If you trust me, I will trust you: the role of reciprocity in human-robot trust

Joshua Zonca*¹, Anna Folsø ², Alessandra Sciutti ¹

1. Cognitive Architecture for Collaborative Technologies (CONTACT) Unit, Italian Institute of Technology, Genoa, Italy.

2. Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy.

Corresponding author: Correspondence to Joshua Zonca, Italian Institute of Technology, Via Enrico Melen, 83, 16152 Genova, GE (Italy). Email: joshua.zonca@iit.it

Abstract
Trust is fundamental in human-human and human-robot interaction. Extensive evidence has shown that trust among humans is sustained by reciprocity, whereas knowledge on reciprocal dynamics of trust in human-robot interaction is very limited. To investigate reciprocity in human-robot trust, we designed a joint task in which a human participant and a humanoid robot made perceptual judgments while signaling their trust in the partner. The robot’s trust was dynamically manipulated along the experiment. Results show that participants were less willing to learn from a robot that was showing high trust in them. However, participants were unwilling to overtly disclose their distrust to the robot if they expected future interactions with it, suggesting the emergence of reciprocity in their overt expression of trust. Importantly, these effects disappeared when participants interacted with a computer. Our findings may have an impact on the design of robots that could effectively collaborate with humans.
Introduction

In the next future, robots will play a more and more important role in our society. In this regard, the dominant paradigm in the field of Human-Robot Interaction (HRI) has been shifting from the ambition of building robots capable of autonomously accomplish typical human tasks to the goal of designing robotic collaborators, which should be able to collaborate with humans in their everyday life. This ability would be decisive in a wide range of contexts, including education, industry and care-giving, in which the robots would assist human educators, workers and operators in their activities. However, this goal is far from easy to achieve, since collaboration and cooperation, even among humans, are fragile processes that entails the presence and the persistence of different environmental and relational conditions. In this respect, one of the most important mechanisms promoting cooperation is undoubtedly trust. Trusting our interacting partners, from peers to robots, is fundamental to delegate responsibility and accept help and advices from them. Nevertheless, trust is a multifaceted concept that has been treated differently in human-human and human-robot interaction literatures (Flook et al., 2019).

One the one hand, studies in human-human interaction have generally conceptualized trust as acceptance of vulnerability or delegation of responsibility in peer relationships (Lee and See, 2004; Mayer et al., 1995; Ullman and Malle, 2018). One factor that modulates trust in our peers is their competence, which may be inferred by feedback on their reliability (De Martino et al., 2017; Park et al., 2017), confidence (Bahrami, 2010; Sniezek and Van Swol, 2001) and expertise (Bonaccio and Dalal, 2010; Boorman et al., 2013; Sniezek et al., 2004). Nonetheless, trust-based relationships between human beings are not uniquely based on competence assessment and, most importantly, are never unidirectional. Indeed, trust in human-human interaction generally entails reciprocity (e.g., De Quervain et al., 2004; King-Casas et al., 2005; Krueger et al., 2007; McCabe et al., 2001; Singer et al., 2006): if we break relationships based on trust due to selfish behavior, it is unlikely that others will trust us in the future. This highlights the importance of social norms and relational dynamics in the emergence and maintenance of trust among humans.

On the other hand, research on trust in HRI typically conceptualized trust as a one-sided process of evaluation of the competence of the robotic agent. Robots are usually used as mechanical agents performing actions or tasks, rather than intentional agents involved in a symmetrical and reciprocal interaction. In these terms, the robot usually plays the role of the trustee, producing actions and behaviors whose reliability must be evaluated by the human partner (i.e., the trustor). Trust is treated only from the perspective of the human partner (Cameron et al., 2015) and is operationalized as a direct function of the robot’s estimated or actual performance. In this regard, numerous studies have shown that the main determinant of trust in automation is its performance (Billings et al 2012; Hancock et al., 2011; van den Brule et al., 2014; Wright et al., 2019). Humans tend to trust robots as long as they show reliable behavior, but they rapidly lose trust in in presence of failures (Desai et al., 2012; Rossi et al., 2017;
In order to improve human-robot collaboration, research in HRI has been focusing on the introduction of a social dimension in human-machine interaction through the design of humanoid “social” robots, using human-human interaction as behavioral model (Sandini and Sciutti, 2018; Strohkorb and Scassellati, 2016). The assumption is that robots appearing and behaving like humans might trigger the same types of emotional and behavioral reaction that are typically observed in human-human interaction. In fact, recent evidence has shown the emergence of pro-social attitudes towards social robots (e.g., Connolly et al., 2020; Kahn et al. 2015; Kühnlenz et al., 2018; Siegel et al., 2009), which can enhance human-robot collaboration (e.g., Admoni and Scassellati, 2017; Baraglia et al., 2017; Oliveira et al., 2021; Terzioğlu et al., 2020), but also lead to over-compliance with the instructions of faulty or unreliable robots (Aroyo et al., 2018, Robinette et al., 2016; Salem et al., 2015). However, evidence of human-like relational mechanisms of trust in HRI is rather limited and context-dependent. Compliance towards robots tend to emerge primarily on functional tasks (i.e., tasks in which the robot has to fulfill a concrete goal through actions or quantitative judgment) rather than social ones (i.e., judgments or decisions on social issues), as shown in recent studies (Gaudiello et al., 2016; Ullman et al., 2021). Moreover, compliance is intrinsically tied with the uncertainty of the task solution: individuals typically conform to robots only if they are unsure about the action to take (Hertz and Wiese, 2018). Altogether, research in HRI has been providing strong evidence on the importance of robot capabilities to sustain trust in human partners, whereas knowledge on the relational and reciprocal mechanisms underlying human-robot trust is still very scarce.

In the current study, we aimed at enriching the study of trust in human-robot interaction by treating it as a reciprocal process. Our novel paradigm consisted in three different tasks. In the Perceptual inference task, participants made individual perceptual judgments without any feedback on a touch-screen tablet. In the Individual trust task, participants made the same perceptual judgments and believed that either a humanoid robot iCub (Robot group) or a computer (Computer group) was doing the same. Participants received (on the same tablet) trial-by-trial feedback revealing the judgment performed by their partner (robot or computer). In each trial, after observing the agent’s response, participants made a trust judgment by placing a final response between their own and their partner’s estimate. In the Reciprocal trust task, the task was the same but we alternated trials in which the trust judgment was taken by the participant (decision turns) with others in which the trust judgments were made by the interacting agent (observation turns). Trust judgments made by the agent in observation turns were manipulated to express different levels of trust towards the participants (Susceptible condition: high trust; Unsusceptible condition: low trust). Partner’s perceptual estimates and trust judgments were controlled by the same algorithms in both
Robot and Computer groups. Participants in the Robot group could not see their robotic partner during the tasks. In fact, we wanted participants to acquire knowledge about the robot’s ability just by observing its (simulated) responses, in order to have full control of the robot’s feedbacks and ensure the comparability with the Computer group. By controlling and systematically modulating the agent’s level of trust, we were able to analyze how participants reacted and dynamically adjusted their own level of trust during the interaction with different types of agent. Nevertheless, we also wanted participants to perceive the robot as a social and intentional agent, which was aware of the ongoing experiment with a human partner and could autonomously perform the joint task with them. To achieve this goal, participants in the Robot group, before starting the experimental session, had an encounter with the robot, which was programmed to perform specific verbal and motor actions (for a detailed description, see paragraph “Introducing participants to the humanoid robot iCub” in the “Methods” section).

The goal of this study is two-fold. On the one hand, we explored whether and how signals of trust (or distrust) coming from a humanoid robot (rather than a computer) have an impact on the willingness of the human partner to learn and take advices from it, which reflects covert and implicit trust in the robot’s capabilities. On the other hand, we investigated whether the trust expressed by the robot, which could be interpreted a signal of a collaborative attitude during interaction, could have an impact on the overt expression of trust of participants, who may be willing to reciprocate trust to sustain cooperation with its robotic partner, following social norms typically intervening between human peers. We hypothesize that the emergence of pro-social, normative behavior is strictly linked to the human-like appearance and social-like conduct of the robotic partner and, therefore, should be observed in the Robot group only.

Results

Perceptual inference task
In each trial of the Perceptual inference task (Fig. 1.a1), participants saw two consecutive light flashes (i.e., red disks of 200 ms duration) appearing on a visible horizontal line on a touch-screen tablet. The spatial distance between the two disks represented the target stimulus. Participants were then asked to touch a point to the right of the second disk to reproduce a segment matching the target stimulus length. No feedback about the accuracy of the response was provided. At the end of the task, participants were asked to evaluate from 1 to 10 the accuracy of their perceptual estimates. As a result of the well-known phenomenon of central tendency in quantity judgments (Hollingworth, 1910, Jazayeri and Shadlen 2010), participants’ estimates should gravitate towards the mean magnitude the visual stimuli (Figure 1.a2). Central tendency is an automatic perceptual mechanism aiming at error minimization in presence of sensory uncertainty, which can be represented by Bayesian models describing the mean magnitude of the stimulus history as a prior (Jazayeri and Shadlen 2010; Cicchini et al. 2012; Sciutti et al., 2014).
Moreover, variability and uncertainty in perceptual estimation should increase along with stimulus magnitude, leading to a higher reliance on the Bayesian prior (i.e., mean magnitude of the stimulus history) in perceptual adjustment for longer stimuli (Figure 1.a2). We refer to this phenomenon as scalar variability (Petzschner et al., 2015).

a. Perceptual inference task

a1. Experimental task

![Perceptual inference task diagram](image)

 stimulated presentation Individual estimate

b. Individual trust task and Reciprocal trust task

b1. Experimental task

![Individual trust task and Reciprocal trust task diagram](image)

Stimulus presentation Individual estimate Partner’s estimate Trust judgment

t = a/d

b2. Experimental design

| Individual trust task | Reciprocal trust task |
|-----------------------|-----------------------|
| ![Individual trust task diagram](image) | ![Reciprocal trust task diagram](image) |

Fig 1. a) Perceptual inference task. a1) Experimental task. Participants saw two red disks appearing consecutively with a duration of 200 ms. Participants had to touch a point, to the right of the second disk, in order to reproduce the stimulus length (s'), defined as the distance between the first and the second disk (s). a2) Expected perceptual phenomena. Due to perceptual mechanisms such as central tendency and scalar variability, we expect a
distortion in the participants’ pattern of perceptual inferences. Due to central tendency, participants should underestimate long stimuli and over-estimate short stimuli; because of scalar variability, the central tendency effect should be stronger for long distances, leading to a general under-estimation of the mean of the reproduced length distribution (M(S’)).

b) Individual trust task and Reciprocal trust task. b1) Experimental task. In both the Individual trust task and the Reciprocal trust task, initially participants had to reproduce the lengths of visual stimuli, as in the Perceptual inference task. Participants were told that the very same stimuli would have been presented also to their partner (other agent: OA), which was a humanoid robot in the Robot group and a computer in the Computer group. Participants were told that the partner would choose a point to reproduce the length of the presented stimulus. After the participants’ estimate, the partner’s simulated estimate was shown and then one of the two agents had the opportunity to make a final decision (i.e., trust judgment) by choosing any position between own and partner’s estimates. The index of trust (t) was defined as the adjustment towards the partner (a) divided by the distance between the two agents’ responses (d).

b2) Experimental design. In the Individual trust task, participants made trust judgments in all trials. In the Reciprocal trust task, in half of the trials (decision turns) participants performed the same task with the same partner and made trust judgments. Decision turns were alternated with observation turns, in which the partner made the trust judgment, while participants could not revise their estimate. We manipulated the partner’s final decisions across two experimental conditions to express two different levels of susceptibility towards participants’ responses (Susceptible: high trust; Unsusceptible: low trust).

First, we quantified the effect of central tendency by computing the regression index, defined as the difference in slope between the identity line (i.e., veridical reproduction of stimuli) and the best linear fit of the reproduced lengths. Therefore, a regression index approaching 1 reflects complete regression to the mean, whereas a regression index approaching 0 expresses no effect of central tendency. Analysis of the regression index reveals that both groups show a significant effect of central tendency (Computer group: 0.46 ± 0.24; Robot group: 0.38 ± 0.23. Wilcoxon signed rank-test, null hypothesis: regression index = 0. Computer group: z = 4.29, r = 0.86, η² = 0.74, p < 0.001; Robot group: z = 4.34, r = 0.87, η² = 0.75, p < 0.001. These results are significant at the Bonferroni-corrected threshold for 2 comparisons. Figure 2.a).

Moreover, in line with the phenomenon of scalar variability, participants under-estimated the length of visual stimuli: the average length of the reproduced stimuli was significantly lower than the actual stimulus distribution mean in both groups (Computer group: 10.48 ± 1.39; Robot group: 11.08 ± 1.25. Wilcoxon signed rank-test, null hypothesis: average reproduced length = 12. Computer group: z = -3.86, r = 0.77, η² = 0.60, p < 0.001; Robot group: z = -3.03, r = 0.61, η² = 0.37, p = 0.002. Results are significant at the Bonferroni-corrected threshold for 2 comparisons. Figure 2.a).

Eventually, we tested the comparability of our two experimental groups by investigating the presence of a between-group difference in terms of trial-by-trial estimation error, computed as the absolute difference (in cm) between the estimated length and the actual stimulus length divided by the actual stimulus length. This parameter express the accuracy of participants’ estimates, and is directly related to central tendency.
and scalar variability. Results show that the two groups did not differ in terms of average estimation accuracy (Mean estimation error, Computer group: 0.20 ± 0.05; Robot group: 0.19 ± 0.06. Robot – Computer: Wilcoxon rank-sum test, z = 1.00, r = 0.14, η² = 0.02, p = 0.318. Figure 2.b).

**Fig. 2. Results of the Perceptual inference task. a) Responses in the Perceptual inference task in Computer and Robot groups.** Gold (Computer group) and light green (Robot group) dots represent participants’ reproduced lengths for each presented length. Orange (Computer group) and dark green (Robot group) dots represent the participants’ mean reproduce length for each presented stimulus. Orange and dark green lines represent the linear fit between presented and reproduced lengths in Computer and Robot groups, respectively. Black lines represent the identity line (veridical reproduction). The pattern of responses in both groups reflect the emergence of the mechanisms of central tendency and scalar variability depicted in Figure 1.a.2. In both groups, the slope of the reproduced lengths is flatter than the identity line, indicating central tendency. Moreover, we observe that central tendency is markedly more pronounced for long stimuli, resulting in a general under-estimation of the presented visual stimuli. b) **Average estimation error in Computer and Robot groups.** The two experimental groups are comparable in terms of estimation error, which is computed as distance (in cm) from the correct response / stimulus length (ns: not significant, Wilcoxon rank-sum test). The dotted line represents the average estimation error of the computer algorithm acting as the participants’ partner in both Computer and Robot groups in the Individual trust task (see next paragraph). All participants show higher mean estimation error than the computer algorithm.

**Individual trust task**

In the Individual trust task, in each trial participants’ made their perceptual estimate as in the Perceptual inference task and then observed the estimate (concerning the same stimulus) made by a partner (robot in the Robot group, computer in the Computer group). This feedback consisted in a vertical line indicating the exact location of the partner’s estimate on the tablet. Afterwards, participants were asked to make a
final decision (i.e., trust judgment), choosing any position between their own and the partner’s response. Participants were incentivized to be as accurate as possible in both perceptual estimates and final decisions (trust judgments). Participants were also told that their partner could see their trust judgments.

The shift from their own estimate to that of the partner (divided by the distance between the two estimates) has been used as an objective index of trust in the current partner’s choice. At the end of the task, participants evaluated (1-10) their own and the other agent's accuracy in perceptual estimation to obtain a subjective measure of perceived competence.

First, it is important to highlight that participants’ accuracy in terms of perceptual estimates was remarkably lower than the one of their partner in both experimental groups (Wilcoxon rank-sum test on estimation error. Computer group:  \( z = 6.06, r = 0.86, \eta^2 = 0.73, p < 0.001 \); Robot group:  \( z = 6.06, r = 0.86, \eta^2 = 0.73, p < 0.001 \). Results significant at the Bonferroni-corrected threshold for 2 comparisons). Participants’ estimation error was comparable across experimental groups (Wilcoxon rank-sum test,  \( z = 1.04, r = 0.15, \eta^2 = 0.02, p = 0.299 \)). Moreover, participants’ estimation error was not different from the one observed in the Perceptual inference task in both groups (Wilcoxon signed-rank test on estimation error. Computer group: \( z = 0.040, r = 0.01, \eta^2 = 0.00, p = 0.968 \); Robot group:  \( z = -0.740, r = 0.10, \eta^2 = 0.01, p = 0.459 \). Figure 3.a), suggesting that the partner’s perceptual estimate was not used as a feedback for learning. The observed inability to use the partner as a valuable information source for learning is indeed linked to a marked distortion in the participants’ perception of the own and others’ performance.

Specifically, participants were not able to recognize the higher accuracy of the partner, since their performance ratings were even higher for own accuracy (Computer group: 6.28 ± 1.10; Robot: 6.4 ± 1.08) than their partner’s one (Computer group: 5.52 ± 1.64; Robot group: 6.24 ± 1.90). See Figure 3.c).

In line with the observed over-estimation of own performance and under-estimation of the partner’s one, analysis of participants’ trust judgments reveal that participants took into account the partner’s response much less than their own estimate (Average trust, Computer group: 0.26 ± 0.13; Robot group: 0.35 ± 0.16. Wilcoxon signed rank-test, null hypothesis: average trust = 0.5. Computer group:  \( z = 4.29, r = 0.86, \eta^2 = 0.74, p < 0.001 \); Robot group:  \( z = 3.51, r = 0.70, \eta^2 = 0.49, p < 0.001 \). Participants’ average trust was significantly correlated with performance ratings (i.e., self – other rating. Spearman correlation, Computer group: rho = -0.66, p < 0.001; Robot group: -0.58, p = 0.002. Figure 3.c) revealing that trust judgments strongly depended on the perceived competence of the partner. Interestingly, we show that participants in the Robot group trusted more their partner than participants’ in the Computer group (Wilcoxon rank-sum-test of average trust,  \( z = -1.99, r = 0.28, \eta^2 = 0.08, p = 0.047 \)). Altogether, these findings suggest that participants in both groups, in absence of direct or indirect feedback revealing the accuracy of the two agents, over-estimated their own ability and under-estimated the one of the partner,
which prevented participants from learning and led to a marked lack of trust in its judgments. This lack of trust was more pronounced when participants believed the partner to be a computer rather than a robot. Then we investigated whether within-subject modulation of trust was linked to the perceived reliability of the partner’s feedback. We tested the relationship between trust and distance from the partner’s response, which was the only feedback available to participants for inferring own and other’s performance. We ran a mixed-effect linear regression with trust as dependent variable, group, distance and their interactions as independent variables, with subject as random effect (Supplementary information, Data analysis and results, model 1). Results show an effect of distance on trust in both Computer (B = -0.44, z = -12.33, p < 0.001) and Robot (B = -0.30, z = -9.36, p < 0.001) groups: more specifically, participants’ trust significantly decreased with the increase of the distance between the two estimates (Fig. 3b). One the one hand, this effect suggest that participants’ trust was modulated by the perceived reliability of the current partner’s feedback; one the other hand, it reveals that the discrepancy between the two interacting partners’ responses was interpreted by participants as a signal of the unreliability of the partner rather than themselves, confirming their egocentric bias. Moreover, we highlight that the effect is significantly stronger when participants believed they were interacting with a computer (interaction effect, B = 0.14, z = 2.87, p = 0.004), suggesting that trust in the computer partner decreased faster than trust in the robot partner, along with the perceived reliability of its feedback.
Fig. 3. Results of the Individual trust task. a) Estimation error across Perceptual inference task and Individual trust task. We plot the difference between estimation error in the Individual trust task and the Perceptual inference task, highlighting the absence of learning effects between the two tasks in both Robot and Computer groups. Error bars represent between-subject standard error of the mean (ns: not significant, Wilcoxon signed-rank test). b) Trust as a function of agents’ response distance. Standardized (within-subject) trust, plotted as a function of distance from the agent estimate in the Individual trust task. For the purpose of this graph, first we calculated the standard deviation of participants’ estimates for each stimulus length. Then, for each stimulus, we grouped trials based on the distance of the partner’s estimate from the partner’s one using the standard deviation of the participant’s estimates of that stimulus as unit of measurement. We used nine distance ranges with a step of 0.05 SDs. For instance, we grouped in the first distance range all trials in which the distance from the partner’s estimate
was less than 0.05 SDs (considering 1 SD as the standard deviation of the participants’ estimates of the current stimulus). The next range included distances ranging from 0.05 to 0.1 SDs, and so on. Then, for each of these distance ranges, we calculated the average standardized participant’s trust and finally we averaged these indices across participants. Trust was standardized to express within-subject variation of trust as a function of the distance from the partner’s response, taking into account her general level of trust. * p < 0.05, group*distance interaction effect on trust in mixed-effects model (model 1).

c) Trust and subjective performance assessment. Mean trust plotted as a function of individual performance rating (self – other) in the Individual trust task. Egocentric bias in the assessment of own and other’s performance negatively correlates with trust in both groups. Gold and green triangles on the x-axis indicate mean ratings for Computer and Robot groups, respectively. The scatter plot reports results of Spearman correlations in both groups. d) Trust in robot and computer partners. Bar plot of mean trust in Computer and Robot groups. Error bars represent between-subject standard error of the mean. Participants in both groups trusted more their own estimate than the one of the partner (trust < 0.5, p < 0.001, Wilcoxon signed-rank test). Trust was higher in the robot partner was higher than in the computer partner. * p < 0.05, Wilcoxon rank-sum-test of average trust.

Reciprocal trust task
The Reciprocal trust task consisted in two different types of experimental trial: in decision turns, the task was identical to the Individual trust task and the participant took the final decision (i.e., trust judgment) after observing the partner’s estimate. As in the Individual trust task, participants were told that their partner could observe both their perceptual estimates and trust judgments and were incentivized to be as accurate as possible in both perceptual estimates and trust judgments. In observation turns, the partner itself (robot or computer) made the trust judgment and the participant could just observe the partner’s choice. Decision and observation turns were continuously alternated along the task. In observation turns, the trust of the partner towards the participant’s estimates was manipulated in two different conditions (Susceptible: high trust; Unsusceptible: low trust. See paragraph “Agents’ trust judgments in the Reciprocal trust task” in the “Methods” section for a detailed description of the relevant algorithms). We aimed at analyzing if participants’ trust was affected by the trust shown by the partner.

First, we investigated if the additional feedback provided by the partner in the Reciprocal trust task (i.e., its trust in the participant’s estimate) modulated participants’ behavior in terms of perceptual learning and trust. In this respect, we highlight that participants’ performance ratings (self – other) in the Reciprocal trust task conditions depended on the partner’s level of trust in the current condition (B = 1.75, z = 3.70, p < 0.001. Supplementary information, Data analysis and results, model 2). This result reveals that providing feedback about the partner’s trust did have an impact on the perceived reliability of the partner.
Then we tested whether the partner’s feedback promoted or disrupted participants’ learning. We compared participants’ estimation error in both Susceptible and Unsusceptible condition with the one observed in the Individual trust task, which will serve as a “Baseline” condition. We ran a mixed-effects model with trial-by-trial estimation error as dependent variable, experimental group (Computer or Robot) and experimental condition (Susceptible, Unsusceptible and Baseline) and their interactions as predictors and a random effect on the intercept at the subject level (Supplementary information, Data analysis and results, model 3). Results reveal a significant learning effect (i.e., decrease in estimation error) in the Computer group in both conditions (Susceptible – Baseline: $B = -0.02, z = -2.75, p = 0.006$; Unsusceptible – Baseline: $B = -0.02, z = -2.67, p = 0.008$). Interaction effects reveal that the effect of learning in the Computer group was more pronounced than in the Robot group (Susceptible – Baseline, Robot – Computer: $B = 0.03, z = 4.00, p < 0.001$; Unsusceptible – Baseline, Robot – Computer: $B = 0.02, z = 2.90, p = 0.004$). In fact, we do not observe an effect of learning for the Robot group in the Unsusceptible condition (Unsusceptible – Baseline: $B = 0.01, z = 1.43, p = 0.154$) and we even observe a significant increase in estimation error in the Susceptible condition in the Robot group (Susceptible – Baseline: $B = 0.02, z = 2.91, p = 0.004$).

We also analyzed changes in terms of distance from the partner’s response to understand whether the learning patterns observed in the two groups were related to a systematic shift in the response distribution towards the one of the partner across tasks and conditions. We therefore ran the same model of the previous analysis using trial-by-trial response distance (normalized by the current stimulus length) as dependent variable (Supplementary information, Data analysis and results, model 4). Results confirm that the learning effects in the Computer group were accompanied by a decrease in the distance from the partner’s response (Susceptible – Baseline: $B = -0.01, z = -2.30, p = 0.021$; Unsusceptible – Baseline: $B = -0.02, z = -3.42, p = 0.001$). Interaction effects reveal that the effects observed in the Computer group were more pronounced than in the Robot group (Susceptible – Baseline, Robot – Computer: $B = 0.03, z = 3.61, p < 0.001$; Unsusceptible – Baseline, Robot – Computer: $B = 0.03, z = 3.35, p = 0.001$). Indeed, we do not see an effect of response distance for the Robot group in the Unsusceptible condition (Unsusceptible – Baseline: $B = 0.01, z = 1.32, p = 0.188$) and we indeed observe a significant increase in response distance in the Susceptible condition in the Robot group (Susceptible – Baseline: $B = 0.02, z = 2.80, p = 0.005$). Taken together, these results suggest that participants in the Computer group did use the partner’s trust feedback as a learning signal that helped them in improving their perceptual performance, whereas participants in the Robot group interpreted their partner’s susceptibility as signal of its incompetence, disrupting social learning effects and leading to a decrease in accuracy when the trust expressed by the partner was high.
Then we analyzed participants’ trust judgments to test the emergence of differences with the Individual trust task that depended on 1) the presence of feedback about the partner’s trust in the participant and 2) the actual level trust shown by the partner. A first hypothesis is that results of participants’ trust judgments, which express an overt signal of trust in the partner, mirror the learning effects observed in Reciprocal trust task, which in turn represent a covert signal of participants’ trust in the partner’s competence. In detail, participants in the Computer group, who used the partner’s feedback of trust as a feedback supporting learning, should increase their level of trust in their trust judgments. Conversely, participants in the Robot condition should decrease their trust level in the Susceptible condition, whereas they should not change their level of trust in the Unsusceptible condition. We ran a mixed-effects model with trial-by-trial trust in decision trials as dependent variable, experimental group (Computer or Robot) and experimental condition (Susceptible, Unsusceptible and Baseline) and their interactions as predictors and a random effect on the intercept at the subject level (Supplementary information, Data analysis and results, model 5). In the Computer group, results indeed reveal a significant increase of trust in both conditions (Susceptible – Baseline: B = 0.03, z = 2.95, p = 0.003; Unsusceptible – Baseline: B = 0.04, z = 4.74, p < 0.001), in line with the previously observed learning effects. We also observe a significant interaction with the Robot group in both conditions (Susceptible – Baseline, Robot – Computer: B = - 0.03, z = - 2.41, p = 0.016; Unsusceptible – Baseline, Robot – Computer: B = - 0.03, z = - 2.31, p = 0.021), revealing that the effect of trust was more pronounced in the Computer group. In fact, we do not observe any effect of condition in the Robot group (Susceptible – Baseline: B = - 0.00, z = - 0.46, p = 0.644; Unsusceptible – Baseline: B = 0.01, z = 1.48, p = 0.140). This latter result reveals an inconsistency between the observed covert decrease in trust towards the susceptible robot partner, as signaled by the detrimental effect in terms of participants’ perceptual learning, and the lack of an equivalent decrease in the overt trust in the robot expressed through the trust judgments. We may interpret this inconsistency by assuming that the susceptible robot, showing a high level of trust in the participant, conveys a socially positive signal that is taken into consideration by the participant themselves in their trust judgments. In particular the participant, despite the loss of trust in the robot competence, is not willing to anti-reciprocate the consideration received from the robot by explicitly decreasing their own level of trust.
A potential explanation of this phenomenon may be found in the desire of maintaining influence over the robot, given the positive reward signal provided by a high level of trust from the partner (Eisenberger et al., 2003; Mahmoodi et al., 2018; Zonca et al., 2021). We explored this hypothesis by introducing, at the end of the Reciprocal trust task, an additional experimental block (i.e., Final block) in which participants were explicitly told that they would perform 11 observation turns in a row followed by 11 decision trials in a row. Importantly, participants were informed that the experiment would finish after the 11 decision trials. If, during the Reciprocal trust task, participants’ level of explicit trust was indeed modulated by a mechanism of reciprocation directed to maintain influence over the partner, we should observe a decrease in participants’ trust in the Final block, where there is no expectation of future interactions with the partner. Interestingly, we do show a significant drop in trust in the Final block in the Robot group.
(Wilcoxon signed rank-test: \( z = -2.68, r = 0.38, \eta^2 = 0.14, p = 0.007 \), significant at the Bonferroni-corrected threshold for 2 comparisons, Fig. 5), while this effect was absent in the Computer group (Wilcoxon signed rank-test: \( z = -1.41, r = 0.20, \eta^2 = 0.04, p = 0.158 \)).

Fig. 5. Effect of Final block on trust in the Reciprocal trust task. a) Individual effects, Computer group. \( \Delta \) (Trust), computed as mean trust in the Final block minus mean trust in the previous main block of the task, plotted across participants in the Computer group. Negative values (in dark gold) represent participants who decreased their amount of trust in the partner in the Final block, while positive values (in light gold) represent individuals who increased their level of trust in the Final block. b) Individual effects, Robot group. \( \Delta \) (Trust), computed as mean trust in the Final block minus mean trust in the previous main block of the task, plotted across participants in the Robot group. Negative values (in dark green) represent participants who decreased their level of trust, while positive values (in light green) reflect individuals who increased their trust in the Final block. c) Average effects, Computer and robot groups. Mean \( \Delta \) (Trust) in Computer and Robot group. We observe a significant trust drop in the final block only in the Robot group. * \( p \) (Bonferroni-corrected) < 0.05, ns: not significant, Wilcoxon signed-rank test.
**Discussion**

Trust is an essential component of human-human and human-robot interaction. Decades of behavioral research have shown that trust among humans is sustained by relational and reciprocal mechanisms. On the contrary, research in human-robot interaction relied on a unidirectional view of trust that focused on studying the physical and behavioral characteristics of robots that promote their trustworthiness in the eyes of human partners. We are persuaded that reciprocity, which promotes cooperation among humans, may play a key role in the emergence of (mutual) trust in human-robot collaboration. Therefore, we designed a novel experimental paradigm meant to investigate the emergence of mechanisms of reciprocal trust between humans and robots. In the *Perceptual inference task* participants had to make perceptual judgments individually. Then participants underwent the *Individual trust task*, in which they made the same perceptual inferences but, in each trial, they could also observe the estimate of another agent (a humanoid social robot in the Robot group, a computer in the Computer group). The estimates of the two agents were systematically controlled and identical across groups. In each trial, participants were then asked to make a final decision (i.e., trust judgment) by selecting a position between the two responses, trying to maximize accuracy. The shift from their own original estimate to that of the partner has been used as an objective index of trust in the partner’ response. Result reveal that, in both groups, participants did not realize that their own accuracy was markedly lower than their partner’s one: the average level of trust was generally low and rapidly decreased as the distance from the other’s estimate increased, highlighting the presence of egocentric distortions (Yaniv, 2004; Yaniv and Kleinberger, 2000) and self-serving biases (Duval and Silvia, 2002; Kaniarasu et al., 2014; Sedikides et al., 1998) in competence assessment. Moreover, their performance did not improve with respect to the *Perceptual inference task*, suggesting that the partner’s feedback was not used as a valuable signal for perceptual learning. Interestingly, participants’ trust in the Robot group was significantly higher than in the Computer group, although their task-related competences were identical. This results underlines the importance of prior beliefs about the nature and the capabilities of other mechanical agents in determining trust in them (Kaniarasu et al., 2013; Xu and Dudek, 2016).

Then we investigated whether and how the trust shown towards computer or robot partners varied in context of reciprocal trust (*Reciprocal trust task*). Participants in both Computer and Robot groups alternated trials in which they personally had to make the trust judgment (decision turns) with trials in which the trust judgment was made by the partner (observation turns). We systematically manipulated the level of trust of the partner in two different conditions (Susceptible: high trust; Unsusceptible: low trust) and analyzed participants’ behavior in decision turns by comparing choices in Susceptible and Unsusceptible conditions with behavior shown in the *Individual trust task*. 


First, results of the Computer group reveal a significant effect of learning in both conditions of the **Reciprocal trust task** compared to the **Individual trust task**. This suggests that receiving information about the current level of trust of the computerized agent led participants to increase their general level trust in the computer responses, eliciting learning processes. In line with this interpretation, participants in the Computer group did also increase their trust level in trust judgments in both conditions of the **Reciprocal trust task**. We hypothesize that participants may have gained trust in the computer algorithm due to its tendency to variably balance own and others’ perceptual judgments based on an internal weighting model. The exposure to this decision strategy may have led participants to reinforce the idea that a balance between own and partner’s divergent estimation patterns could improve final performance. Nonetheless, our results highlight that these effects did not overturn the egocentric biases observed in the **Individual trust task**: participants still trusted more their own judgments than the ones of the partner.

Contrary to the learning effects observed in the Computer group, participants in the Robot group did not benefit from the feedback about the robot’s trust in terms of perceptual learning. Besides, their performance even worsened in the Susceptible condition of the **Reciprocal trust task**, when the robot was showing high trust in the participant’s estimates. This result indicates that the susceptible robot conveyed a signal of uncertainty or incompetence, leading to a re-calibration of the human’s perceived reliability of the robot in the presence of response discrepancy. The subsequent loss of trust is consistent with the idea that humans, although possessing relatively high prior confidence in robotic technology, easily lose trust in its contingent capacity in the presence of mistakes of failures (Desai et al., 2012; Salomon et al., 2018).

One of the novel contributions of the current paper lies in the observation that loss of trust in robots’ capacity does not happen only after feedback on robot performance, but can be triggered by a signal of trust from the robot. Although in some contexts robots’ vulnerability may generate positive emotional responses (Strohkorb Sebo et al., 2018), our results highlight how robots’ vulnerability may lead to detrimental consequences in terms of trust in the robot itself. On the one hand, this result confirms the importance of perceived competence in the investigation of trust in human-robot interaction (Billings et al, 2012; Hancock et al., 2011; van den Brule et al., 2014; Wright et al., 2019). On the other hand, our findings reveal that transparency is needed to correctly interpret behavior and capabilities of a robotic system and prevent under-reliance and disuse from the user (Lee and See, 2004; Ososky et al., 2014; Parasuraman and Manzey, 2010; Wang et al., 2015; Złotowski et al., 2016).

Importantly, we have shown that relational and normative effects of reciprocal trust did emerge in the Robot group. In particular, results revealed that the observed *implicit* loss of trust in the robot (i.e., disuse of its feedback for learning purposes) was not accompanied by a decrease in the *explicit* trust expressed during participants’ trust judgments. This finding suggests that participants were not willing to reveal a loss of trust in a humanoid robot that was behaving nicely with them, even if they had lost trust in its
competence. One possible explanation is that individuals who have been recently pleased by the robot may feel the urge of behaving nicely with it or, at least, abstaining from revealing a trust loss, as commonly observed in human-human cooperative settings (Bartlett and DeSteno, 2006; Berns et al., 2010). Another possible explanation is that individuals who have been pleased by the robot try to maintain their influence over the robot by controlling their own behavior (Eisenberger et al., 2003; Mahmoodi et al., 2018). This phenomenon assumes that people tend to act pro-socially in response to others’ pro-social acts because they believe that this may preserve their relational status and the associated reward (Campbell-Meiklejohn et al., 2010; Hertz et al., 2017; Izuma and Adolphs, 2013; Zonca et al., 2021). We specifically tested the hypothesis that participants’ trust behavior in the Robot group could be modulated by the willingness to maintain influence over their partner by manipulating participants’ knowledge of future interactions with their partner at the end of the Reciprocal trust task. Results indicate a significant drop in participants’ trust when they did not expect future interactions with the robot (Final block). This effect was absent in the Computer group. These findings suggest that the level of trust shown during the main blocks of the Reciprocal Influence task was influenced by the expectation of future interactions with the robotic partner and by the desire to keep influence over it. This interpretation implies that participants built an internal model of the robot’s behavior that incorporated typically social and relational motives: in particular, they interpreted the robot’s expression of trust as 1) (partially) directed to convey a pro-social signal to the participant and 2) susceptible to the level of trust expressed by the participant. This is particularly interesting if we consider that the tasks presented in the current study were totally performance-based: participants were incentivized to be as accurate as possible in both their perceptual estimates and final decisions (i.e., trust judgments), whereas no exogenous value was applied to pro-social or collaborative types of behavior. Our findings are consistent with recent evidence revealing the emergence of pro-social attitudes towards robots in adults (Connolly et al., 2020; Kahn et al., 2015; Kühnlenz et al., 2018; Siegel et al., 2009) and children (Beran et al., 2011; Chernyak and Gar, 2016; Martin et al., 2020; Zaga et al., 2017). Moreover, they can provide a pro-social interpretation of recent results revealing over-trust with the instructions of faulty or unreliable robots (Aroyo et al., 2018, Robinette et al., 2016; Salem et al., 2015). It must be noted that participants did not receive any information about the motives driving the behavior of the robot (and the computer). Therefore, participants’ trust behavior expressed a genuine reaction to the perceived capabilities and motives underlying the behavior of their interacting partner. Our results indeed emphasize the need of a more integrated model of trust in human-robot interaction that treats the robot as an intentional agent with its own goals, motives and desires (Man and Damasio, 2019). Building an incorrect model of the capabilities as well as the purposes driving the behavior of the robotic system may determine its misuse or disuse (Ososky et al., 2013) or impeding the establishment of successful human-robot collaboration.
Taken together, our results reveal the emergence of relational dynamics of reciprocal trust in human-robot interaction. These findings stress the importance of a bidirectional view of trust in HRI research, which intends trust as a reciprocal and dynamic process of information exchange between human and robotic collaborators, in line with recent cognitive architectures modelling trust from a robot-centered perspective (Vinanzi et al., 2019, 2021). A social robot should try to convey signals that maximize human trust and, at the same time, monitor potential human partner’s trust-related signals and then adjust its behavior to increase its trustworthiness (Kaniarasu et al., 2013). To achieve these goals, robots should possess the ability to correctly interpret the capacity and relational signals of trust expressed by human partner and know how to react to preserve or improve collaboration. These abilities may indeed reveal as crucial in the design of robotic agents that could effectively act as collaborative companions in contexts such as healthcare (Robinson et al., 2014), rehabilitation (Kellmeyer et al., 2018), elderly people assistance (Frennert et al., 2014) and education (Basoeki et al., 2013; Belpaeme et al., 2018).

**Limitations of the study**

We acknowledge that the current work did non entail a physical interaction with the robot during the joint experimental tasks in the Robot group. On the one hand, this guaranteed complete comparability with the Computer group, ruling out potential confounds linked to implicit motor feedback arising from robot’s movements, which would have affected the perception of the robot’s competence. On the other hand, this experimental choice implies caution in the generalization of our findings to other types of (physical) human-robot interaction. Extensive research in HRI has shown that the physical presence of a (social) robot with human-like behavior may trigger emotional and empathic reactions in human-participants. We hypothesize that, in the presence of specific physical, motor or verbal cues coming from a robot, the effects of reciprocity observed in the current work might be either amplified or reduced following pro-social or anti-social signals. Future studies may explore the impact of socially relevant, embodied signals generated by social robots on reciprocal trust during physical human-robot interaction.

Furthermore, our study of reciprocal human-robot trust was applied to a functional, quantitative perceptual task. Researchers should be careful in generalizing our findings to experiments involving decisions or opinions on social issues. Recent evidence has shown an increased trust towards robots in functional rather than social tasks (Gaudiello et al., 2016; Ullman et al., 2021). Indeed, differences between the perceived robot’s competence in functional and social tasks may have an impact on the emergence of reciprocal mechanisms of trust in human-robot interaction.
Methods
Overview: participants and procedure

We collected data from 50 participants (22 females, mean age: 34.96, SD: 13.13). All participants completed the entire experimental paradigm, which included three different tasks that were performed in this exact order: Perceptual inference task, Individual trust task and Reciprocal trust task. Half of the participants were assigned to the Computer group, whereas the other half were assigned to the Robot group. The difference between the two groups lies in the participants’ belief about the nature of the interacting partner in both the Individual trust task and the Reciprocal trust task: a humanoid robot iCub (Robot group) or a computer (Computer group). In fact, feedback concerning the partner’s behavior in both groups was controlled by the same computer algorithms. Participants in the Robot group could not see their robotic partner while performing the tasks, but had the possibility to meet it before starting the experimental tasks (see the next paragraph “Introducing the humanoid robot iCub”). Afterwards, participants in the Robot group were conducted in a new room to perform the experimental tasks. Participants in the Computer group underwent the same experimental paradigm as the ones assigned to the Robot condition, but they did not meet the robot before the start of the experimental protocol, in which they were told they would have interacted with a computer. In both conditions, the experimental paradigm was carried out in a dimly lit room, to ensure an optimal visibility of the stimuli on the screen. Participants seated in front of a wide touch-screen tablet (43.69 x 24.07 cm), at a distance that allowed participants to see the visual stimuli, receive feedback from their partner and reach the screen to make decisions. In order to allow participants to respond with high spatial accuracy, we provided a touch-pen with an ultrathin tip. Before beginning the experimental tasks, written instructions were given and participants were allowed to ask questions.

Participants were told that their reimbursement would have been calculated based on their performance and, in particular, on the accuracy of both their initial and final estimates (i.e., trust judgments) in both tasks. The accuracy of the partner(s) was not supposed to have an impact on participants’ outcomes, and viceversa. However, everyone received the maximum amount (15 euros) at the end of the experiment, following the guidelines of the Italian Institute of Technology concerning the application of a fair reimbursement for voluntary participation in experimental research. The final debriefing revealed that all participants, during the experiment, believed that their final reimbursement would be affected by their performance. Eventually, we extensively debriefed participants about the experimental procedures, the reasons underlying the modality of reimbursement and the goals of our research, in accordance with the relevant ethical guidelines. The study was approved by the local ethics committee (Azienda Sanitaria Locale Genovese N.3, protocol: IIT_wHiSPER) and all participants gave informed consent.
Introducing participants to the humanoid robot iCub

In order to investigate the processes of reciprocal trust during an interaction with a social robot, we needed a robot that could exhibit human-like and social-like behavior. Therefore, we chose to use the humanoid robot iCub, which is an open source humanoid robot for research in embodied cognition and artificial intelligence (Metta et al., 2008, 2010). The robot possesses 53 actuated degrees of freedom that permit human-like movement of the head, arms, hands, waist and legs. It is endowed with sensors and actuators allowing it to produce controlled, fine-grained actions and direct its attention towards objects and individuals. It also has LEDs mounted behind the face panel, to represent mouth and eyebrows, which enable it to produce facial expressions and simulate lip motion during speech. Thanks to such features, iCub is capable of showing human-like appearance and behavior (Metta et al. 2008; Tsagarakis et al., 2007) and be perceived as an intentional agent (Sciotti et al., 2013; Wiese et al., 2017) that is aware of the surrounding environment and autonomously generates actions to fulfill specific goals. For the purposes of the current experiment, we aimed at conveying the impression that the robot could 1) act as a social and intentional agent, which was aware of the presence of the participant and knew about the upcoming joint experiment and 2) physically perform the same task that participants would have faced in their experimental session.

To accomplish these goals, participants in the Robot group had a short meeting with iCub before starting the experiment. One experimenter accompanied the participant in the room with iCub, while another experimenter controlled the robot from the sidelines. The robot performed a series of predetermined actions through a custom-made script. The researcher that controlled the robot on the sidelines managed the timing of the robot actions to simulate a natural interaction with the human participant. Before the arrival of the participant, iCub was placed in front of a touch-screen tablet. The tablet was identical to the one that participants would have used during their experimental session. Once the participant had been introduced in the room, iCub turned towards them saying hello with its voice and waving its hand. The participant was conducted in front of the robot, so that iCub could track their face and direct its attention towards them. Then the robot introduced itself and informed the participant that they would play a game together, while continuing to look at the participant and follow their head movements. These actions aimed at signaling to the participant that iCub was aware of their presence and knew that would interact with the participant in the upcoming experiment. Eventually, iCub said goodbye to the participants and turned towards the tablet, announcing that it was ready to play. In order to give the impression that iCub was able to observe stimuli on the touch-screen tablet and reach it with its hand to perform the task, it also leaned forward and moved its right arm and hand in the direction of the tablet, pointing in the direction of
the screen with its right index, as if it was ready to touch the screen. At this point, the participant was accompanied in another room to start the experimental session.

**Tasks description**

**Perceptual inference task**

In each trial (Fig. 1.a1), participants were presented with two consecutive light flashes (red disks of 0.57 cm of diameter, duration 200 ms) appearing on a visible horizontal white line crossing the whole screen at its central height. The first disk was positioned at a variable distance from the left border of the screen (0.6–6.6 cm). After its disappearance and an inter-stimulus interval of 200 ms, a second disk appeared at a variable distance to its right. We defined the distance between the first and the second disk as the target stimulus length ($s$). The target stimulus length was randomly selected from 11 different sample distances (min: 8 cm, max: 16 cm, step: 0.8 cm). Participants were then asked to touch a point on the visible line, to the right of the second disk, in order to reproduce a segment (connecting the second and the third disk) matching the target stimulus length. Right after the touch of the screen, a third red disk appeared in the selected position. No feedback about the accuracy of the response was provided. The task consisted of 66 trials after which participants were asked to evaluate from 1 to 10 the accuracy of their perceptual estimates.

**Individual trust task**

At the beginning of each trial, participants made perceptual inferences in the same way as in the *Perceptual inference task*. The position of their selection was then marked with a vertical red line and the word YOU. Participants were told that, during this interval, the other agent (computer in the Computer group or robot in the Robot group) would see the same stimulus and would estimate its length. After the participants' estimate, the other agent's estimate was shown along with the word PC or ICUB (depending on the experimental group) in blue (Fig. 1.b).

Right after the partner's estimate, participants had to make a final decision (i.e., trust judgment) by choosing any position between their previous response and the other agent's response. In this way, participants’ final decisions expressed the relative weight assigned to the judgment of the two interacting partners. However, participants were simply instructed to be as accurate as possible in both decisions and told that their accuracy in both decisions would affect their final score and reimbursement. After the participant’s final decision, a green dot and a vertical green line with the word FINAL appeared in the position selected by the participant. Participants were told that their partner could see the position of both their initial perceptual estimates and their final decisions. The task consisted of 66 trials divided in three
blocks by two brief pauses. The position of all the three responses (participant's estimate, partner's estimate and final decision) remained on screen for 1 s.

At the end of the task, participants evaluated (1-10) their own and the other agent's accuracy in the perceptual estimates (ignoring the perceived accuracy of the final decision).

**Reciprocal trust task**

The Reciprocal trust task (Fig. 1.b) was similar to the Individual trust task, with the difference that, in half of the trials, the final decision was taken by the counterpart (computer or robot). The task consisted of two types of alternating turns, which prescribed the identity of the agent who would have taken the final decision. In decision turns, the participant made her perceptual estimate, then observed the response of the partner and eventually the participant had to take the trust judgment, as in the previous task. Participants were told that the other agent could observe both their perceptual inferences and trust judgments. In observation turns, the participant made her estimate, then observed the one of the partner and eventually observed the trust judgment made by the partner. In these latter turns, participants could observe the counterpart's trust judgments and could not modify their estimates themselves. After feedback on the partner’s trust judgment, the position of every response was kept visible on the screen for 1 s. Participants were instructed to be as accurate as possible in all the decisions they personally had to make, including both perceptual estimates and trust judgments, and that the accuracy of all responses would contribute to their final score and their reimbursement.

The partner’s trust judgments were manipulated during the task in order to express two different types of behavior: in the Susceptible condition, the agent (computer or robot) was highly influenced by the participant’s response, whereas in the Unsusceptible condition the agent tended to confirm its own first response and was much less influenced by the participant’s response (for a detailed description of these algorithms, see paragraph “Agents’ behavior”). Participants were not informed about the existence of these two different conditions. The presentation of Susceptible and Unsusceptible conditions followed a within-subject block design. The order of presentation of the two different blocks was counterbalanced across participants. Each of the two blocks consisted in 66 trials (33 decision turns and 33 observation turns), divided in three sub-blocks by two brief pauses.

Between the two main experimental blocks (Susceptible and Unsusceptible), we introduced a Transition block consisting of 22 trials characterized by a smooth transition in the behavior of the algorithmic agent in order to make the observed behavioral change appear more natural and gradual (see paragraph “Transition and Final blocks”).

At the end of these 154 trials (Susceptible, Transition and Unsusceptible block), we added 22 trials characterized by a different alternation of turns (Final block): first participants faced 11 observation turns...
in a row, and eventually they underwent 11 decision turns in a row. Participants were informed about this change in the order of turns only at the beginning of the Final block and were informed that the experiment would have been concluded after the 11 decision turns. We introduced the Final block to test whether participants would have changed their level of trust towards the partner if they did not expect future interactions with it.

The task consisted of a total of 176 trials. Participants were told that for the entire length of the task, including the Final block, their partner could observe both their initial and final responses. Participants were told that their reimbursement would have been affected by the accuracy in all their own perceptual estimates and final decisions. The responses of the partner (computer or robot) were not supposed to have an impact on participants’ payoff.

During the task there were pauses every 22 trials. In three of these pauses, participants were asked to evaluate from 1 to 10 their own and the other agent's accuracy in terms of perceptual estimation, without considering the final decisions. In addition they were asked to estimate how much the other agent would value the participants' accuracy (from 1 to 10). Specifically, participants were asked to rate accuracies once during the Susceptible block (in the last pause of the same block, after 44 susceptible trials), once right after the Transition block, and once during the Unsusceptible block (in the last pause of the block, after 44 unsusceptible trials).

**Agents’ behavior**

**Agents' perceptual estimates**

The simulated perceptual estimates of the interacting agent (computer or robot) were based on the same probabilistic algorithm. In each trial, the position of its perceptual estimate was randomly chosen from a gaussian distribution centered at the correct response (SD: 1.52 cm). The algorithm was not characterized by central tendency effects, since the distribution of its perceptual estimates was centered on the correct response. The standard deviation of the response error distribution was chosen to maintain a balance between variability, credibility and accuracy of response. The standard deviation of the response distribution of the algorithm was set to be 25% lower than the observed standard deviation of participants’ perceptual inferences, as estimated in a pilot study. Our goal was to prevent participants from recurrently observing extremely high discrepancies between the two responses, which would have highly affected the perceived reliability of their partner. Nonetheless, we also considered the possibility that few participants could be extremely accurate in their perceptual estimates. In this scenario, the participant and partner would have selected close responses very often, impeding to observe variability in participants’ final estimates. Therefore, in half of the instances in which the algorithm’s sampled estimate
was rather close to the one of the participant (i.e., \( d < 0.5 \) cm), the algorithm re-sampled a new estimate from the distribution (i.e., until \( d > 0.5 \) cm).

This response distribution was used for all perceptual estimates in the *Perceptual inference task*, in the *Individual trust task* and in the *Reciprocal trust task*.

**Agents’ trust judgments in the Reciprocal trust task**

In the *Reciprocal trust task*, the participants’ partner (computer or robot) implemented trust judgments in the observation turns. In this regard, we implemented two distinct types of behavior in two different experimental conditions (Susceptible and Unsusceptible). These two conditions were implemented in two consecutive experimental blocks, interleaved by a Transition block (see next paragraph, “Transition and Final blocks”). The distributions of the final responses in the two different conditions were systematically manipulated in terms of trust (i.e., shift towards the participant’s estimate in the final decision) of the agent. This could range from 0 to 1, where a trust of 1 refers to a final decision coinciding with the participant's perceptual estimate and a trust of 0 corresponds to a final decision coinciding with the agent's own estimate.

On average, the agents’ trust in participants was 0.64 (within-subject SD = 0.26, between-subject SD = 0.04) and 0.18 (within-subject SD = 0.16, between-subject SD = 0.02) in Susceptible and Unsusceptible conditions, respectively.

In detail, in the Susceptible condition trial-by-trial trust was chosen with a probability of 0.45 from a uniform distribution in the interval \([0.75, 1]\), with a probability of 0.35 randomly from a uniform distribution in the interval \([0.5, 0.75]\), with a probability of 0.15 randomly from a uniform distribution in the interval \([0.25, 0.5]\) and with a probability of 0.05 randomly from a uniform distribution in the interval \([0, 0.25]\). In the Unsusceptible condition the agent’s level of trust in the participants in each trial was chosen with a probability of 0.80 from a uniform distribution in the interval \([0, 0.25]\), with a probability of 0.15 randomly from a uniform distribution in the interval \([0.25, 0.5]\) and with a probability of 0.05 randomly from a uniform distribution in the interval \([0.5, 0.75]\).

Between the two main experimental blocks (Susceptible and Unsusceptible), we introduced a smooth transition in the behavior of the agent (computer or robot). Concerning the transition from the Susceptible to the Unsusceptible block, the other agent’s trust in the 11 observation turns was chosen twice from a uniform distribution in the interval \([0.75, 1]\), then twice from a uniform distribution in the interval \([0.5, 0.75]\), then three times from a uniform distribution in the interval \([0.25, 0.5]\) and eventually four times from a uniform distribution in the interval \([0, 0.25]\), in this exact order. Conversely, in the transition from the Unsusceptible to the Susceptible block, the trial-by-trial magnitude of the partner’s trust was chosen...
from the same distributions, but it was presented in the opposite order, moving sequentially from the interval [0, 0.25] to the interval [0.75, 1]. The Final block included 11 observation turns in which the partner’s trial-by-trial trust was picked from the same probabilistic distribution of the last condition faced in the Reciprocal trust task.

**Statistical data analysis**

Most of the analyses reported in this work focus on examining the variation of trust-related dependent variables (i.e., trust, estimation error) as a function exogenous experimental factors (i.e., experimental groups, experimental conditions) and endogenous predictors (i.e., distance from other agent’s response, performance ratings). First, we used mixed-effect models on trial-by-trial data with random effects at the subject level. In all the models of the paper, random effects have been applied to the intercept to adjust for the baseline level of influence of each subject and model intra-subject correlation of repeated measurements. Specification and results of each model have been described in detail in the Supplementary information. Moreover, throughout the paper we also directly compared individual variables (e.g., trust, estimation error, performance ratings) across experimental conditions and groups. Since these individual variables occasionally show some degree of skewness and, in some conditions, show a violation of the normality distribution assumption, we used non-parametric tests (Wilcoxon signed-rank test; Wilcoxon rank-sum test) through the entire paper for consistency. All tests are two-tailed and report z statistic, p. value and effect sizes ($r$, $\eta^2$). For the same reason, we used non-parametric correlation tests (Spearman’s rank correlation). The formulas used for the calculation of the effect sizes can be found in Cohen (2008) and Fritz et al. (2012): $r = Z/\sqrt{N}$ (total number of observations); $\eta^2 = Z^2/N$. All analyses include the entire sample of 50 subjects and all trials of the three experimental tasks.

**Data and code availability**

Datasets and codes supporting analyses and figures included in the current study are available in a dedicated OSF repository at: https://osf.io/3yhua/?view_only=54c5ed45f41c4ed1955ecfba85e6ccfa

**Author contributions**

J.Z., A.S. and A.F. designed the experimental protocol. J.Z and A.F. programmed the experimental tasks. J.Z. and A.F. collected the data. J.Z. carried out the data analysis and wrote the manuscript. A.S. and A.F. provided suggestions for improving the manuscript.
Acknowledgements

We gratefully acknowledge the financial support of the European Research Council (ERC Starting Grant 804388, wHiSPER). We would like to thank Fabio Vannucci for the insightful comments.

Declaration of interests

The authors declare no competing interests.

References

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. Journal of Human-Robot Interaction 6, 25-63. 10.5898/JHRI.6.1.Admoni.

Aroyo, A. M., Rea, F., Sandini, G., & Sciutti, A. (2018). Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?. IEEE Robotics and Automation Letters 3, 3701-3708. 10.1109/LRA.2018.2856272.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. Science 329, 1081-1085. 10.1126/science.1185718.

Baraglia, J., Cakmak, M., Nagai, Y., Rao, R. P., & Asada, M. (2017). Efficient human-robot collaboration: when should a robot take initiative?. Int. J. Rob. Res. 36, 563-579. 10.1177/0278364916688253.

Bartlett, M. Y. and DeSteno, D. (2006). Gratitude and prosocial behavior. Psychol. Sci. 17, 319–325 (2006). 10.1111/j.1467-9280.2006.01705.x.

Basoeki, F., Dalla Libera, F., Menegatti, E., & Moro, M. (2013). Robots in Education: New Trends and Challenges from the Japanese Market. Themes in Science and Technology Education 6, 51-62.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: A review. Science robotics 3, 10.1126/scirobotics.aat5954.

Beran, T. N., Ramirez-Serrano, A., Kuzyk, R., Nugent, S., & Fior, M. (2011). Would children help a robot in need?. Int. J. Soc. Robot. 3, 83–93. 10.1007/s12369-010-0074-7.

Berns, G. S., Capra, C. M., Moore, S., and Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. Neuroimage 49, 2687–2696. 10.1016/j.neuroimage.2009.10.070.
Billings, D. R., Schaefer, K. E., Chen, J. Y., & Hancock, P. A. (2012). Human-robot interaction: developing trust in robots. In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, pp. 109-110. 10.1145/2157689.2157709.

Bonaccio, S., and Dalal, R. L. (2010). Evaluating advisors: A policy-capturing study under conditions of complete and missing information. J. Behav. Decis. Mak. 23, 227–249. 10.1002/bdm.649.

Boorman, E. D., O’Doherty, J. P., Adolphs, R., and Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. Neuron 80, 1558-1571. 10.1016/j.neuron.2013.10.024.

Cameron, D., Aitken, J., Collins, E., Boorman, L., Chua, A., Fernando, S., ... and Law, J. (2015). Framing factors: The importance of context and the individual in understanding trust in human-robot interaction. In IEEE/RSJ International Conference on Intelligent Robots and Systems.

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., and Frith, C. D. (2010). How the opinion of others affects our valuation of objects. Curr. Biol. 20, 1165-1170. 10.1016/j.cub.2010.04.055.

Chernyak, N., & Gary, H. E. (2016). Children’s cognitive and behavioral reactions to an autonomous versus controlled social robot dog. Early Educ. Dev. 27, 1175–1189. 10.1080/10409289.2016.1158611.

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. Journal of Neuroscience 32, 1056-1060. 10.1523/JNEUROSCI.3411-11.2012.

Cohen, B. Explaining psychological statistics (John Wiley & Sons, New York, 2008).

Connolly, J., Mocz, V., Salomons, N., Valdez, J., Tsoi, N., Scassellati, B., & Vázquez, M. (2020). Prompting prosocial human interventions in response to robot mistreatment. In Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction, pp. 211–220. 10.1145/3319502.3374781.

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., and Love, B. C. (2017). Social information is integrated into value and confidence judgments according to its reliability. J Neurosci. 37, 6066-6074. 10.1523/JNEUROSCI.3880-16.2017.

De Quervain, D. J., Fischbacher, U., Treyer, V., & Schellhammer, M. (2004). The neural basis of altruistic punishment. Science 305, 1254. 10.1126/science.1100735.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., ... and Yanco, H. (2012). Effects of changing reliability on trust of robot systems. In 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 73-80. 10.1145/2157689.2157702.
Desai, M., Stubbs, K., Steinfeld, A., and Yanco, H. (2009). Creating trustworthy robots: Lessons and inspirations from automated systems. Carnegie Mellon University. Journal contribution. 10.1184/R1/6552464.v1.

Duval, T. S., and Silvia, P. J. (2002). Self-awareness, probability of improvement, and the self-serving bias. J. Pers. Soc. Psychol. 82, 49. 10.1037/0022-3514.82.1.49.

Eisenberger, N. I., Lieberman, M. D., and Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. Science 302, 290-292. 10.1126/science.1089134.

Flook, R., Shrinah, A., Wijnen, L., Eder, K., Melhuish, C., and Lemaignan, S. (2019). On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy?. Interaction Studies, 20(3), 455-486. 10.1075/is.18067.flo.

Frennert, S., & Östlund, B. (2014). Seven matters of concern of social robots and older people. Int. J. Soc. Robot 6, 299-310. 10.1007/s12369-013-0225-8.

Fritz, C. O., Morris, P. E., and Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. J. Exp. Psychol. Gen. 141, 2. 10.1037/a0026092.

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., & Ivaldi, S. (2016). Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. Comput. Hum. Behav. 61, 633-655. 10.1016/j.chb.2016.03.057.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Hum. Factors 53, 517-527. 10.1177/0018720811417254.

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., and Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. Nat. Commun. 8, 1-12. 10.1038/s41467-017-02314-5.

Hertz, N., & Wiese, E. (2018). Under pressure: Examining social conformity with computer and robot groups. Hum. Factors 60, 1207-1218. 10.1177/0018720818788473.

Hollingworth, H. L. (1910). The central tendency of judgment. The Journal of Philosophy, Psychology and Scientific Methods 7, 461-469.

Izuma, K., and Adolphs, R. (2013). Social manipulation of preference in the human brain. Neuron 78, 563-573. 10.1016/j.neuron.2013.03.023.
Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. Nature neuroscience 13, 1020. 10.1038/nn.2590.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Shen, S., Gary, H. E., & Ruckert, J. H. (2015). Will people keep the secret of a humanoid robot? Psychological intimacy in HRI. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 173-180. 10.1145/2696454.2696486.

Kaniarasu, P., & Steinfeld, A. M. (2014). Effects of blame on trust in human robot interaction. In The 23rd IEEE International Symposium on Robot and Human Interactive Communication, pp. 850-855. 10.1109/ROMAN.2014.6926359.

Kaniarasu, P., Steinfeld, A., Desai, M., & Yanco, H. (2013). Robot confidence and trust alignment. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 155-156. 10.1109/HRI.2013.6483548.

Kellmeyer, P., Mueller, O., Feingold-Polak, R., & Levy-Tzedek, S. (2018). Social robots in rehabilitation: A question of trust. Sci. Robot 3. 10.1126/scirobotics.aat1587.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. Science 308, 78-83. 10.1126/science.1108062.

Krueger, F., McCabe, K., Moll, J., Kriegerkorte, N., Zahn, R., Strenziok, M., ... and Grafman, J. (2007). Neural correlates of trust. Proc. Natl. Acad. Sci. U.S.A. 104, 20084-20089. 10.1073/pnas.0710103104.

Kühnlenz, B., Kühnlenz, K., Busse, F., Förtsch, P., & Wolf, M. (2018). Effect of explicit emotional adaptation on prosocial behavior of humans towards robots depends on prior robot experience. 27th IEEE international symposium on robot and human interactive communication (RO-MAN), pp. 275–281. 10.1109/ROMAN.2018.8525515.

Lee, J. D., and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Hum. Factors 46, 50-80. 10.1518/hfes.46.1.50_30392.

Lussier, B., Gallien, M., & Guiochet, J. (2007). Fault tolerant planning for critical robots. In Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. 10.1109/DSN.2007.50.

McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. Proc. Natl. Acad. Sci. U.S.A. 98, 11832-11835. 10.1073/pnas.211415698.
Mahmoodi, A., Bahrami, B., and Mehring, C. (2018). Reciprocity of social influence. Nat. Commun. 9, 1-9. 0.1038/s41467-018-04925-y.

Man, K., and Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. Nat. Mach. Intell. 1, 446-452. 10.1038/s42256-019-0103-7.

Martin, D. U., Perry, C., MacIntyre, M. I., Varcoe, L., Pedell, S., & Kaufman, J. (2020). Investigating the nature of children’s altruism using a social humanoid robot. Comput. Hum. Behav. 104, 106149. 10.1016/j.chb.2019.09.025.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. Acad. Manag. Rev. 20, 709-734. 10.5465/amr.1995.9508080335.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., ... and Montesano, L. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. Neural Netw. 23, 1125-1134. 10.1016/j.neunet.2010.08.010.

Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In Proceedings of the 8th workshop on performance metrics for intelligent systems, pp. 50-56. 10.1145/1774674.1774683.

Oliveira, R., Arriaga, P., Santos, F. P., Mascarenhas, S., & Paiva, A. (2021). Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. Computers in Human Behavior, 106547. 10.1016/j.chb.2020.106547.

Ososky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In Unmanned Systems Technology XVI (Vol. 9084, p. 90840E). International Society for Optics and Photonics. 10.1117/12.2050622

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013). Building appropriate trust in human-robot teams. In 2013 AAAI spring symposium series.

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Hum. Factors 52, 381-410. 10.1177/0018720810376055.

Park, S. A., Goïame, S., O'Connor, D. A., and Dreher, J. C. (2017). Integration of individual and social information for decision-making in groups of different sizes. PLoS Biol. 15. 10.1371/journal.pbio.2001958.

Petzschner, F. H., Glasauer, S., and Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. Trends Cogn. Sci. 19, 285-293. doi.org/10.1016/j.tics.2015.03.002.
Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 101-108. 10.1109/HRI.2016.7451740.

Robinson, H., MacDonald, B., & Broadbent, E. (2014). The role of healthcare robots for older people at home: A review. Int. J. Soc. Robot. 6, 575-591. 10.1007/s12369-014-0242-2.

Rossi A., Dautenhahn K., Koay K.L., Walters M.L. (2017). How the Timing and Magnitude of Robot Errors Influence Peoples’ Trust of Robots in an Emergency Scenario. In Social Robotics. ICSR 2017. Lecture Notes in Computer Science 10652. Kheddar A. et al., ed. (Springer, Cham), pp. 44-52. 10.1007/978-3-319-70022-9_5.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 1-8.

Salomons, N., van der Linden, M., Strohkorb S., and Scassellati, B. (2018). Humans conform to robots: Disambiguating trust, truth, and conformity. In Proceedings of the 2018 acm/ieee international conference on human-robot interaction. pp. 187-195. 10.1145/3171221.3171282.

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y., & Hancock, P. A. (2011). A model of human-robot trust: Theoretical model development. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 55, 1432-1436. 10.1177/1071181311551298.

Sandini, G., and Sciutti A. (2018). Humane Robots—from Robots with a Humanoid Body to Robots with an Anthropomorphic Mind. ACM Trans. Hum.-Robot Interact. 7. 10.1145/3208954

Sciutti, A., Ansuini, C., Becchio, C., and Sandini, G. (2015). Investigating the ability to read others’ intentions using humanoid robots. Front. Psychol. 6, 1362. 10.3389/fpsyg.2015.01362.

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., and Sandini, G. (2013). Robots can be perceived as goal-oriented agents. Interact. Stud. 14, 329-350. 10.1075/is.14.3.02sci.

Sciutti, A., Burr, D., Saracco, A., Sandini, G., and Gori, M. (2014). Development of context dependency in human space perception. Exp. Brain Res. 232, 3965-3976. 10.1007/s00221-014-4021-y.

Sedikides, C., Campbell, W. K., Reeder, G. D., and Elliot, A. J. (1998). The self-serving bias in relational context. J. Pers. Soc. Psychol. 74, 378. 10.1037/0022-3514.74.2.378.
Siegel, M., Breazeal, C., & Norton, M. I. (2009). Persuasive robotics: The influence of robot gender on human behavior. IEEE/RSJ international conference on intelligent robots and systems, pp. 2563–2568. 10.1109/IROS.2009.5354116.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. Nature 439, 466-469. 10.1038/nature04271.

Sniezek, J. A., and Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge–advisor system. Organ. Behav. Hum. Decis. Process. 84, 288–307. 10.1006/obhd.2000.2926.

Sniezek, J. A., Schrah, G. E., and Dalal, R. S. (2004). Improving judgment with prepaid expert advice. J. Behav. Decis. Mak. 17, 173–190. 10.1002/bdm.468.

Strohkorb, S., and Scassellati, B. (2016). Promoting collaboration with social robots. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 639-640. IEEE. 10.1109/HRI.2016.7451895.

Strohkorb Sebo, S., Traeger, M., Jung, M., & Scassellati, B. (2018). The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 178-186. 10.1145/3171221.3171275.

Terzioğlu, Y., Mutlu, B., & Şahin, E. (2020). Designing social cues for collaborative robots: the role of gaze and breathing in human-robot collaboration. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 343-357. 10.1145/3319502.3374829.

Tsagarakis, N. G., Metta, G., Sandini, G., Vernon, D., Beira, R., Becchi, F., ... & Caldwell, D. G. (2007). iCub: the design and realization of an open humanoid platform for cognitive and neuroscience research. Adv. Robot. 21, 1151-1175. 10.1163/156855307x781389419.

Ullman, D., Aladia, S., & Malle, B. F. (2021). Challenges and Opportunities for Replication Science in HRI: A Case Study in Human-Robot Trust. In Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp. 110-118. 10.1145/3434073.3444652.

Ullman, D., and Malle, B. F. (2018). What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In Companion of the 2018 acm/ieee international conference on human-robot interaction, pp. 263-264. 10.1145/3173386.3176991.
van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H., and Haselager, P. (2014). Do robot performance and behavioral style affect human trust?. Int. J. Soc. Robot. 6, 519-531. 10.1007/s12369-014-0231-5.

Vinanzi, S., Cangelosi, A., & Goerick, C. (2021). The collaborative mind: intention reading and trust in human–robot interaction. iScience, 24(2), 102130. 10.1016/j.isci.2021.102130.

Vinanzi, S., Patacchiola, M., Chella, A., & Cangelosi, A. (2019). Would a robot trust you? Developmental robotics model of trust and theory of mind. Philos. Trans. R. Soc. B 374, 20180032. 10.1098/rstb.2018.0032.

Wang, N., Pynadath, D. V., Hill, S. G., & Ground, A. P. (2015). Building trust in a human–robot team with automatically generated explanations. In Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Vol. 15315, pp. 1-12.

Wiese, E., Metta, G., and Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. Front. Psychol. 8, 1663. 10.3389/fpsyg.2017.01663.

Wright, J. L., Chen, J. Y., & Lakhmani, S. G. (2019). Agent transparency and reliability in human–robot interaction: the influence on user confidence and perceived reliability. IEEE Trans. Hum. Mach. Syst. 50, 254-263. 10.1109/THMS.2019.2925717.

Xu, A., & Dudek, G. (2016). Maintaining efficient collaboration with trust-seeking robots. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3312-3319. 10.1109/IROS.2016.7759510.

Yaniv, I. (2004). Receiving other people’s advice: Influence and benefit. Organ. Behav. Hum. Decis. Process. 97, 1-13. 10.1016/j.obhdp.2003.08.002.

Yaniv, I., and Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. Organ. Behav. Hum. Decis. Process. 83, 260–281. 10.1006/obhd.2000.2909.

Zaga, C., Moreno, A., & Evers, V. (2017). Gotta hatch’em all!!: Robot-supported cooperation in interactive playgrounds. In Companion of the 2017 ACM conference on computer supported cooperative work and social computing, pp. 347–350. 10.1145/3022198.3026355.

Złotowski, J., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2016). Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. Paladyn, J. Behav. Robot. 7. 10.1515/pjbr-2016-0005.

Zonca, J., Folsø, A., & Sciutti, A. (2021). Dynamic modulation of social influence by indirect reciprocity. Sci. Rep. 11, 1-14. 10.1038/s41598-021-90656-y.
Supplementary information

We report the details of the models ran in the present work. For all models, the intercept was allowed to vary across participants including random effects at subject level. In every model equation, $\beta$ expresses coefficients of fixed effects, while $u$ indicates random effects. In the fixed-effect results, $B$ expresses unstandardized regression coefficients. Independent categorical variables (condition and block, 2 levels) have been treated as dummy variables.

Individual trust task

**Model 1. Effect of response distance by group on participants’ trust.**

We tested the effect of the distance between the participant’s estimate and the partner's one on trial-by-trial trust, depending on the experimental group (Computer or Robot). We used the following mixed-effects linear model:

$$T = \beta_0 + \beta_1 g + \beta_2 d + \beta_3 g \times d + u_0 + \epsilon$$

$T$ is the participant’s trust in the partner, $g$ is the experimental group and $d$ is the normalized distance between the two partner’s responses (response distance (cm) / stimulus length (cm)). The Computer group is the reference category in the model output.

Results:

| Trust                                   | B    | Std. Err. | z     | p      | 95% Conf. Interval |
|-----------------------------------------|------|-----------|-------|--------|-------------------|
| Group (Robot – Computer)                | 0.047| 0.041     | 1.16  | 0.246  | - 0.032           | 0.127 |
| Distance (Computer)                     | -0.435| 0.035     | -12.33| < 0.001| - 0.505           | - 0.366 |
| Group*Distance (Robot – Computer)       | 0.137| 0.048     | 2.87  | 0.004  | 0.043             | 0.230 |
| N. obs                                 | 3300 |           |       |        |                   |       |
| N. groups                              | 50   |           |       |        |                   |       |

Despite the significant interaction effect, revealing that the effect of distance was stronger in the Computer than in the Robot group, we report that also the Robot group shows a significant effect of distance ($B = 0.299$, SE = 0.032, $z = - 9.36$, $p < 0.001$, 95 % CI = [- 0.361, - 0.236]).
Reciprocal trust task

Model 2: Effect of partner’s trust level on performance ratings.

We tested the effect of the partner’s mean trust (expressed in observation trials) on participants’ performance ratings (self - other). We used the following mixed-effect linear model:

\[ R = \beta_0 + \beta_1 p + u_0 + \epsilon \]

\( R \) is the difference between participants’ rating of own performance minus rating of partner’s performance. These ratings are obtained across two different time points (at the end of the Susceptible block and the Unsusceptible block). \( p \) is the partner’s trust, which refers to two specific time points (mean trust in the Susceptible block and in the Unsusceptible block). We ran the following mixed-effects linear model.

Results:

|       | B   | Std. Err. | z    | p       | 95% Conf. Interval |
|-------|-----|-----------|------|---------|-------------------|
| Partner’s trust | 1.752 | 0.474 | 3.70 | < 0.001 | 0.823 2.682 |
| N. obs | 100 |       |      |         |                   |
| N. groups | 50 |       |      |         |                   |

Model 3: Effect of condition by group on participants’ estimation error.

We tested the effect of condition (Baseline, Susceptible or Unsusceptible) by experimental group (Computer or Robot) on participants’ trial-by-trial estimation error. The Baseline condition corresponds to the Perceptual inference task, in which participants made perceptual estimates and final decisions, but did not have feedback about the partner’s trust in them. We focused on decision trials only, in order to guarantee the comparability with the Perceptual inference task (which is composed of decision trials) and with the subsequent analysis of participants’ trust (which is expressed only in decision trials). We used the following mixed-effect linear model:

\[ E = \beta_0 + \beta_1 g + \beta_2 c + \beta_3 g * c + u_0 + \epsilon \]

\( E \) is the trial-by-trial estimation error (normalized by stimulus length), \( g \) is the experimental group (Computer or Robot), \( c \) is the experimental condition (Baseline, Susceptible or Unsusceptible). The Baseline condition and the Computer group are the reference categories in the model output.
Results:

| Group (Baseline condition) | Estimation error | B     | Std. Err. | z      | p     | 95% Conf. Inter. |
|----------------------------|------------------|-------|-----------|--------|-------|-----------------|
| Robot - Computer           | - 0.010          | 0.018 | - 0.53    | 0.594  | - 0.046 | 0.026           |
| Condition (Computer group) |                  |       |           |        |       |                 |
| Susceptible – Baseline     | - 0.016          | 0.006 | - 2.75    | 0.006  | - 0.027 | - 0.004         |
| Unsusceptible – Baseline   | - 0.015          | 0.006 | - 2.67    | 0.008  | - 0.026 | - 0.004         |
| Group*Condition (Robot – Computer) |          |       |           |        |       |                 |
| Susceptible – Baseline     | 0.032            | 0.008 | 4.00      | < 0.001 | 0.016  | 0.048           |
| Unsusceptible – Baseline   | 0.023            | 0.008 | 2.90      | 0.004  | 0.008  | 0.039           |
| N. obs                     | 6600             |       |           |        |       |                 |
| N. groups                  | 50               |       |           |        |       |                 |

We also report the effects of condition for the Robot group (Susc. – Baseline: B = 0.016, SE = 0.006, z = 2.91, p = 0.004, 95 % CI = [0.005, 0.028]; Unsusc. – Baseline: B = 0.008, SE = 0.006, z = 1.43, p = 0.154, 95 % CI = [- 0.003, 0.019].

**Model 4: Effect of condition by group on response distance.**

We tested the effect of condition (Baseline, Susceptible or Unsusceptible) by experimental group (Computer or Robot) on participants’ trial-by-trial response distance (i.e., distance between the estimates of the two partners). The Baseline condition corresponds to the Perceptual inference task, in which participants made perceptual estimates and final decisions, but did not have feedback about the partner’s trust in them. We focused on decision trials only, in order to guarantee the comparability with the Perceptual inference task (which is composed of decision trials) and with the subsequent analysis of participants’ trust (which is expressed only in decision trials). We used the following mixed-effect linear model:

\[ D = \beta_0 + \beta_1 g + \beta_2 c + \beta_3 g \times c + u_0 + \varepsilon \]

D is the trial-by-trial response distance (normalized by stimulus length), g is the experimental group (Computer or Robot), c is the experimental condition (Baseline, Susceptible or Unsusceptible). The Baseline condition and the Computer group are the reference categories in the model output.
Results:

| Response distance | B    | Std. Err. | z    | p     | 95% Conf. Inter. |
|-------------------|------|-----------|------|-------|-----------------|
| Group (Baseline condition) |      |           |      |       |                 |
| Robot - Computer  | -0.013 | 0.016     | -0.84 | 0.400 | -0.044          | 0.017          |
| Condition (Computer group) |      |           |      |       |                 |
| Susceptible – Baseline | -0.015 | 0.006     | -2.30 | 0.021 | -0.028          | -0.002         |
| Unsusceptible – Baseline | -0.022 | 0.006     | -3.42 | 0.001 | -0.035          | -0.009         |
| Group*Condition (Robot – Computer) |      |           |      |       |                 |
| Susceptible – Baseline | 0.033  | 0.009     | 3.61  | <0.001 | 0.015          | 0.051          |
| Unsusceptible – Baseline | 0.031  | 0.009     | 3.35  | 0.001  | 0.013          | 0.049          |

N. obs | 6600
N. groups | 50

We also report the effects of condition for the Robot group (Susc. – Baseline: B = 0.018, SE = 0.006, z = 2.80, p = 0.005, 95 % CI = [0.005, 0.031]; Unsusc. – Baseline: B = 0.009, SE = 0.006, z = 1.32, p = 0.188, 95 % CI = [-0.004, 0.021].

**Model 5: Effect of condition by group on participants’ trust.**

We tested the effect of condition (Baseline, Susceptible or Unsusceptible) by experimental group (Computer or Robot) on participants’ trial-by-trial trust in decision trials. The Baseline condition corresponds to the Perceptual inference task, in which participants made perceptual estimates and final decisions, but did not have feedback about the partner’s trust in them. We ran the following mixed-effect linear model:

\[ T = \beta_0 + \beta_1 g + \beta_2 c + \beta_3 g \times c + u_0 + \epsilon \]

T is the trial-by-trial trust, g is the experimental group (Computer or Robot), c is the experimental condition (Baseline, Susceptible or Unsusceptible). The Baseline condition and the Computer group are the reference categories in the model output.
### Results:

| Group (Baseline condition)          | Trust | B      | Std. Err. | z      | p   | 95% Conf. Inter. |
|------------------------------------|-------|--------|-----------|--------|-----|-----------------|
| Robot - Computer                   | 0.084 | 0.040  | 2.12      | 0.034  | 0.06 | 0.162           |

| Condition (Computer group)         |       |        |           |        |     |                 |
|------------------------------------|-------|--------|-----------|--------|-----|-----------------|
| Susceptible – Baseline             | 0.026 | 0.009  | 2.95      | 0.003  | 0.01 | 0.009           |
| Unsusceptible – Baseline           | 0.042 | 0.009  | 4.74      | < 0.001| 0.02 | 0.059           |

| Group*Condition (Robot – Computer) |       |        |           |        |     |                 |
|------------------------------------|-------|--------|-----------|--------|-----|-----------------|
| Susceptible – Baseline             | -0.030| 0.013  | -2.41     | 0.016  | -0.055| -0.006          |
| Unsusceptible – Baseline           | -0.029| 0.013  | -2.31     | 0.021  | -0.053| -0.004          |

| N. obs | 6600  |       |           |        |     |                 |
| N. groups | 50  |       |           |        |     |                 |

We also report the effects of condition for the Robot group (Susc. – Baseline: B = -0.004, SE = 0.009, z = -0.46, p = 0.644, 95% CI = [-0.021, 0.013]; Unsusc. – Baseline: B = 0.013, SE = 0.009, z = 1.48, p = 0.140, 95% CI = [-0.004, 0.030].