes to help select an action. In the latter case, a reaction or feedback from a social partner contributes to updating one's internal representations of the decision problem and the

related variables. Several studies examined the effects of robot signals when displayed prior to participants' decisions. For instance, robot gaze/head orientation and gestures have been used to provide cues aimed to inform or influence human choices in various HRI settings (Chidambaram, Chiang, & Mutlu, 2012; Ghazali, Ham, Barakova, & Markopoulos, 2018; Perez-Osorio, Abubshait, & Wykowska, 2021; Romat, Williams, Wang, Johnston, & Bard, 2016). Moreover, eye contact with a robot prior to decision-making has been shown to delay decisions, affect neural activity, and influence strategies in a competitive game (Belkaid, Kompatsiari, De Tommaso, Zablith, & Wykowska, 2021). In contrast, less is known about social signals provided by robots after decisions. Robot verbal feedback has been shown to influence participants' choices (Ham & Midden, 2014). Nevertheless, robot feedback could also be provided through non-verbal behavior, like facial expressions, gaze behavior, and gestures (Ciardo & Wykowska, 2022; Gonsior et al., 2011; Parenti, Lukomski, De Tommaso, Belkaid, & Wykowska, 2023). Post-decision feedback is likely to influence upcoming decisions, and thus interact with pre-decision signals. Therefore, as the number of social signals emitted by the robot increases, the question of how humans interpret them and integrate them in their decision process becomes more complex.

On the one hand, during interactions with robots, people could rely on social expectations grounded in experience with human–human interaction (Edwards, Edwards, Westerman, & Spence, 2019). Such social expectations, and the resulting interaction patterns, are thought to play a major role in guiding interpersonal communication. Indeed, extensive research in social psychology highlighted the importance of prior expectations based on experiences and social norms in anticipating others' behavior (Burgoon, 1993). Concurrently, it has been suggested that people tend to interact with machines using the same interaction patterns they developed to interact with other humans (Edwards et al., 2019) and similar cognitive mechanisms that have developed for interacting with other humans (Parenti et al., 2023; Wykowska, 2020). On the other hand, people's social expectations about robots' behavior (Kahn et al., 2011) might to some extent differ from expectations regarding human behavior. Such expectations may be related to high-level cognitive, social, and emotional capabilities, but also to lower-level properties. For instance, robots may be expected to exhibit mechanistic movements rather than smooth motion trajectories and how people perceive the robot may depend on whether these expectations are met (Ghiglino, Willemse, De Tommaso, & Wykowska, 2021; Parenti, Marchesi, Belkaid, & Wykowska, 2021b). As suggested by previous studies, robots' anticipated behavior can be influenced by a variety of factors such as their appearance, social setting, prior experience, and familiarity (Kahn et al., 2011; Kwon, Jung, & Knepper, 2016).

In social psychology, the "Expectancy Violation" theory proposes to consider social interaction according to the extent to which interpersonal experiences deviate positively or negatively from one's expectations (Burgoon, 1993). We argue that this is also particularly relevant in the context of HRIs. Indeed, studies suggest that humans have high expectations about what robots are able to achieve and about their interaction skills (Horstmann & Kramer, 2019, 2020). Yet, because the current state of the art in robotics does not match these expectations, negative violation of expectation is more likely to occur (Horstmann & Kramer, 2019; Kwon et al., 2016; Marchesi, Bossi, Ghiglino, De Tommaso, & Wykowska, 2021). This can

negatively affect HRIs. For instance, negative violation of expectations has been linked to detrimental communication outcomes and uncertainty in the interaction (Bartholow, Fabiani, Gratton, & Bettencourt, 2001; Mendes, Blascovich, Hunter, Lickel, & Jost, 2007). This can result in a variety of cognitive and behavioral responses, including increased arousal and negative affect (Burgoon, 1993; Burgoon & Hale, 1988). These effects can be particularly pronounced in situations where the violation is unexpected or particularly salient (Proulx, Sleegers, & Tritt, 2017). Violation of expectations can have negative effects on performance. This can be due to the cognitive load and attentional resources required to process the unexpected stimulus or the violation of the learned sequence (Browining & Harmer, 2012; Ferdinand, Mecklinger, & Opitz, 2015).

This paper aimed to address the question of how expectations regarding social signals of an artificial agent affect human decision-making processes. More specifically, we focused on the (in)congruency between signals that can be interpreted as cues toward which decision to take and those that are delivered as feedback regarding the decision taken. This question is quite relevant for the field, as we are developing artificial agents and social robots that are endowed with more and more complex repertoire of social behaviors. Thus, it is important to understand the relationship between those various behaviors and their impact on human cognition, decision-making specifically.

To address the aims of our study, we implemented a two-alternative-choice task (cups-and-ball game) online with a between-subjects design where participants were asked to make a decision regarding which, out of two presented simultaneously cups, contains a ball hidden beneath it. Importantly, we manipulated the robot's cue before the decision (directional pointing toward one of the cups) and the robot's feedback ("thumb up," or no feedback) after the decision. Our goal was (i) to examine the effects of these social signals on participants' decision processes and the potential interaction between pre- and post-decision signals, and (ii) to assess whether the observed effects would be similar to when the signals were delivered by a human. Cue validity was manipulated such that the robot indicated the correct cup in 80% (high validity) or 20% (low validity) of the trials. Social feedback was manipulated such that the robot displayed either a positive reaction (social feedback) or no reaction (no social feedback) to successful guesses (Experiment 1). Because positive feedback could signal cooperation, we hypothesized that participants would follow the robot cue more frequently in the presence of positive feedback (H1). We also expected participants to be slower in the condition with low validity and social feedback (H2). This is because we reasoned that participants would expect pre- and post-decision social signals to be congruent and that reacting positively to successful guesses after providing an invalid cue (incongruent scenario) would violate such social expectations. To further test whether the observed effect was (i) indeed related to social expectations (rather than other, non-social processes) and (ii) similar to what we could observe with human social signals, two additional experiments were conducted. In Experiment 2, the cue was given by a flashlight—to control for lower-level attentional mechanisms, while in Experiment 3, the robot was replaced by a human—to compare across agent types. We hypothesized that the robot condition would be similar to the human condition and different from the flashlight condition (H3), as the robot had a human-like appearance and thus would be likely perceived as a social agent.

2. Methods

Since three experiments were designed to test our hypotheses, in the following section, we report methods and materials that are common to the three experiments and then describe the aspects that are specific to each experiment. Experiments were implemented in PsychoPy3 (v2020.1.3; Peirce et al., 2019) and were made available to users via a link in Pavlovia.org and then embedded in Prolific.com call to participation.

2.1. Participants

For all three experiments, we recruited 400 participants in total, online through Prolific (www.prolific.com), of which 396 were included in the final analysis (age: 26.8 ± 8.4, m/f: 243/153, student status y/n: 208/188), 50 for each of the eight experimental conditions (four conditions in Experiment 1, two conditions in Experiment 2, and two conditions in Experiment 3). We ran an a priori power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) for mixed Analysis of Variance (ANOVA) using medium effect size ($f = 0.25$), an alpha of 0.05, power set to 0.95. The power analysis suggested a total sample of 176 participants for Experiment 1, meaning 44 participants for each of the four conditions. We rounded it to 50 participants per group in Experiment 1 to account for potential dropouts and maintained this sample size for consistency in Experiments 2 and 3. Four participants were excluded from the final analyses because they did not complete the entire task. Prolific provides participants with an allocated time for completing an online experiment, which is typically set at three times the estimated duration of the experiment. When participants do not complete the task within the allotted time, Prolific automatically closes the session and registers participation based on their guidelines. We identified that four participants began the task but did not actively complete it, and the session closed after the time limit, in adherence to Prolific participation rules. Inclusion criteria were right-handedness, age between 18 and 64, fluency in English, and normal or corrected-to-normal vision. Experiments were conducted online from January 2021 to May 2021, with participants drawn from the Prolific participant pool. To prevent multiple participation across studies, we excluded participants with matching Prolific Identification Codes (IDs) from previous studies in this line. The eight experimental conditions were run sequentially, ensuring participant independence. To prevent multiple participation across the different conditions (between-subjects) of this study, we excluded participants with matching Prolific IDs from previous conditions of this experiment. Participants were paid £6.5 for their participation in the experiment, and the task took 30 minutes on average to be completed. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria) and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki, 1964) and Prolific policies (prolific.com).

2.2. Design and procedure

2.2.1. Task

The task (cups and ball game) was loosely inspired by the Shell Game (Britannica Encyclopedia, 2023). The "cups-and-balls" experiment, while sharing similarities with the task

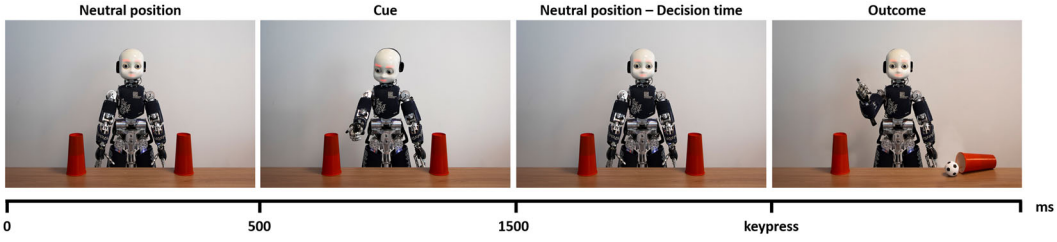

Fig. 1. Trial structure. The cue from the robot represents the pre-decision signal and the feedback (thumb up) represents the post-decision signal once the actual ball position is revealed (outcome).

used in Parenti et al. (2023), represents a more engaging adaptation of a simple gambling task previously employed in HRI studies (e.g., see Abubshait et al., 2021). We selected this task, as it addresses decision-making processes in ambiguous situations, where one needs to "bet" on an option. Since such an "ambiguous" context elicits a large degree of uncertainty during decision-making, we deemed it a well-suited game for observing the influence of social signals (and social partners) on the decision-making processes. This kind of paradigm represents a real-life situation when one needs to make decisions under uncertainty (e.g., choosing between two menu items at a restaurant, or choosing one type of perfume over another as a gift for a friend). In such situations, due to a high degree of ambiguity and uncertain outcome, we might be prone to being influenced by others, through their explicit (verbal) recommendations but also often by means of implicit social signals.

We presented the participants with a sequence of photos, in which two cups were located on a table and an agent was positioned on the other side of the table game. Instructions stated that the ball could change position on each trial and that the agent (a robot in Experiment 1, a flash in Experiment 2, or a human in Experiment 3) would be guessing together with the participant where the ball was. The robot we used in this study is iCub, a humanoid robot designed to serve as a social robot (Metta, Sandini, Vernon, Natale, & Nori, 2008). We asked participants to be as accurate as possible in finding the correct location of the ball, without time constraints. The game started with the agent looking at the participant, the agent was then always hinting with a social cue toward a cup and then returning to the starting position (see Fig. 1). The cue consisted of the robot pointing toward one or the other cup by using its arm and hand gestures (see the second panel in Fig. 1). The frame with the agent starting with a neutral position lasted 500 ms, the cue lasted 1000 ms. After that, the agent went back in the neutral position and the participant was able to choose the "left" or "right" cup. No time constraint was set for the participant's decision time. After the participant's choice, the ball position was revealed by lifting the chosen cup. The agent could give positive feedback ("thumb up") to the participant's hit or remain neutral. The agent's feedback lasted 1500 ms (see Fig. 1). Participants were instructed that at the end of each trial, the ball would be automatically reshuffled randomly under one of the two cups even if the shuffle was not visually happening on screen. Each participant completed 100 trials and a final debriefing.

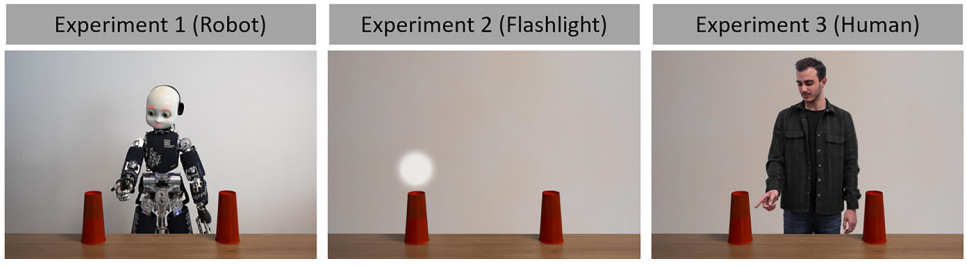

Fig. 2. Examples of the different experiments. Here, we present the cue stimuli for the three different experimental conditions.

2.2.2. *Experiment 1*

Experiment 1 was composed of four conditions following a between-subject design. For each condition, we manipulated the validity of iCub cues (80% vs. 20%) and the possibility to receive a positive feedback to "hits" from the robot (feedback present vs. feedback absent). We recruited 50 participants for each of the four experimental conditions for a total of 200 participants recruited in Experiment 1.

2.2.3. *Experiment 2*

Experiment 2 was composed of two conditions (feedback present vs. feedback absent) in a between-subjects design. We focused only on the 20% validity condition. The experiment involved a flashlight, instead of the robot, as the initial cue for each trial. As in Experiment 1, in one of the two conditions the robot was giving participants a positive feedback after they hit the right cup. We recruited 50 participants for each of the two experimental conditions for a total of 100 participants recruited in Experiment 2.

2.2.4. *Experiment 3*

Experiment 3 was composed of two conditions (feedback present vs. feedback absent) in which the virtual partner was a human. As in Experiment 2, we focused only on the 20% validity condition. In Experiment 3, the human, rather than a robot or a flashlight, gave cues to the participants and in one condition showed a positive feedback after participants' hits. We recruited 50 participants for each of the two experimental conditions for a total of 100 participants recruited in Experiment 3.

2.3. *Stimuli*

Stimuli consisted of an agent's picture (e.g., robot or human in Experiments 1 and 3) or a flashlight (Experiment 2) always in the same setting: behind a table with two red cups on top of it (see Fig. 2). In Experiment 1, the robot was turning its head, gazing, and pointing toward a specific cup to produce the cue. In Experiment 2, a light appearing just above one cup or the other produced the cue. In Experiment 3, the human agent was mimicking the robot cue from Experiment 1. The social feedback displayed after the decision was implemented by making

the robot (Experiments 1 and 2) or the human (Experiment 3) produce a thumbs up (with always the same arm) and a positive facial expression (smiling).

Pictures and data used in the three studies can be found in the public repository under the anonymized link: https://osf.io/97ex8/?view_only = 1a53da66cc5b423b85fc2dbf73d17394.

2.4. Data analysis

Data analysis was performed using RStudio (RStudio Team, 2020) and was conducted on demographic information and behavioral data from the task. Response times (RTs) slower than 2.5 standard deviations from the sample mean were considered outliers and excluded for each experimental condition. RTs faster than 100 ms were considered outliers and were excluded (Ratcliff, 1993). We did not apply any transformation on the RTs, given that the majority of the observations were distributed between 0 and 1 s (Ratcliff, 1993). Participants' accuracy rates were calculated based on the percentage of hits in the task (hits on 100 trials). The rate of following the robot's cue (hereafter: following rate) was calculated based on the percentage of trials in which participants were following the cues and responding congruently with the cue. Comparisons across conditions and across studies were made using ANOVAs. We planned a priori separate comparisons to investigate differences across the three experiments and feedback conditions (results Section 3.5), in order to address our second and third hypotheses (H2 and H3). Throughout the paper, multiple comparisons were corrected and *p*-values were reported according to Tukey's correction. Eta-squared equations were used to calculate effect sizes for ANOVAs.

3. Results and discussion

In this section, we present the results and discussion for each of the three studies separately. The comparison of the three studies will then be reported in a separate section.

3.1. Experiment 1

3.1.1. Accuracy and following rate

Participants' average performance rates were calculated on the entire game session and submitted separately to a two-way ANOVA with hint validity (between-subjects) and the presence of social feedback (between-subjects) as factors. We found a main effect of validity such that participants from the 20% validity condition were significantly less accurate ($F(1,195) = 94.603$, $p < .001$, $\eta^2 = 0.323$) than the participants in the 80% validity condition as shown in Fig. 3a. Participants in the 20% validity group also followed the robot hints less ($F(1,195) = 458.533$, $p < .001$, $\eta^2 = 0.699$) than participants in the 80% validity group as shown in Fig. 3b. However, our results did not confirm our first hypothesis H1, in that no main effect of social feedback emerged (Following rate: $F(1,195) = 2.018$, $p = .157$; Accuracy: $F(1,195) = 0.073$, $p = .787$) or interaction with validity (Following rate: $F(1,195) = 0.146$, $p = .702$; Accuracy: $F(1,195) = 2.995$, $p = .085$).

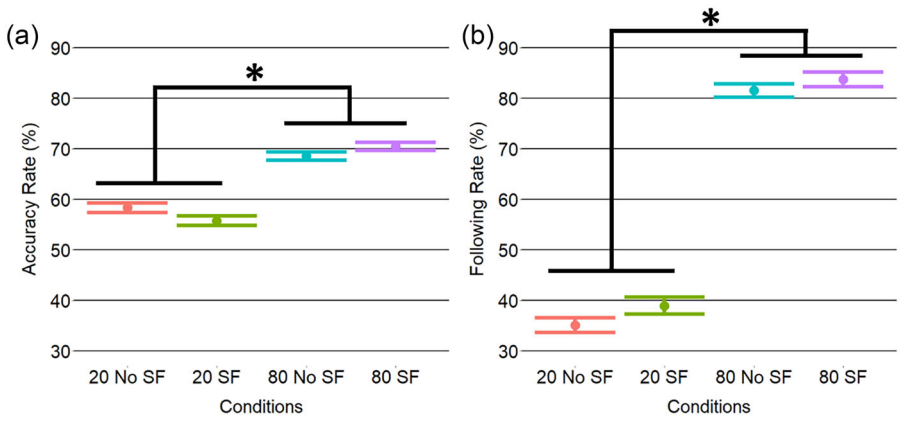

Fig. 3. Participants' performance rates. SF = social feedback; No SF = no social feedback; "20" denotes 20% validity, while "80" denotes 80% validity. (a) Accuracy rate: percentage of hits among 100 trials, (b) following rate: percentage of responses following robot cue among 100 trials. * indicates a group-wise difference in which p-value is below .05.

3.1.2. RTs

RTs were averaged for each participant and then submitted to a between-subjects two-way ANOVA. The analysis revealed a main effect of validity ($F(1,195) = 5.944$, $p = .016$, $\eta^2 = 0.029$), with 20% validity condition yielding slower RTs than 80% validity condition, a trend associated with the presence of social feedback ($F(1,195) = 1.273$, $p = .059$, $\eta^2 = 0.018$) and no interaction ($F(1,195) = 1.273$, $p = .261$, $\eta^2 = 0.006$; see Fig. 4a). Moreover, because social feedback happens after the decision of the current trial, it could potentially influence the following trials. Thus, we sought to examine participants' RTs throughout the experiment using the following non-linear model with parameters A, B, and C to fit the decrease of RT over trials:

$$y = A.exp(Bx) + C.$$

Curve fitting qualitatively suggested a possible effect of time where the difference between conditions increases over trials to become more marked toward the end of the experiment. Therefore, we split the data into five time bins. We then submitted RTs to a mixed three-way ANOVA where validity and presence of social feedback were between-subject factors with two levels each and time bins as a within-subject factor with five levels. Results showed a main effect of bin ($F(4,975) = 12.637$, $p < .001$, $\eta^2 = 0.047$), main effect of validity ($F(1,975) = 19.107$, $p < .001$, $\eta^2 = 0.018$), main effect of the presence of social feedback ($F(1,975) = 12.966$, $p < .001$, $\eta^2 = 0.012$), and a significant interaction between validity and presence of social feedback ($F(1,975) = 4.478$, $p = .035$, $\eta^2 = 0.004$). Post hoc analysis showed that participants were significantly slower in the 20-SF condition, compared to the other three conditions (all $p_{tukey} < .001$; see Fig. 4b).

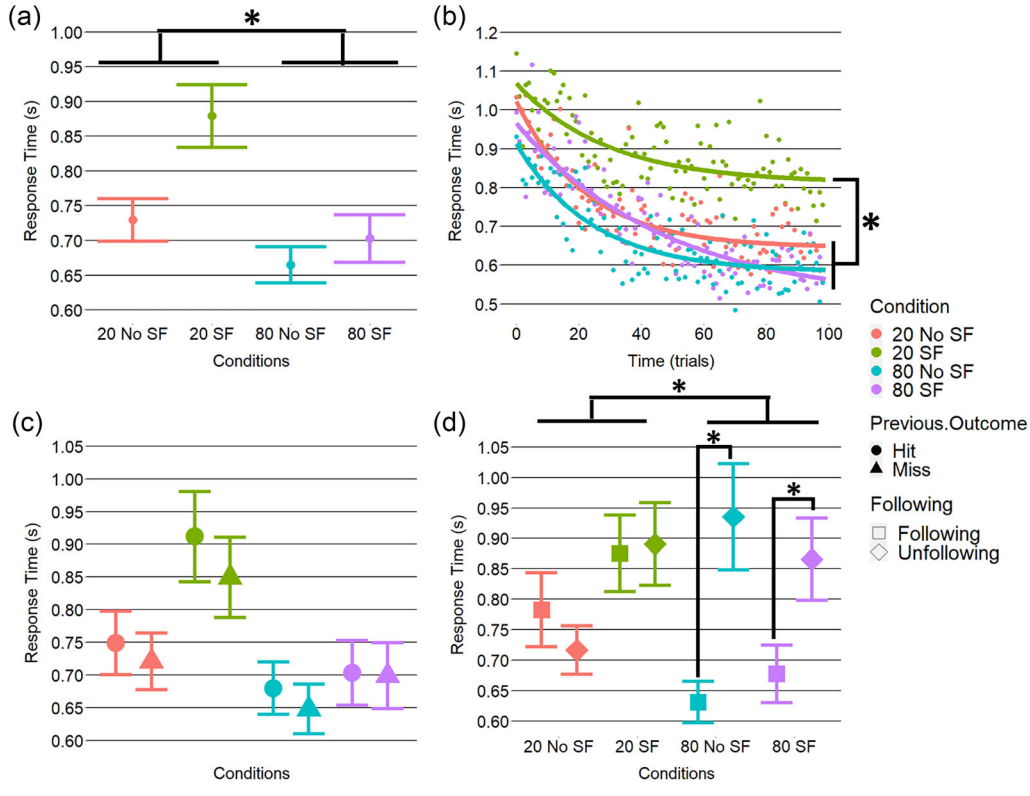

Fig. 4. (a) Averaged response times (RTs) among the four experimental conditions. SF = social feedback; No SF = no social feedback; "20" denotes 20% validity, while "80" denotes 80% validity; (b) curve fitting on each of the four conditions over time; (c) effect of previous outcome on RTs among the four experimental conditions; (d) effect of following robot cues on RTs. * indicates a group-wise difference in which *p*-value is below .05.

Our second hypothesis H2 was that the incongruence between invalid cues and positive feedback upon successful guesses might violate participants' social expectations, which could explain why we observed slower responses in the 20% validity with social feedback (20-SF) condition. However, other explanations could also be considered. For instance, one of the possible explanations for later responses in the 20-SF condition is that, because iCub only reacted to hits (successful trials), the absence of feedback after misses (unsuccessful trials) delayed responses in subsequent trials. Those trials being more frequent when cues were 20% valid would result in overall slower RTs in the 20-SF condition. To evaluate this hypothesis, we tested the effect of previous outcomes on RTs in subsequent trials in each condition through a three-way ANOVA, splitting the data based on whether the previous trial was successful and thus followed by positive feedback from the robot. As reported in Fig. 4c, no main effect of the previous outcome was found ($F(1,390) = 0.764$, $p = .383$, $\eta^2 = 0.002$) neither other interaction (all $p > .135$), thus not supporting this interpretation.

Yet another alternative explanation could be that the observed delay was due to cognitive control, required to inhibit the cued response and choose the opposite side. If this had been

the case, participants would have been slower when they did not follow the hint, that is, in the trial where they inhibited the cued response. To assess this interpretation, we submitted the data to a three-way ANOVA with a hint following as a third factor. The analysis showed a main effect of hint following ($F(1,382) = 6.717$, $p = .01$, $\eta^2 = 0.017$) and an interaction effect of hint following and cue validity ($F(1,382) = 10.182$, $p = .002$, $\eta^2 = 0.025$), cf. Fig. 3d. No other interaction effects were found to be significant (all $p > .09$).

Post hoc analysis showed that when validity was at 80%, participants were indeed faster when following, compared to trials where they were not following the cues ($t = -4.034$, $p_{\text{tukey}} < .001$) as shown in Fig. 3d, turquoise, diamond versus square. Moreover, when cue validity was at 20%, participants were slower in following than at 80% validity ($t = 2.934$, $p_{\text{tukey}} = .019$), indicating that the 20% conditions did require stronger cognitive control than the 80% condition. However, within the 20% validity conditions, no difference in RTs was found between following and unfollowing (all $p > .127$). In other words, the slower responses observed in the 20-SF condition cannot be entirely explained by cognitive control aiming to inhibit the cued response.

3.2. Experiment 2

Having established that the effect of RT observed in Experiment 1 could not be entirely be explained by the expectation of the presence (or absence) of feedback or by cognitive control, we designed follow-up experiments aiming to further examine whether this effect could indeed be attributed to the violation of social expectations as we hypothesized in H2. In Experiment 2, we were interested in testing whether the difference we observed in RTs was due to purely attentional mechanisms triggered by the cue rather than it being a social effect. To address this question, we replaced the social cue of the robot with an attention-capturing non-social signal of a flashlight (see Fig. 2). Here, we only included the two 20% validity conditions, as this condition showed the strongest effect in combination with the social feedback when the time course of the experiment was taken into account (Fig. 3b). Although the flashlight replaced the social cues, social feedback was still given by the robot. This was done in order to prevent the modification of two factors simultaneously.

3.2.1. Accuracy and following rate

Participants' average performance rates were calculated on the entire game session and submitted separately to an independent *t*-test with the presence of social feedback as a factor. No effect was found between the two flashlight conditions on accuracy ($t(98) = -0.613$, $p = .541$) or following rates ($t(98) = 0.518$, $p = .606$). Accuracy mean and standard deviation was 56.1 ± 14.9 for the non-social feedback and 57.8 ± 13.5 for the social feedback condition. Following rate mean and standard deviation were 36.8 ± 23.1 for the non-social feedback condition and 34.5 ± 20.3 for the social feedback condition.

3.2.2. RT

RTs were averaged for each participant and then submitted to an independent sample *t*-test with social feedback/no feedback as a factor. No effect was found ($t(98) = -0.740$,

$p = .461$) where the mean RT was 0.7 ± 0.4 for the no social feedback condition and 0.7 ± 0.2 for the social feedback condition. Similarly to Experiment 1, we then submitted RTs to a mixed two-way ANOVA where the presence of social feedback is a between-subject factor with two levels and time bins as a within-subject factor with five levels. Results showed a main effect of time ($F(4,488) = 4.664$, $p < .001$, $\eta^2 = 0.037$). However, no main effect of the presence of social feedback ($p = .176$) or significant interaction with time ($p = .746$) was found, indicating that the influence of the social feedback found in Experiment 1 disappears when the pre-decision cue is not social.

3.3. Experiment 1 versus 2

3.3.1. Accuracy and following rate

Participants' accuracy rates for the 20 validity conditions were calculated on the entire game session and submitted separately to a mixed two-way ANOVA with the type of cue (Experiment 1 vs. Experiment 2) and the presence of social feedback as factors. We did not find any effect of the type of cue or feedback on accuracy or following rates (all p-values $>$ .085).

3.3.2. RT

RTs were averaged for each participant and then submitted to a mixed two-way ANOVA to examine differences across conditions. The analysis showed no effect of social feedback. However, it revealed a main effect of the type of cue ($F(1,196) = 5.501$, $p = .02$, $\eta^2 = 0.027$) and significant interaction between the type of cue and presence of social feedback ($F(1,196) = 3.916$, $p = .049$, $\eta^2 = 0.019$). The comparison between the robot condition (Experiment 1) and flashlight condition (Experiment 2) is illustrated in Fig. 4. Next, we performed post hoc tests showing a difference when the social feedback was presented between robot and flash conditions ($t = -3.058$, $p_{\text{tukey}} = .013$). No other post hoc t-test was found to be significant (all p-values $>$ .09). This result confirms that the effect is not purely related to attention and supports the hypothesis that violation of social expectations about congruent cues and feedback affects decision-making processes.

3.4. Experiment 3

If the effect observed in Experiment 1 is indeed related to the social aspect of the cue–feedback relationship, we reasoned that we should observe a similar effect if social signals were provided by a human. Therefore, in Experiment 3, we replaced the robot stimuli with human stimuli (see Fig. 2). As in Experiment 2, we only included the 20% validity condition. Our goal was: (1) to examine whether the effect on RTs found in Experiment 1 replicates when the two social signals (i.e., pre-decision cue and post-decision feedback) are provided by a social agent; and (2) to assess whether the robot's signals were indeed acting as social signals and had similar effects to human signals (H3).

3.4.1. Accuracy and following rate

Participants' averaged performance rates were calculated on the entire game session and submitted separately to an independent samples *t*-test with the presence of social feedback as a factor. No effect was found between the two conditions on accuracy ($t(97) = -0.623$, $p = .535$) or following rates ($t(97) = 0.307$, $p = .759$). Accuracy mean and standard deviation were 56.9 ± 11.723 for the non-social feedback and 58.286 ± 10.344 for the social feedback condition. Following rate mean and standard deviation were 35.82 ± 17.354 for the non-social feedback condition and 34.776 ± 16.423 for the social feedback condition.

3.4.2. RT

RTs were averaged for each participant and then submitted to an independent samples *t*-test with social feedback (present/absent) as a factor. No effect was found ($t(97) = -0.689$, $p = .492$). We then submitted RTs to a mixed two-way ANOVA where the presence of social feedback was a between-subject factor with two levels and time bins were a within-subject factor with five levels (five bins). As in Experiment 2, there was a main effect of time ($F(4,483) = 5.634$, $p < .001$, $\eta^2 = 0.044$) but no main effect of the presence of social feedback ($p = .140$) nor significant interaction between time and social feedback ($p = .992$). Thus, the incongruence effect as a function of time found in Experiment 1 was not replicated when a human exhibited social signals.

3.5. Comparison across 20% validity conditions across all three experiments

3.5.1. Accuracy and following rate

Participants' averaged performance rates were submitted to a mixed two-way ANOVA with type of cue (human, robot, and flashlight) and the presence of social feedback as factors. No effect of the two factors was found on accuracy (all *p*-values $> .368$) or on following rate (all *p*-values $> .463$). Descriptive statistics of the behavioral measures for the three experiments are summarized in Table 1.

3.5.2. RT

Mean participants' RTs were submitted to a mixed two-way ANOVA with the presence of social feedback and type of cue as between-subject factors. This analysis revealed a main effect of type of cue ($F(2,293) = 4.212$, $p = .016$, $\eta^2 = 0.027$) with flashlight yielding the fastest RTs, followed by robot condition and then the human condition (see Table 1, RTs row). However, no effect of the presence of social feedback ($p = .213$) nor interaction ($p = .178$) was found as shown in Fig. 5. Planned post hoc comparisons showed that, while the difference between the flashlight and the human conditions was significant ($t = -2.760$, $p_{\text{tukey}} = .017$), there was no difference between the robot and human conditions ($t = 0.611$, $p_{\text{tukey}} = .814$). Additionally, although it did not reach significance, there was a trend toward a difference between robot and flashlight conditions ($t = -2.155$, $p_{\text{tukey}} = .081$). Consistently with our hypothesis H2 that the effect on RTs is due to the incongruence between the two social signals (cue and feedback), it appears that the difference between flashlight and robot is mainly driven by conditions where the social feedback was provided (Fig. 5). Considering

Table 1
The table summarizes the descriptive statistics for the 20% validity conditions across the three experiments.

	Human		Robot		Flashlight	
	No SF	SF	No SF	SF	No SF	SF
Accuracy	56.9 ± 11.723	58.286 ± 10.344	58.3 ± 9.844	55.76 ± 9.899	56.06 ± 14.926	57.816 ± 13.45
Following rate	32.82 ± 17.354	34.776 ± 16.423	35.08 ± 14.829	38.92 ± 16.807	36.78 ± 23.149	34.51 ± 20.32
RTs	0.809 ± 0.459	0.867 ± 0.372	0.73 ± 0.306	0.882 ± 0.452	0.715 ± 0.374	0.668 ± 0.242

Note. For each measure, we reported mean and standard deviation.
Abbreviations: No SF, condition with no social feedback; RTs, response times; SF, condition where the agent showed social feedback.

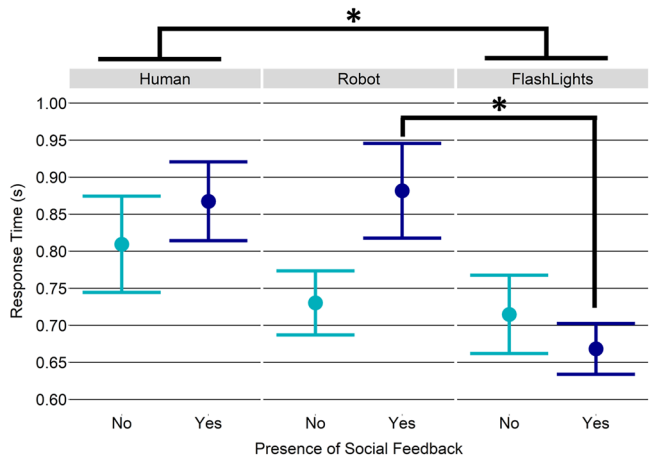

Fig. 5. Twenty percent validity conditions from the tree studies divided by the type of cue (top labels) and presence of social feedback (bottom labels). * indicates a group-wise difference in which *p*-value is below .05.

only those conditions with social feedback, we found a significant difference between robot and flashlight ($t = -2.917$, $p_{\text{tukey}} = .011$) while this difference is not significant between robot and human ($t = -0.193$, $p_{\text{tukey}} = .980$). On the other hand, in conditions with no social feedback, no difference was found between robot and flashlight ($t = -0.201$, $p_{\text{tukey}} = .978$) nor between robot and human ($t = 1.029$, $p_{\text{tukey}} = .560$).

4. General discussion

This series of studies had two objectives: (i) to examine the effects of a combination of pre- and post-decision social signals on participants' decision processes, and (ii) to assess whether robot signals affected performance similarly to human signals. We performed three studies using a two-alternative choice task where we manipulated the congruency of the pre-decision cue and the presence of post-decision social feedback. The studies differed in the type of agent involved: pre-decision cues were given by a robot (Experiment 1), a flashlight (Experiment 2), or a human (Experiment 3); and post-decision social feedback was given by a robot (Experiments 1 and 2) or a human (Experiment 3).

Based on previous research employing verbal social praise as feedback, we hypothesized that positive reactions from the robot to participants' wins would increase their trust in the robot and that it might in turn increase their tendency to follow its cues. Our results did not confirm this hypothesis, not only in the robot condition but also in the human condition. One explanation could be that social praise can only be mediated by verbal feedback. Alternatively, it could be that the task—that is, following valid cues and unfollowing invalid cues—was too simplistic for the participants' choices to be altered by the other agent's feedback. In addition, for the human condition, we presented participants only with 20% validity conditions. Thus, the cues were highly unreliable (from participants' perspective). Therefore, in case of such high unreliability of the cues, participants simply might have chosen the strategy to not

follow the cues, and the feedback could not have influenced that strategy. In the robot condition, following rate was strictly related to validity (there was in fact a main effect of validity). However, this was not modulated by social feedback. Thus, it seems that for the rate of following, strategic control was the main mechanism influencing participants' behavior (following rate), and the social feedback was not potent enough to modulate the control mechanism. This is in line with the results of Kompatsiari et al. (2022) where the authors showed that in a gaze cueing paradigm, strategic control is a more potent factor than social signals, regarding orienting of attention. Finally, it might also be that the agents being represented as 2D photographs were not naturalistic enough to evoke reactions to the feedback they provided, especially since the study was conducted online. Further investigation is needed to disentangle these questions and to better understand how robot social feedback may influence human trust and decision-making processes.

In line with our second hypothesis, our main results are related to participants' RTs, which showed a slower decrease over the course of the experiment in the condition with low cue validity and social feedback—that is, when the robot was mostly giving invalid cues but providing positive feedback after successful guesses (Experiment 1, Fig. 3b). RTs can be particularly informative in the study of decision-making (Gold & Shalden, 2007; Ratcliff, 2013). Typically, choices made instinctively are much faster than those involving a high degree of deliberation (Rubinstein, 2007). In this study, we expected slower responses in the condition with low cue validity and social feedback as a result of the violation of the expectation of congruence between cue and feedback. Although this effect was not observed when considering RTs averaged over trials, we found that it emerged over the course of the experiment. This might be related to participants' realization only over time of the validity of the cue. The validity of the cue is not obvious at the beginning of the experiment and requires the accumulation of a sufficient amount of "samples" to understand that the agent is providing mostly invalid cues. In consequence, the discrepancy between the cue validity and the social feedback also becomes evident only as the experiment progresses.

To further test whether this effect was due to the relationship between the cue and the feedback as social signals, we replaced the robot cue with a non-social cue (flashlight) in Experiment 2. As a result, the effect of increased RTs for incongruent cues, emerging over the course of Experiment 1 was not observed anymore. This confirmed that the effect could not be attributed to non-social processes, such as attention orienting to the saliency of the cue. Overall, the analyses conducted in Experiments 1 and 2 indicate that the effect is less likely to be explained by either attentional mechanisms, feedback expectation, or cognitive control. Instead, our findings suggest that the effect related to (in)congruency of social cues and social feedback was driven by the expectations of consistency between the social cue and the feedback communicated by the robot. The violation of such expectation may have led to the engagement of cognitive processes related to reasoning about the robot's actions and intentions, or to a higher effort needed to suppress the cued response it has become evident that it is in contradiction with the feedback, most of the time (see similar results by Belkaid et al., 2021).

However, the effect did not replicate when the cue and feedback were provided by a human in Experiment 3. In addition, our third hypothesis that robot signals would elicit similar

responses to human signals was only moderately supported by comparing the three experiments. This suggests that people have different social expectations about humans and robots. For instance, one can speculate that, as machines, robots are expected to behave in a more consistent, predictable way, compared to humans. In this case, the violation of the expectation of congruent cues and feedback by a human agent is thus less surprising and salient than by a robot agent.

Overall, this study highlights the importance of studying HRIs from the perspective of experimental psychology. Indeed, previous research suggests that people can perceive robots as social agents (Hortensius & Cross, 2018; Marchesi et al., 2019; Sheridan, 2020). Among other non-verbal signals, extensive research has underscored the role of eye contact and gaze behavior in HRI (Kompatsiari et al., 2019, 2021, 2022) and the consequences of those signals on participants' performance (i.e., prolonged RTs) and overall quality of the interaction (reduced sensitivity to outcomes; Belkaid et al., 2021). Concurrently, there is an increasing number of foreseen applications for interactive robots, from collaborative manufacturing to daily assistance or therapy. For instance, robots equipped with non-verbal signals have been employed for both adult (Fasola & Mataric, 2013) and children (Ghiglino et al., 2023; Scassellati et al., 2018) trainings in a clinical context. Understanding how users perceive signals exhibited by these machines is therefore critical for designing effective technologies that meet the users' needs and application requirements (Belkaid et al., 2021).

The present study illustrates how expectations about robots and humans during social interactions may differ. This needs to be taken into account to interpret results from robot-based paradigms. On the other hand, it also emphasizes the opportunity presented by robots as behaviorally complex social agents that are yet not necessarily subject to the same social norms as those at play in human–human interactions. Indeed, there may be a mismatch between what we expect from a machine versus a social agent. We could be more forgiving of a human who breaks a social rule but less forgiving of a robot that is expected to be programmed correctly. For instance, Dietvorst and colleagues described how "algorithm aversion" leads humans to be less forgiving of AI forecasting algorithms when making errors, compared to human forecasters (Dietvorst, Simmons, & Massey, 2014). On the contrary, we could tolerate more errors from a machine that did not learn our social norms. Interestingly, robots can be designed to behave according to certain norms or to break them. Interestingly, Leib and colleagues tested how AI-generated advice could corrupt people's behavior during a task as much as a human advice could do (Leib, Köbis, Rilke, Hagens, & Irlenbusch, 2021). Moreover, social norms and expectations are constantly updated throughout interactions. By exploiting differences in people's expectations about humans and robots, and by manipulating how participants perceive the robot and how the robot behaves, researchers can cast novel insights into how social expectations and norms are formed and updated in human interactions.

4.1. Limitations and future research

One limitation of our study is the absence of tangible incentives for participants. While we aimed to examine the impact of non-verbal cues in a controlled environment, the absence

of real-world consequences or monetary gains may have attenuated the effects on decision-making. Future work could investigate how actual gains (monetary or otherwise) affect motivation, performance, and cue following, providing a more ecologically valid perspective on decision-making in HRI. Another limitation to consider is that our study utilized a specific robot representation, one that might be perceived as cute and approachable. HRIs can involve a wide range of robot designs, from utilitarian to anthropomorphic. The extent to which non-verbal cues impact decision-making might vary with different robot designs. Future research could explore how diverse robot representations influence the interpretation and effects of non-verbal cues in interaction.

5. Conclusion

The current study focused on investigating how robot signals affect human decision-making processes and how this compares to decision-making in other social and non-social scenarios. We showed that violation of expectations regarding congruency of pre- and post-decision social signals expressed by the robot affects performance in decision-making. Interestingly, the human agents providing the social signals were not scrutinized with the same degree of expectations. These findings highlight the importance of studying people's expectations toward social robots and the potential effects of the violation of those expectations in behavioral and psychological terms. As a relatively new technology, expectations about social robots may vary. This can be due to a lack of understanding of their behavioral and cognitive capabilities, but also to a mismatch between what we expect from a machine versus a social agent. Examining the factors underlying such expectations and the effects of deviating from them is essential to develop effective and intuitive ways for humans and robots to interact and work together.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant awarded to Agnieszka Wykowska, titled "InStance: Intentional stance for social attunement." G.A. no.: ERC-2016-StG- 715058). The content of this paper is the sole responsibility of the authors. The European Commission or its services cannot be held responsible for any use that may be made of the information it contains.

Conflict of interest

The authors declare no competing interests.

References

Abubshait, A., & Wiese, E. (2017). You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in Psychology*, *8*, 1393.

Abubshait, A., Momen, A., & Wiese, E. (2020). Pre-exposure to ambiguous faces modulates top-down control of attentional orienting to counterpredictive gaze cues. *Frontiers in Psychology*, *11*, 2234.

Abubshait, A., Beatty, P. J., McDonald, C. G., Hassall, C. D., Krigolson, O. E., & Wiese, E. (2021). A win-win situation: Does familiarity with a social robot modulate feedback monitoring and learning? *Cognitive, Affective, & Behavioral Neuroscience*, *21*, 763–775.

Abubshait, A., Pérez-Osorio, J., De Tommaso, D., & Wykowska, A. (2023). Conflicting demands during a handover task with a robot affect the neural pattern of the human cognitive system. Available at: https://doi.org/10.31219/osf.io/derq7 accessed on September 10, 2023.

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, *6*(1), 25–63.

Bartholow, B. D., Fabiani, M., Gratton, G., & Bettencourt, B. A. (2001). A psychophysiological examination of cognitive processing of and affective responses to social expectancy violations. *Psychological Science*, *12*(3), 197–204.

Belkaid, M., Kompatsiari, K., De Tommaso, D., Zablith, I., & Wykowska, A. (2021). Mutual gaze with a robot affects human neural activity and delays decision-making processes. *Science Robotics*, *6*(58), eabc5044.

Boucher, J. D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., Domiey, P. F., & Ventre-Dominey, J. (2012). I reach faster when I see you look: Gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in Neurorobotics*, *6*, 3.

Britannica, T. (2023). Editors of Encyclopaedia. cups and balls trick. Encyclopedia Britannica. Available at: https://www.britannica.com/art/cups-and-balls-trick

Browning, M., & Harmer, C. J. (2012). Expectancy and surprise predict neural and behavioral measures of attention to threatening stimuli. *Neuroimage*, *59*(2), 1942–1948.

Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology*, *12*(1-2), 30–48.

Burgoon, J. K., & Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communications Monographs*, *55*(1), 58–79.

Burgoon, J. K., Guerrero, L. K., & Floyd, K. (2016). *Nonverbal communication*. New York: Routledge.

Burgoon, J. K., Manusov, V., & Guerrero, L. K. (2010). *Nonverbal Communication* (1st ed.). Routledge. https://doi.org/10.4324/9781315663425

Chidambaram, V., Chiang, Y. H., & Mutlu, B. (2012). Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. *Proceedings of 2012 ACM/IEEE international conference on Human-Robot Interaction*, Boston, MA.

Ciardo, F., & Wykowska, A. (2022). Robot's social gaze affects conflict resolution but not conflict adaptations. *Journal of Cognition*, *5*(1), 2.

Ciardo, F., De Tommaso, D., & Wykowska, A. (2022). Human-like behavioral variability blurs the distinction between a human and a machine in a nonverbal Turing test. *Science Robotics*, *7*(68), eabo1241.

Diederich, S., Brendel, A. B., Morana, S., & Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, *23*(1), 96–138.

Dietvorst, B. J., Simmons, J., & Massey, C. (2014). Understanding algorithm aversion: Forecasters erroneously avoid algorithms after seeing them err. In *Academy of management proceedings* (Vol. *2014*, p. 12227). Briarcliff Manor, NY 10510: Academy of Management.

Edwards, A., Edwards, C., Westerman, D., & Spence, P. R. (2019). Initial expectations, interactions, and beyond with social robots. *Computers in Human Behavior*, *90*, 308–314.

Fasola, J., & Matarić, M. J. (2013). A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction*, *2*(2), 3–32.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, *132*, 138–161.

Ferdinand, N. K., Mecklinger, A., & Opitz, B. (2015). Learning context modulates the processing of expectancy violations. *Brain Research*, *1629*, 72–84.

Fiore, S. M., Wiltshire, T. J., Lobato, E. J. C., Jentsch, F. G., Huang, W. H., & Axelrod, B. (2013). Toward understanding social cues and signals in human-robot interaction: Effects of robot gaze and proxemic behavior. *Frontiers in Psychology*, *4*, 859. https://doi.org/10.3389/fpsyg.2013.00859

Ghazali, A. S., Ham, J., Barakova, E., & Markopoulos, P. (2018). The influence of social cues in persuasive social robots on psychological reactance and compliance. *Computers in Human Behavior*, *87*, 58–65.

Ghiglino, D., Willemse, C., De Tommaso, D., & Wykowska, A. (2021). Mind the eyes: Artificial agents' eye movements modulate attentional engagement and anthropomorphic attribution. *Frontiers in Robotics and AI*, *8*, 642796.

Ghiglino, D., Floris, F., De Tommaso, D., Kompatsiari, K., Chevalier, P., Priolo, T., & Wykowska, A. (2023). Artificial scaffolding: Augmenting social cognition by means of robot technology. *Autism Research*, *16*(5), 997–1008.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision-making. *Annual Review Neuroscience*, *30*, 535–574

Gonsior, B., Sosnowski, S., Mayer, C., Blume, J., Radig, B., Wollherr, D., & Kühnlenz, K. (2011). Improving aspects of empathy and subjective performance for HRI through mirroring facial expressions. *2011 RO-MAN*, Atlanta, GA (pp. 350–356).

Ham, J., & Midden, C. J. (2014). A persuasive robot to stimulate energy conservation: The influence of positive and negative social feedback and task similarity on energy-consumption behavior. *International Journal of Social Robotics*, *6*(2), 163–171.

Horstmann, A. C., & Krämer, N. C. (2020). Expectations vs. actual behavior of a social robot: An experimental investigation of the effects of a social robot's interaction skill level and its expected future role on people's evaluations. *PLoS One*, *15*(8), e0238133.

Horstmann, A. C., & Krämer, N. C. (2019). Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in Psychology*, *10*, 939.

Hortensius, R., & Cross, E. S. (2018). From automata to animate beings: The scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences*, *1426*(1), 93–110.

Ito, A., Hayakawa, S., & Terada, T. (2004). Why robots need body for mind communication-an attempt of eye-contact between human and robot. RO-MAN 2004*,* 13th IEEE International Workshop on Robot and Human Interactive Communication, Kurashiki, Japan (pp. 473–478).

Kahn, P. H. Jr., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J. H., & Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. *Proceedings of the 6th International Conference on Human-Robot Interaction*, Lusanne, Switzerland (pp. 159–160).

Knapp, M. L., & Harrison, R. P. (1972). Observing and recording nonverbal data in human transactions. *Annual Convention of the Speech Communication Association*, Chicago, IL.

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Boston, MA: Cengage Learning.

Kompatsiari, K., Bossi, F., & Wykowska, A. (2021). Eye contact during joint attention with a humanoid robot modulates oscillatory brain activity. *Social Cognitive and Affective Neuroscience*, *16*(4), 383–392.

Kompatsiari, K., Ciardo, F., & Wykowska, A. (2022). To follow or not to follow your gaze: The interplay between strategic control and the eye contact effect on gaze-induced attention orienting. *Journal of Experimental Psychology: General*, *151*(1), 121.

Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G., & Wykowska, A. (2019). It's in the eyes: The engaging role of eye contact in HRI. *International Journal of Social Robotics*, *13*, 525–535. https://doi.org/10.1007/s12369-019-00565-4

Kwon, M., Jung, M. F., & Knepper, R. A. (2016). Human expectations of social robots. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Christchurch, New Zealand (pp. 463–464).

Leathers, D. G. (1976). *Nonverbal communication systems*. Boston, MA: Allyn and Bacon.

Eaves, M., & Leathers, D. G. (2017). *Successful nonverbal communication: Principles and applications*. London: Routledge.

Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2021). The corruptive force of AI-generated advice. Available at: http://arxiv.org/abs/2102.07536 accessed on September 4, 2023.

Lombardi, M., Roselli, C., Kompatsiari, K., Rospo, F., Natale, L., & Wykowska, A. (2023). The impact of facial expression and communicative gaze of a humanoid robot on individual Sense of Agency. *Scientific Reports*, *13*(1), 10113.

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology*, *10*, 450. https://doi.org/10.3389/fpsyg.2019.00450

Marchesi, S., Bossi, F., Ghiglino, D., De Tommaso, D., & Wykowska, A. (2021). I am looking for your mind: Pupil dilation predicts individual differences in sensitivity to hints of human-likeness in robot behavior. *Frontiers in Robotics and AI*, *8*, 653537.

Mendes, W. B., Blascovich, J., Hunter, S. B., Lickel, B., & Jost, J. T. (2007). Threatened by the unexpected: Physiological responses during social interactions with expectancy-violating partners. *Journal of Personality and Social Psychology*, *92*(4), 698.

Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The iCub humanoid robot: An open platform for research in embodied cognition. *Performance Metrics for Intelligent Systems (PerMIS) Workshop*, Gaithersburg, MD (pp. 50–56). https://doi.org/10.1145/1774674.1774683

Moon, A., Troniak, D. M., Gleeson, B., Pan, M. K., Zheng, M., Blumer, B. A., & Croft, E. A. (2014). Meet me where I'm gazing: How shared attention gaze affects human-robot handover timing. *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, Bielefeld, Germany (pp. 334–341).

Palinko, O., Rea, F., Sandini, G., & Sciutti, A. (2016). Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, Korea (pp. 5048–5054).

Parenti, L., Belkaid, M., & Wykowska, A. (2021a). Understanding how social signals affect decision-making: Current challenges and robot-based methodological solutions. https://doi.org/10.31219/osf.io/2g5sf

Parenti, L., Lukomski, A. W., De Tommaso, D., Belkaid, M., & Wykowska, A. (2023). Human-likeness of feedback gestures affects decision processes and subjective trust. *International Journal of Social Robotics*, *15*(8), 1419–1427.

Parenti, L., Marchesi, S., Belkaid, M., & Wykowska, A. (2021b). Exposure to robotic virtual agent affects adoption of intentional stance. *Proceedings of the 9th International Conference on Human-Agent Interaction*, New Orleans, LA (pp. 348–353).

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. 10.3758/s13428-018-01193-y

Perez-Osorio, J., Marchesi, S., Ghiglino, D., Ince, M., & Wykowska, A. (2019). More than you expect: Priors influence on the adoption of intentional stance toward humanoid robots. In *Social robotics. ICSR 2019. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 119–129). Cham: Springer. https://doi.org/10.1007/978-3-030-35888-4_12 accessed on February 2, 2023.

Perez-Osorio, J., Abubshait, A., & Wykowska, A. (2021). Irrelevant robot signals in a categorization task induce cognitive conflict in performance, eye trajectories, the n2 component of the EEG signal, and frontal theta oscillations. *Journal of Cognitive Neuroscience*, *34*(1), 108–126.

Proulx, T., Sleegers, W., & Tritt, S. M. (2017). The expectancy bias: Expectancy-violating faces evoke earlier pupillary dilation than neutral or negative faces. *Journal of Experimental Social Psychology*, *70*, 69–79.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510.

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review*, *120*, 281–292

Romat, H., Williams, M. A., Wang, X., Johnston, B., & Bard, H. (2016). Natural human-robot interaction using social cues. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Christchurch, New Zealand (pp. 503–504).

RStudio Team (2020). RStudio: Integrated development for R. RStudio, PBC. Available at: http://www.rstudio.com/ accessed on February 16, 2023.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, *117*(523), 1243–1259.

Sanfey, A. G. (2007). Social decision-making: Insights from game theory and neuroscience. *Science*, *318*(5850), 598–602. https://doi.org/10.1126/science.1142996

Scassellati, B., Boccanfuso, L., Huang, C. M., Mademtzi, M., Qin, M., Salomons, N., Ventola, P., & Shic, F. (2018). Improving social skills in children with ASD using a long-term, in-home social robot. *Science Robotics*, *3*(21), eaat7544.

Sheridan, T. B. (2020). A review of recent research in social robotics. *Current Opinion in Psychology*, *36*, 7–12.

Spatola, N., Marchesi, S., & Wykowska, A. (2022). Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism. *Frontiers in Robotics and AI*, *9*, 863319.

Szafir, D., & Mutlu, B. (2012). Pay attention! Designing adaptive agents that monitor and improve user engagement. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, TX (pp. 11–20).

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PLoS ONE*, *7*(9), e45391. https://doi.org/10.1371/journal.pone.0045391

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, *8*, 1663.

Wykowska, A. (2020). Social robots to test flexibility of human social cognition. *International Journal of Social Robotics*, *12*(6), 1203–1211.

Wykowska, A. (2021). Robots as mirrors of the human mind. *Current Directions in Psychological Science*, *30*(1), 34–40.