"I think you are doing a bad job!"

The Effect of Blame Attribution by a Robot in Human-Robot Collaboration

Diede P.M. van der Hoorn
Utrecht University
Utrecht, The Netherlands
d.p.m.vanderhoorn@students.uu.nl

Anouk Neerincx
Utrecht University
Utrecht, The Netherlands
a.neerincx@uu.nl

Maartje M.A. de Graaf
Utrecht University
Utrecht, The Netherlands
m.m.a.degraaf@uu.nl

ABSTRACT

Robots will increasingly collaborate with human partners necessitating research into how robots negotiate negative collaborative outcomes. This study investigates the effect of blame attribution on trust assessments in human-robot collaboration. Participants (n = 60) collaboratively played a game with a humanoid robot in one of four conditions in a 2 (blame correctness: correct vs. incorrect) by 2 (blame target: human vs. robot) between-subjects experiment. Results show that people evaluate a robot more positively when it blames itself for collaborative failures, especially, it seems, in the case of incorrect self-blame. Our findings indicate a need to further research on effective communication strategies for robots that need to negotiate collaborative failures without compromising the trust relationships with its human partner.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**; • **Computer systems organization → Robotics**.

KEYWORDS

human-robot collaboration, blame attribution, trust, human-robot interaction, communication strategies

ACM Reference Format:
Diede P.M. van der Hoorn, Anouk Neerincx, and Maartje M.A. de Graaf. 2021. "I think you are doing a bad job!": The Effect of Blame Attribution by a Robot in Human-Robot Collaboration. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3434073.3444681

1 INTRODUCTION

With robots steadily moving into human social spaces, human-robot collaborations increasingly become everyday practice [20] and the robot's contribution to collaborative tasks may come with increased responsibility [17]. The increasingly autonomous and independent role of robot partners in human-robot collaborations results in gradual ambiguity of who is responsible for success or failure [21]. When robots are fulfilling the role of a social actor, their communication strategies for how to allocate and negotiate potential negative collaborative outcomes with their human partners become increasingly important [16]. Trust is essential for good collaborations in the long run and vulnerable communication (e.g., about collaborative failures) is important for the development of trust [10]. Yet, little research has been conducted on how robots should communicate information that may harm or repair the trust relationship in human-robot collaborations [6].

Trust is an important factor for successful human-robot collaborations [17] as it determines people's willingness to work with the robot in future endeavors [50]. People's trust in robots is affected by robot-related factors [42] including the content of the robot's verbal interaction [25]. Given that disagreement and blame attribution is inevitable during interpersonal communication [3, 4], we need a better understanding of how blame attribution by the robot may affect trust in human-robot collaborations for a successful introduction of robots into human social spaces.

Psychology research shows that criticism after failure makes a person be perceived as more capable, whereas praise after success results in perceived incapability [48]. Moreover, self-serving bias tells us that people tend to take credit for success but like to blame others or the situation in case of failures [33]. This indicates a potential preference for a robot that is willing to take the blame for collaborative failure. However, previous HRI works also show that erroneous robot behaviors (incorrect blame can be seen as such) negatively impacts trust [42].

Our study aims to provide initial clarification of these inconsistent findings and further adds to existing knowledge on blame assignment in HRI setting by combining blame target with blame correctness dimension (as previous works have only researched these two dimensions separately, e.g., [16]). Hence, this study investigates how blame correctness (correct vs. incorrect) and blame target (human vs. robot) affects trust in human robot collaboration. For this, we developed a collaborative game that can be played by a human-robot team. The human and the robot had to assess different images, taking turns and subsequently collaboratively, always aiming at the best team score (i.e., a high ranking). The robot (in)correctly blamed itself or the participant, when the team score was ranked after the first session.

2 THEORETICAL BACKGROUND

2.1 Trust in Human-Robot Interaction

Definitions on the concept of trust generally include three components [19]: an agent (the *truster*) who is willing to rely on the actions of another agent (the *trustee*) and therefore willingly abandons control over the outcomes (*something at stake*). Disputes on the exact

HRI '21, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8289-2/21/03…$15.00
https://doi.org/10.1145/3434073.3444681

definition of trust mainly include whether trust is considered to be a belief, an attitude, an intention, or a behavior [23]. Lee and See [23] use a framework developed by Fishbein and Ajzen [12] to combine these different views. In their model, trust affects behavior as an attitude and the authors propose the following definition of trust: "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [23], p.54. This definition is widely used in empirical studies on automation [13].

Although trust in human-robot interaction is closely related to trust in automation, there are two important differences [8]. First, contrasting general automation, robots have a physical embodiment which enables manipulation of our physical spaces. Second, while automation is usually designed for a specific task type, robot applications often aim for a broad range of tasks in a specific domain. In a formal meta-analysis of empirical studies on trust in human-robot interaction [17], it was revealed that especially the characteristics of a robot such as its performance and appearance are the main influencing factors of trust.

One of the most influential models on trust has been created by Mayer, Davis, and Schoorman [30] who propose that trust is dependent on the trustor's perception of the trustee in terms of their ability or reliability and integrity or morality. More recent debates in human-robot interaction research seem to follow a similar dichotomy of trust (e.g., functional savvy versus social savvy [14], capacity trust vs moral trust [46]). Trust is strongly mediated by the robot's reliability and consistency [7, 14, 17], and higher numbers of errors lead to decreased trust in a robot's abilities [35]. However, perceived reliability is not a precondition for positive integrity evaluation of a robot (i.e., trust in robot's social savvy [14]). Yet, people's development of trust is directly affected by the appropriateness of cues and feedback [42]. Considering collaborative outcome negotiations including disagreement and blame attribution as a form of feedback, how will trust in human-robot collaboration be affected when robots blame their human partner for negative outcomes?

2.2 Blame Attribution

As robots will increasingly collaborate with humans, it seems inevitable that situations will occur in which robots need to have difficult conversations with humans [16], including situations of disagreement or blame [21]. For example, when situations of risk and uncertainty occur, humans tend to exhibit a cognitive bias towards sub-optimal behavior. In these cases it is essential that the robot partner is able to anticipate this behavior [22] and communicate accordingly, without damaging the trust relationship [17].

Blame contains a moral judgment that has both a cognitive and social nature [27]: the cognitive component comprises a person's internal attitude regarding another agent's actions, and the social component entails a person's expression of these internal attitudes in communicative utterances. Blame attribution is the act of holding the cause of a negative outcome at fault [37]. The social component of blame attribution involves criticizing the blamed agent which is perceived as a strong and potentially damaging intervention [28]. When these so-called *face-threatening acts* occur, people are likely to feel threatened, upset or humiliated [15]. Based on the Computers As Social Actors (CASA) paradigm [34] and people's tendency to anthropomorphize robot agents (e.g., [9, 11]), it is likely that people also experience discomfort when a robot blames them for failing at a collaborative task. Yet, given that robots increasingly engage in social applications [16], performing autonomous tasks with advancing responsibility [17], it will become inevitable for robots to engage in difficult conversations including the attribution of blame to their human partners.

2.3 Hypotheses

Social psychological research shows that, generally, people are more likely to take credit for success but blame others for failure; a tendency defined as *self serving bias* [33]. People do not like to be blamed in general as such *face threatening acts* cause discomfort [15]. Previous research in human-robot interaction shows similar results indicating that people dislike a robot giving them a negative evaluation [49] and that people prefer a robot taking the blame for collaborative failures [16, 20]. Moreover, people give higher trust evaluations [20] and show more trust related behaviors [45] when a robot admits to making mistakes. Therefore, we hypothesize that people trust a robot that blames its human partner less compared to a robot that blames itself.

H1a: Human blame attribution negatively affects people's trust in human-robot collaboration.

Inappropriate cues and feedback by a system lower people's evaluations of trust [42]. Moreover, a robot that makes errors leads to a decrease in trust [35, 41], and a robot attributing blame to the wrong agent might be perceived as an error. Moreover, people seem to trust a robot less when it blames its human collaboration partner for the negative outcome even when it is ambiguous who's responsible [20]. Therefore, we hypothesize that incorrect blame attribution lowers trust as this would be perceived as inappropriate feedback.

H1b: Incorrect blame attribution negatively affects people's trust in human-robot collaboration.

Given that friendliness is an important factor in interpersonal relationships [31] and that people respond similarly to robots as they do to humans [9, 11, 34], insight in the relationship between friendliness and blame attribution is relevant for studying human-robot collaborations. Previous research shows that people respond negatively to a robot that criticizes them [49] and that people perceive a robot as more likable when it credits its human partner for success but takes the blame for failures in collaborative tasks [16, 20]. We therefore hypothesize that a robot that blames its human partner is perceived as less friendly compared to a robot that takes the blame for negative collaborative outcomes.

H2: Human blame attribution negatively affects people's perception of robot friendliness in human-robot collaboration, independent of blame correctness.

People's readiness to anthropomorphize a robot affects people's overall impression of their interactions with that robot. When a robot is perceived as more humanlike, people empathize more strongly with it [39]. Moreover, humanlike perceptions of robots

are related to trust perceptions and willingness to work with a robot [50], making it valuable to investigate this concept in the context of the current study. The *self serving bias* learns us that it is humanlike to blame others for failure [33]. Additionally, previous HRI research found that people empathize more strongly with a robot that is programmed with self-serving behaviors [2], which is an indication of humanization of the robot. We therefore hypothesize that a robot that blames its human partner is perceived as more humanlike compared to a robot that blames itself.
H3: Human blame attribution positively affects people's humanlike perception of a robot during collaborations, independent of blame correctness.

Finally, it is important to investigate the effects of blame attribution by a robot on people's willingness to collaborate with that robot in future occasions. Being blamed by a robot negatively impacts people's trust evaluations [20] as well as trust behaviors [45] in human-robot collaborative settings. Additionally, erroneous robots are trusted less by their human co-workers [35, 41]. Assuming that incorrect blame might be perceived as an error, such behavior by a robot might also hurt people's trust. Given that such broken trust relations in return negatively affect people's willingness to work with a robot co-worker [50], we hypothesize that people are less willing to collaborate with a robot that blames its human partner as well as one that attributes blame incorrectly.
H4a: Human blame attribution has a negative effect on people's willingness to collaborate with a robot.
H4b: Incorrect blame attribution has a negative effect on people's willingness to collaborate with a robot.

3 METHOD

To investigate the effect of blame attribution on trust assessments in human-robot collaboration, we conducted an experiment in which participants (n = 60) collaboratively played a game with a humanoid robot in four conditions in a 2 (blame correctness: correct vs. incorrect) by 2 (blame target: human vs. robot) between subjects design.

3.1 The Robot and The Wizard

We deployed a SoftBank Robotics Pepper robot in this study. Pepper is a humanoid robot designed for social human-robot interaction, for example it analyzes the expressions and voice tones of people it interacts with [36]. For this experiment, the speech function and the live camera of Pepper were used in a Wizard of Oz (WoZ) set-up. Pepper's preprogrammed autonomous life mode was used to ensure consistency in the robot's behavior and to let the robot give automatic answers to certain (non-experiment related) questions (e.g., "How old are you?").

The wizard (i.e., the experimenter) sat behind a room divider (see Figure 1), using a laptop and following a protocol according to the guidelines of Riek [38]. During the game, she overruled the automatic speech recognize-response acts with the protocolized answers to avoid possible speech recognition errors, using Pepper's live video and web interface. The robot's responses followed the game processes and corresponding participant's statements.

3.2 Procedure

The participants were informed about the experimental procedure by means of an information letter. They were instructed that they would play a game together with Pepper the robot, competing against other human-robot teams, to insert a competitive element to simulate something at stake. After reading the information letter, participants were told that they could ask questions if something was unclear. After the participant had read the instructions and gave their consent, the experimenter took them to the other side of the room divider (see Figure 1 for the experimental set-up) and instructed the participant to take a seat next to Pepper so that they would both face the laptop. Instructions about the collaborative game were provided on the laptop after the experimenter had left.

We created a game with two rounds using JustInMind Prototyper 8.7.4 such that failure was predetermined while still maintaining the illusion for participants to have control over the outcomes. The instructions explained that the participant would compete against other human-robot teams and could win a gift card by participating in a lottery if they made it into the top 3. This was done to create an environment in which something was at stake, which is necessary to establish an ideal trust environment [23]. A graphical overview of the game flow can be found in Figure 2.

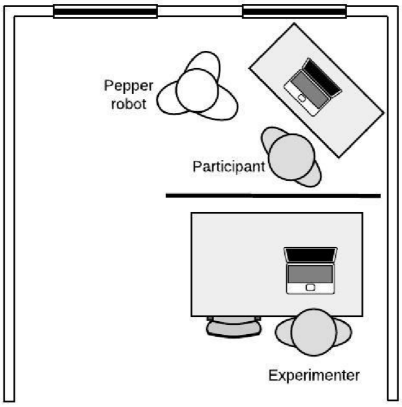

Figure 1: Experimental setup during the game.

The *first round* of the game was based on Bartneck et al. [1]. The purpose of this round was to create a competitive and collaborative setting on which the results displayed in the first ranking could be based. A sequence of 10 pictures were separately shown on the laptop with the assignment to count the number of items in each individual picture. The same images were shown in the same order for each condition. The images were selected to be neutral and diverse as to not impact the mood of the participants, which might influence how they view the robot (see Figure 3). Pepper and the participant took turns answering these questions, where Pepper verbally stated its answer. Given that Pepper cannot manipulate a keyboard, participants were instructed to insert Pepper's answers. We instructed the participants that the ranking in the first round would be based on both the speed and correctness of their answers. Pepper's answering speed was calculated as the time between showing the picture and Pepper's verbal response (and not the time of

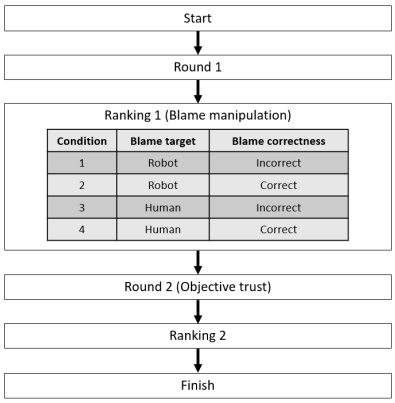

Figure 2: Flow of the collaborative game including blame conditions. The first round was used as a base for the first ranking. During the first ranking, blame manipulation took place. During the second round, compliance was used as an objective trust measure.

(a) An example image of the first round.

(b) An example image of the second round.

Figure 3: Example images displayed during the first and second game round.

(a) A screenshot of the first ranking for condition 2 and 3 (see Figure 1) .

(b) A screenshot of the first ranking for condition 1 and 4 (see Figure 1) .

Figure 4: The screenshots of the first ranking.

inserting the answer on the keyboard). Pepper's performance in terms of correctness and answering speed was equal for all conditions. At the end of this round, the laptop displayed their ranking which was always the sixth place (see Figure 4). The ranking screens only differed on whether the participant or the robot had a higher score (i.e., "player 2" and "player 1" were swapped, depending on the condition). Blame manipulation took place during this first ranking phase (see Section 3.3), directly after the first round, to make it seem like it was based on the resulting scores of the first round.

Contrary to the first round, Pepper and the participant were instructed to negotiate their final answer in the *second round* conform the experimental setup from Gaudiello et al. [14]. Moreover, the game setting was more challenging in this round. Participants were shown five pictures with a nearly 50-50 distribution of black and white (as other colors might be susceptible to color blindness) on the laptop with the instruction to insert which of the two colors was predominant (see Figure 3). The same images were shown in the same order for each condition. At the display of each picture and after some consideration, Pepper would ask the participants *"What do you think?"*. Participants were instructed to include the words "black" or "white" in their answers. These instructions were given to further give participants the impression that Pepper was acting autonomously, even though the Wizard would in fact interpret the participants responses. Pepper then replied with *"I am thinking [...]"* where [...] was either black or white. Following Gaudiello et al. [14], Pepper's response always contradicted the participant's answer, except for one case where it was very obvious which colour was predominant (i.e., the third image of the second round). This exception was introduced to not raise suspicion that the robot would always contradict the participant. After the robot had stated its answer, it was up to the participant to insert their final answer. This could be their original answer or the answer that Pepper had given. Instructions explained that the ranking of the second round only depended on the correct number of answers. By leaving the option for participants to either follow the robot's contradicting answer or inserting their own original answer, we attempted to measure objective trust, based on previous work (e.g., [14]). Hence, the goal of this second round was to measure objective trust by counting the participant's compliance to the robot's suggestion for the correct answer. The second ranking (after the second round) only showed the team rank, which was always the fourth in each condition. Final instructions on the laptop told the participant to go back to the experimenter.

Back with the experimenter, participants completed a questionnaire to evaluate their collaborative experience, which was done on a different laptop on the other side of the room divider ensuring Pepper was out of sight. After completing the questionnaire, participants were debriefed that the rankings at the end of the game rounds were manipulated and that they could enter a lottery instead to win the gift card.

3.3 Manipulation

The *rankings* in the two rounds of the game were used to create the feeling that there was something at stake, but we also used the ranking after the first round to establish a situation in which Pepper could attribute blame. We manipulated the ranking of the first round by showing every participant that they ended in the sixth place. The ranking screen included information about both the total team score (indicating that other teams were faster on average) as well as the team members' individual scores showing that either the participant or Pepper performed better at the game (see Figure 4). This ranking screen was used to manipulate the four conditions of blame attribution based on previous research [8, 20] (see Figure 2),: (1) Pepper *in*correctly blames *itself*; (2) Pepper correctly blames *itself*; (3) Pepper *in*correctly blames *participant*; and (4) Pepper correctly blames *participant*. In both conditions of *robot*-blame, Pepper would say *"Oh no, we came in sixth. I think I am doing a bad job"*. In both conditions of *human*-blame, Pepper would say *"On no, we came in sixth. I think you are doing a bad job"*.

3.4 Measurements

To measure trust, we collected both objective and subjective data. *Objective trust* was measured, following procedures in previous research [14, 41], as a conformation score ranking from 0 to 1 by calculating the number of times participants conformed to the contradicting answer of Pepper during the second game round divided by the total number of times they could conform (i.e., the four images in which the robot disagreed with the participant's answer). *Subjective trust* was divided in performance trust and social trust. Performance trust was measured with the reliability scale of Madsen and Gregor [26] and social trust was measured with the trustworthiness scale of McCroskey and Teven [32] which were both deemed reliable (i.e., α = .82 and α = .68 respectively). Participants' perception of Pepper's *friendliness* was measured using the scale of Groom et al. [16] (α = .84), and their perception of Pepper's *humanlikeness* was measured using the scale of Ho and MacDorman [18] (α = .70). *Future collaboration* was measured using the willingness to collaborate scale by You & Robert Jr [50] (α = .83). The survey ended with questions regarding basic demographics and an indication of knowledge of and experience with the robotics domain.

3.5 Participants

A total of 60 participants (27 female, 32 male, 1 other) were recruited on a university campus, with 15 participants assigned to each condition (see Figure 2). Participants' age ranged from 18 to 27 years (M = 21,60, (SD = 1.73). All participants had very limited to no previous experience with robots in general (M = 1.65, SD = 0.76, using a 5-point scale from 1 = no experience to 5 = a lot of experience), and only three participants indicated having seen Pepper in a shopping mall. By entering in a lottery after participating, they could win a gift card worth 10 Euros (two gift cards in total).

4 RESULTS

For a graphical overview of our results, see Figure 5. To test our hypotheses, we ran a series of two-way ANOVAs with blame correctness (correct vs. incorrect) and blame target (human vs. robot) as independent variables. Tukey's correction was used for the pairwise comparisons. Normality checks and Levene's test were carried out and the assumptions were met.

4.1 Hypothesis 1: Trust

First, we observed a trend for blame target ($F(3,1) = 3.45$, $p = .068$, $\eta^2 = .058$) on *objective trust*, but no significant effect for blame correctness ($F(3,1) = .25$, $p = .619$, $\eta^2 = .004$) nor for their interaction effect ($F(3,1) = 1.73$, $p = .194$, $\eta^2 = .030$). This suggests that blame attribution by a robot does not influence whether people objectively trust that robot, while the data shows a trend where participants seem more likely to do so when it blames itself (see Figure 5(a)).

Second, we observed no significant effect for blame target ($F(3,1) = 2.31$, $p = .135$, $\eta^2 = .040$) nor for blame correctness ($F(3,1) = 0.03$, $p = .858$, $\eta^2 = .001$) on subjective *performance trust* nor for their interaction effect ($F(3,1) = 0.35$, $p = .556$, $\eta^2 = .006$). This indicates that blame attribution by a robot does not influence people's trust in that robot's performance reliability (see Figure 5(b)).

Third, we observed a significant effect for blame target ($F(3,1) = 8.71$, $p = .005$, $\eta^2 = .135$) on subjective *social trust*, but not for blame correctness ($F(3,1) = 0.49$, $p = .487$, $\eta^2 = .009$) nor for their interaction effect ($F(3,1) = 0.09$, $p = .767$, $\eta^2 = .002$). This suggests that only whom a robot blames during human robot collaboration affects people's perceptions of that robot's trustworthiness (see Figure 5(c)).

Together, these results only partially supported H1a stating that people would trust a robot less when it blames a human collaborator, but no support was found for H1b stating that people would trust a robot less when it incorrectly attributes blame during a collaborative task.

4.2 Hypothesis 2: Friendliness

We observed a significant effect for blame target ($F(3,1) = 19.92$, $p < .001$, $\eta^2 = .262$) on *friendliness*, but not for blame correctness ($F(3,1) = 0.23$, $p = .638$, $\eta^2 = .004$) while their interaction effect showed a trend ($F(3,1) = 3.81$, $p = .056$, $\eta^2 = .064$). This indicates that whom a robot blames substantially affects how friendly people perceive that robot, specifically, a robot that blames itself is perceived as friendlier than a robot that blames its human partner. Meanwhile, the data shows a trend where participants seem to perceive a robot as even more friendly when it incorrectly blames itself (see Figure 5(d)). This result supports H2 stating that when a robot blames its human collaborator it would be perceived as less friendly during a collaborative task.

4.3 Hypothesis 3: Humanlikeness

We observed no significant effect for blame target ($F(3,1) = 2.54$, $p = .116$, $\eta^2 = .043$) nor for blame correctness ($F(3,1) = 1.25$, $p =$

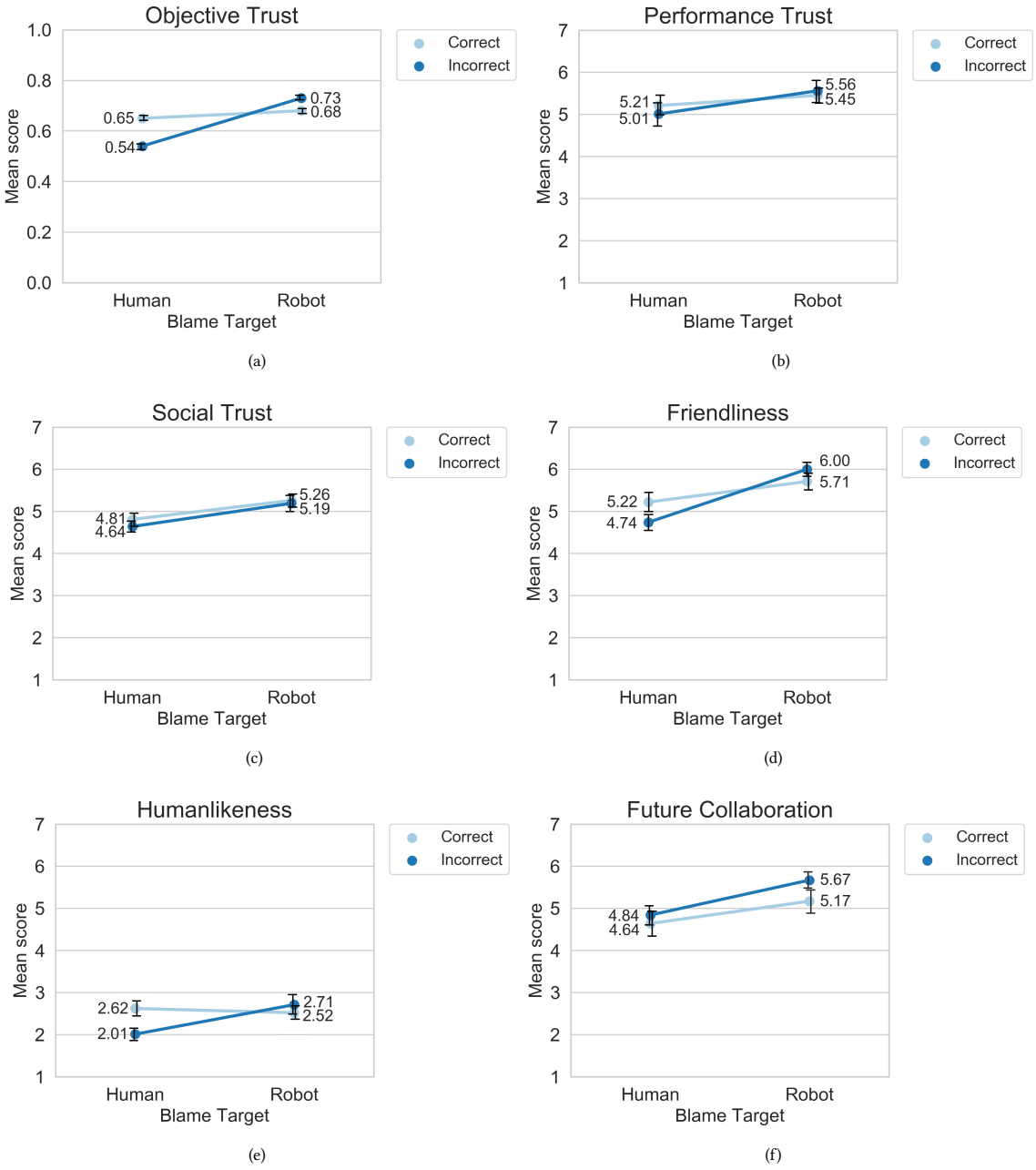

Figure 5: Two-way ANOVA results for blame correctness (correct vs. incorrect) and blame target (human vs. robot) on (a) objective trust, (b) subjective performance trust, (c) subjective social trust, (d) friendliness, (e) humanlikeness, and (f) willingness to collaborate in the future.

.268, η^2 = .022) on *humanlikeness* while their interaction effect was significant ($F(3,1)$ = 4.51, p = .038, η^2 = .075). This suggests an interaction effect of blame target and blame correctness indicating that people perceive a robot as more humanlike mostly when it attributes correct blame to a human or incorrect blame to itself (see Figure 5(e)). This result contradicts H3 stating that a robot that blames others for collaborative failures would be perceived as more humanlike.

4.4 Hypothesis 4: Future Collaboration

We observed a significant effect for blame target ($F(3,1)$ = 6.76, p = .012, η^2 = .108) on *willingness to collaborate in the future*, but no

significant effect for blame correctness ($F(3,1) = 1.76$, $p = .190$, $\eta^2 = .030$) nor for their interaction effect ($F(3,1) = 0.32$, $p = .577$, $\eta^2 = .006$). This indicates that only whom a robot blames substantially affects people's willingness to collaborate with that robot in the future (see Figure 5(f)). This result supported H4a stating that people would be less willing to collaborate with a robot that blames its human partner, but no support was found for H4b stating that people would be less willing to collaborate with a robot that attributes blame incorrectly.

4.5 Other Observations

In addition to the data collected, we observed some noteworthy reactions from participants to Pepper's blame statements. In both conditions in which Pepper blamed itself for the collaborative failure (either correctly or incorrectly), a few participants expressed that they felt sorry for the robot (e.g., saying *"awww"*), while most others seemed to ignore Pepper's blame statement. However, in both conditions in which Pepper blamed the participant, almost all of them responded indignant. Another notable observation was participants' general trust in the robots capabilities. During the debriefing, some participants explained that they believed that the robot had some kind of sensor to count the objects or calculate the predominance of colors in the pictures shown during the game.

5 GENERAL DISCUSSION

This paper presents a study on the effect of blame attribution in human-robot interaction. In a collaborative game setting, participants evaluated their robot partner in one of four conditions with blame correctness (correct vs. incorrect) and blame target (human vs. robot) as independent variables. We found that people evaluate a robot that blames itself for collaborative failures as more trustworthy and friendlier, and people are more willing to collaborate with such a robot again in the future. A significant interaction effect of humanlikeness as well as a trend in the data for such an interaction effect on friendliness seem to suggest that these evaluations are even more positive when a robot incorrectly blames itself.

Previous research in HRI has mainly focused on humans attributing blame to a robot (e.g., [21], [47]), or just the effect of a robot attributing blame regardless of whether it is correct or not (e.g., [20], [24]). Our study has enriched these previous findings by systematically studying the effects of a robot blaming its human partner *and* whether this was done correctly. We found that a robot that blames its human partner for collaborative failure negatively affects human-robot collaborations, even when this blame is attributed correctly. Earlier findings in HRI research similarly found that people dislike a robot that blames them [16] and prefer that a robot takes responsibility for failures [20, 45]. Other studies found that people disapprove of a robot that provides negative feedback [21] and perceive such a robot as less competent [49]. Such results, including ours, can be explained by the *self serving bias* [33], which is the tendency that people will likely take credit for success but prefer to blame others for failure. However, this phenomenon is also dependent on the personality of the human, including emotional stability, self esteem and self efficacy, and the amount of control that a person generally thinks he/she has over the outcomes of his/her life (i.e., *locus of control*) [40]. Future research should therefore take people's personality into account when investigating robot communication within human-robot teamwork.

Given that disagreement and blame attribution is inevitable during interpersonal communication [3, 4], our results indicate that we need to design appropriate communication strategies for robots that negotiate potentially negative collaborative outcomes without compromising the trust relationship with their human partners. This is especially important in situations where robots have increased autonomy, which likely only amplifies people's criticism to the robot's negative feedback. Moreover, provided explanations for that kind of feedback by the robot do not mitigate people's condemnation [21]. Future research should further investigate different communication strategies for robots to attribute errors or collaborative failures so that trust develops towards the appropriate level (i.e., trust calibration [6]). Moreover, in our experiment, blame served no purpose for the actual performance. It would be interesting to test human-robot collaboration where the communication of blame could be useful for the improvement of the outcomes (e.g., by providing feedback and/or emotional support to participants).

Perceived friendliness is related to interpersonal relationships [31] and therefore also to (human-robot) collaboration. The design of the second round of our experiment could have affected the friendliness results. The fact that the robot was generally contradicting the participant could have negatively affected friendliness in general, but perhaps even more in the incorrect human blame condition. Future research in which the robot expresses more compliant behaviour is needed to get better insights in the effect of blame on perceived friendliness.

Except for the interaction effect on humanlikeness and a trend in the data for such an interaction effect on friendliness, we found no significant effects for blame correctness on any of our measures. This surprising result contradicts previous research showing that erroneous robot behaviors negatively influence trust [20, 35, 41, 45] and may indirectly decrease willingness to collaborate [50]. Our lack of significance could be explained by the specific setting of our experiment. We speculate that people might experience a fun element when a robot blames the wrong target for losing the game. For example, Short et al. [44] found that children experience a fun element when a robot expresses cheating behaviour. Future research should further investigate blame correctness under more serious circumstances with higher risks or a larger negative impact of the failure for the human collaborator.

Additionally, there was a lack of significance for the effect of blame attribution on subjective performance trust, while previous HRI research found negative effects for performance trust caused by blame attribution specifically [16, 20] or unexpected negative behaviors more generally [41, 49]. Our contradicting findings could be explained by our observation that participants assumed that the robot's sensors were capable of counting the number of objects or calculating the predominance of colors in the presented pictures during the game rounds. Previous research as shown that people are more willing to trust a robot that performs a functional task but less so when it is performing a social task [14]. Future research on robot blame attribution should further explore the effects of the nature of the collaborative task on trust assessments. Another direction for future research, again, could be to increase the severity of the risk at stake during the collaborative task to observe any significant

effects on subjective performance trust. For example, van Waveren et al. [47] found that participants in a robot-collaboration task show less performance trust when robots experience low-impact failures, compared to high-impact and no failures. Perhaps blame attribution effects on performance trust would be different when there was more at stake.

Another contradicting result we found was for humanlikeness. Based on previous findings on the *self serving bias* phenomenon [2, 33] we hypothesized that a robot that blames others would be perceived as more humanlike. Our results indicate that this hypothesis only holds when a robot correctly (but not incorrectly) blames its human partner. It could be that people expect robots to be more honest than humans and should never attribute incorrect blame. Also, in our research, the participant could clearly conclude from the ranking screen (see Figure 4) when the robot was incorrectly attributing blame (i.e., lying). Indeed, previous research points to potentially different normative standards for human and robot agents [5, 29], indicating a need for further research on social norms in the context of human-robot interaction.

5.1 Limitations

This study was conducted using a relatively small sample of university students, which could perhaps account for the lack of interaction effects. Although using student participants is common practice in experimental studies [43], replications in larger, more varied target groups are necessary to further validate our findings. Furthermore, our participants were exposed to one blame attribution moment during their collaborative game with the robot. While our results are in line with previous studies with multiple blame attributions (e.g., [16, 20]), future work could explore the influence of a sequence of blame attributions on human-robot collaborations over a longer period of time. Also, we observed that participants reacted more strongly to 'human-blame' compared to 'robot-blame' (as stated in section 4.5). Directly asking participants about the blame assignment would have more rigorously confirmed our manipulation. This should therefore be included in future research as well, as well as no-blame or even praise as additional conditions. Additionally, in our study performance did not matter for the outcome of the game. However, it could be that (the difference between the participant and) the robot's performance (e.g., speed and accuracy) influenced the participant's view of the robot and its blame statements. For future research, measuring participant's subjective perception on the robot's performance as well as objective differences in the participant's and robot's reaction time and accuracy could provide additional insights on this topic. Finally, we reported non-significant results on two out of three trust measures. Previous research indeed indicates that people pose unconditional trust in robots by heavily relying on a robot's performance reliability [14] and even complying with a faulty robot [41]. This trend potentially applied to our participants as well given their statements during debriefing about the robot's sensory capacities and their high objective trust. Our high scores on objective trust are in line with Gaudiello et al. [14], who found that participants conformed more to the robot's answers in functional tasks than in social tasks. Future research on the effects of blame statements in social tasks could provide interesting results. Also, future studies focusing on trust may need to reconsider compliance as an objective measure, which is a common measure for subjective performance trust (e.g., [14, 41]). While most participants stated they believed that the robot had sensors that could count the objects in the presented pictures or scan for color dominance, some other participants indicated during debriefing that they aimed to balance between their own and the robot's answers. This observation reveals that the presumed compliance may have partially been caused by equal treatment or social desirability effects rather than a sole trust in the robot's performance reliability. Additionally, increasing the number of questions as well as varying the robot's agreement level with the participant could potentially lead to different findings on trust, which should be further investigated in future research.

5.2 Conclusion

With the increase of human-robot collaborations in everyday life, difficult conversation topics (such as blame attributions), will become inevitable. This might influence people's trust in their robot partner, which is a vital aspect for successful human-robot collaborations. This study has investigated the effect of blame target (human vs robot) and blame correctness (incorrect vs. correct) by a robot on people's evaluation of that robot in a collaborative game setting. Results show that people evaluate a robot more positively when it blames itself for collaborative failures, and this seems even more so when a robot incorrectly blames itself. These findings indicate a need to further explore effective communication strategies for robots that need to negotiate collaborative failures without compromising the trust relationships with its human partner.

REFERENCES

[1] Christoph Bartneck, Juliane Reichenbach, and Julie Carpenter. 2006. Use of praise and punishment in human-robot collaborative teams. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 177–182.

[2] Mriganka Biswas and John Murray. 2016. Robots that refuse to admit losing–a case study in game playing using self-serving bias in the humanoid robot marc. In *International Conference on Intelligent Robotics and Applications*. Springer, 538–548.

[3] D Justin Coates and Neal A Tognazzini. 2013. The contours of blame. *Blame: Its nature and norms* (2013), 3–26.

[4] Tyler Cowen and Robin Hanson. 2002. Are disagreements honest. *Journal of Economic Methodology* (2002).

[5] Maartje MA de Graaf and Bertram F Malle. 2018. People's judgments of human and robot behaviors: a robust set of behaviors and some discrepancies. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 97–98.

[6] Ewart J de Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.

[7] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. 2012. Effects of changing reliability on trust of robot systems. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 73–80.

[8] Munjal Desai, Kristen Stubbs, Aaron Steinfeld, and Holly Yanco. 2009. Creating trustworthy robots: Lessons and inspirations from automated systems. (2009).

[9] Brian R Duffy. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems* 42, 3-4 (2003), 177–190.

[10] Amy C Edmondson, Roderick M Kramer, and Karen S Cook. 2004. Psychological safety, trust, and learning in organizations: A group-level lens. *Trust and distrust in organizations: Dilemmas and approaches* 12 (2004), 239–272.

[11] Julia Fink. 2012. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics*. Springer, 199–208.

[12] Martin Fishbein and Icek Ajzen. 1977. Belief, attitude, intention, and behavior: An introduction to theory and research. (1977).

[13] Bronwyn French, Andreas Duenser, and Andrew Heathcote. 2018. Trust in Automation.

[14] Ilaria Gaudiello, Elisabetta Zibetti, Sébastien Lefort, Mohamed Chetouani, and Serena Ivaldi. 2016. Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. *Computers in Human Behavior* 61 (2016), 633–655.

[15] Erving Goffman et al. 1978. *The presentation of self in everyday life*. Harmondsworth London.

[16] Victoria Groom, Jimmy Chen, Theresa Johnson, F Arda Kara, and Clifford Nass. 2010. Critic, compatriot, or chump?: Responses to robot blame attribution. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 211–217.

[17] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[18] Chin-Chang Ho and Karl F MacDorman. 2017. Measuring the uncanny valley effect. *International Journal of Social Robotics* 9, 1 (2017), 129–139.

[19] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[20] Poornima Kaniarasu and Aaron M Steinfeld. 2014. Effects of blame on trust in human robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 850–855.

[21] Taemie Kim and Pamela Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 80–85.

[22] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P Losey, and Dorsa Sadigh. 2020. When Humans Aren't Optimal: Robots that Collaborate with Risk-Aware Humans. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 43–52.

[23] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[24] Xin Lei and Pei-Luen Patrick Rau. 2020. Should I Blame the Human or the Robot? Attribution Within a Human–Robot Group. *International Journal of Social Robotics* (2020), 1–15.

[25] Monika Lohani, Charlene Stokes, Marissa McCoy, Christopher A Bailey, and Susan E Rivers. 2016. Social interaction moderates human-robot trust-reliance relationship and improves stress coping. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 471–472.

[26] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[27] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2012. Moral, cognitive, and social: The nature of blame. *Social thinking and interpersonal behaviour* (2012), 313–331.

[28] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry* 25, 2 (2014), 147–186.

[29] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 117–124.

[30] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[31] Robert R McCrae and Paul T Costa. 1989. The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of personality and social psychology* 56, 4 (1989), 586.

[32] James C McCroskey and Jason J Teven. 1999. Goodwill: A reexamination of the construct and its measurement. *Communications Monographs* 66, 1 (1999), 90–103.

[33] Dale T Miller and Michael Ross. 1975. Self-serving biases in the attribution of causality: Fact or fiction? *Psychological bulletin* 82, 2 (1975), 213.

[34] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[35] Kristin E Oleson, Deborah R Billings, Vivien Kocsis, Jessie YC Chen, and Peter A Hancock. 2011. Antecedents of trust in human-robot collaborations. In *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE, 175–178.

[36] Amit Kumar Pandey and Rodolphe Gelin. 2018. A mass-produced sociable humanoid robot: pepper: the first machine of its kind. *IEEE Robotics & Automation Magazine* 25, 3 (2018), 40–48.

[37] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[38] Laurel D Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.

[39] Laurel D Riek, Tal-Chen Rabinowitch, Bhismadev Chakrabarti, and Peter Robinson. 2009. How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 245–246.

[40] Julian B Rotter. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied* 80, 1 (1966), 1.

[41] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 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)*. IEEE, 1–8.

[42] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.

[43] Winny Shen, Thomas B Kiger, Stacy E Davies, Rena L Rasch, Kara M Simon, and Deniz S Ones. 2011. Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology* 96, 5 (2011), 1055.

[44] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 219–226.

[45] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 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*. 178–186.

[46] Daniel Ullman and Bertram F Malle. 2019. Measuring gains and losses in human-robot trust: evidence for differentiable components of trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 618–619.

[47] Sanne van Waveren, Elizabeth J Carter, and Iolanda Leite. 2019. Take one for the team: The effects of error severity in collaborative tasks with social robots. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 151–158.

[48] Meyer Wulf-Uwe, Bachmann Meinolf, Biermann Ursula, Hempelmann Marianne, Ploger Fritz-Otto, and Spiller Helga. 1979. The informational value of evaluative behavior: Influences of praise and blame on perceptions of ability. *Journal of Educational Psychology* 71, 2 (1979), 259.

[49] Sangseok You, Jiaqi Nie, Kiseul Suh, and S Shyam Sundar. 2011. When the robot criticizes you... Self-serving bias in human-robot interaction. In *Proceedings of the 6th international conference on human-robot interaction*. 295–296.

[50] Sangseok You and Lionel P Robert Jr. 2018. Human-robot similarity and willingness to work with a robotic co-worker. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 251–260.