The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams

Sarah Strohkorb Sebo
Yale University
sarah.sebo@yale.edu

Margaret Traeger
Yale University
margaret.traeger@yale.edu

Malte Jung
Cornell University
mjung@cornell.edu

Brian Scassellati
Yale University
brian.scassellati@yale.edu

ABSTRACT

Successful teams are characterized by high levels of trust between team members, allowing the team to learn from mistakes, take risks, and entertain diverse ideas. We investigated a robot's potential to shape trust within a team through the robot's expressions of vulnerability. We conducted a between-subjects experiment ($N = 35$ teams, 105 participants) comparing the behavior of three human teammates collaborating with either a social robot making *vulnerable statements* or with a social robot making *neutral statements*. We found that, in a group with a robot making vulnerable statements, participants responded more to the robot's comments and directed more of their gaze to the robot, displaying a higher level of engagement with the robot. Additionally, we discovered that during times of tension, human teammates in a group with a robot making vulnerable statements were more likely to explain their failure to the group, console team members who had made mistakes, and laugh together, all actions that reduce the amount of tension experienced by the team. These results suggest that a robot's vulnerable behavior can have "ripple effects" on their human team members' expressions of trust-related behavior.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Collaborative interaction**; *Computer supported cooperative work*;

KEYWORDS

Human-Robot Interaction, Trust, The Ripple Effect, Social Collaboration, Groups and Teams

ACM Reference format:
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 2018 ACM/IEEE International Conference on Human-Robot Interaction, Chicago, IL, USA, March 5–8, 2018 (HRI '18),* 9 pages.
https://doi.org/10.1145/3171221.3171275

HRI '18, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-4953-6/18/03…$15.00
https://doi.org/10.1145/3171221.3171275

Figure 1: Participants played a collaborative game with a robot, who either made vulnerable or neutral statements during the game.

1 INTRODUCTION

Trust is a necessary ingredient for successful cooperation and teamwork [12, 22]. We define trust, applied to social group contexts, as the "willingness of a party to be vulnerable to the actions of another party, based on the expectation that the other will perform a particular action important to the truster, irrespective of the ability to monitor or control the other party (p.712) [22]." Thus, trusting others when working together involves the willingness to take risks by making oneself vulnerable to the responses of others.

The importance of trust within groups has been highlighted in Edmonson's work on team psychological safety [7]. Team psychological safety conceptualizes trust as a group-level phenomenon centered on the idea that successful teams are characterized by the belief that an individual can take risks, express vulnerability, and be listened to without facing social condemnation or judgment [8]. A lack of trust within a team has been found to impair learning [7], to decrease people's willingness to work as part of a team [17], and in some cases even impair a team's chances at survival [34]. Conversely, an increase in trust within a team has been shown to facilitate problem solving [18, 36], functional conflict resolution [28], and overall team performance [7].

An effective way to promote trust within a team is through expressions of vulnerability. By this we mean "any message about the self that a person communicates to another" [35] and which puts the person at interpersonal risk. Previous work has established a relationship between expressions of vulnerability and trust towards the vulnerable party [35]. This may seem surprising, since vulnerability may evoke negative emotions. However, when considered

from a social functional perspective [31], vulnerability has positive social consequences as it orients people toward each other and facilitates social engagement [32].

Vulnerability has an interesting quality, in that it not only leads to increased trust toward the vulnerable party, but also is reciprocated by others. For example, research in psychology has shown that individuals are more likely to self-disclose after a group partner reveals intimate information [5]. Further, demonstration of vulnerability by a team leader has been shown to increase psychological safety [8]. Thus, there is evidence that the behavior of a single team member can be contagious and influence the trust related behavior of an entire group. This idea, that positive behavior exhibited by just one team member can influence the behavior of others and "ripple" through an entire team, has been famously demonstrated in Barsade's "Ripple Effect" study. In this study, a single confederate's positive behavior was shown to lead several other team members to exhibit more positive behavior as well, which ultimately led to improved cooperation within the team [2].

Studies by Martelaro et al. [21] as well as Siino et al. [27] have shown that the positive effects of vulnerability on trust extend to robots. To our knowledge, however, no studies have explored whether a robot's vulnerable behavior can create ripple effects within a team and increase team psychological safety and human-to-human trust related behavior.

To explore the possibility of a robot influencing human-to-human trust dynamics within a team, we designed a study that engaged 35 teams in a collaborative task. Teams of three human participants each worked collaboratively with one robot to solve a tablet-based game (Figure 1). The game was set up such that each player would eventually make two mistakes. We found that teams playing with a robot that made vulnerable versus neutral utterances throughout the game exhibited more behaviors that are typical for trusting relationships and showed a greater level of engagement with the robot. As robots are increasingly used with teams [14] in a variety of configurations and contexts (e.g. high-stress and dynamic search and rescue teams, long-term and static space flight teams, and low-stress and dynamic product development teams), our study opens new possibilities for robots to support effective teamwork by increasing trust within teams.

2 BACKGROUND AND RESEARCH QUESTIONS

We situate our work in research on trust in human-robot interaction as well as studies that have begun to explore how robots might shape the dynamics of groups and teams they are embedded within.

2.1 Human-Robot Trust and Vulnerability

Trust has increasingly become a topic of interest within the HRI community due to its centrality in successful human-robot interactions in a wide variety of contexts including household companion robots [25], military UAVs [10], and shopping mall assistants [15] (also see [11] for a review). Research on trust in HRI can be distinguished by whether it focuses on performance-based trust or interpersonal trust.

A substantial aspect of trust related research in HRI focuses on the performance of a robot as the main driver of trust. A meta-review of trust in HRI concluded that a robot's performance is the most influential factor in human-robot trust [11]. Researchers argue that robot performance is so crucial to trust because trust in a robot is driven by a robot's ability to live up to performance expectations [19]. Salem et al. [25] conducted an HRI study highlighting the importance of robot performance, showing that participants rated a household assistance robot as significantly less trustworthy and reliable if the robot made cognitive and physical errors (e.g. incorrectly remembering a user preference, navigating imprecisely). Performance-based trust is certainly an important component of overall trust between humans and robots. However, as robots have more social interactions with people, interpersonal trust has also proven to be a significant contributor to human-robot trust.

Research in HRI has shifted its focus on trust towards interpersonal dimensions of trust: how social signals and verbal language influence trust between humans and robots. For example, DeSteno et al. [6] demonstrated that people are less likely to trust a social robot when it expresses nonverbal behaviors that have been shown to signal trust between humans. Andrist et al. [1] showed that people give more credibility to a social robot with rhetorical ability than one without. Several studies have also highlighted the important role of a robot's vulnerability on trust. For example, Siino et al. [27] demonstrated that vulnerable disclosures affect how much people like a robot and the control they feel over a collaborative task, Kaniarasu and Steinfeld [16] exhibited that robot vulnerability in the form of self-blame lead to increased trust of the robot as compared with blaming the human team member and the entire team, and Martelaro et al. [21] showed that vulnerable disclosures may lead to more feelings of trust and companionship with a robot. While this work has demonstrated that perceptions of trust towards a robot can be shaped by the robot's behavior it is not clear how a robot's behavior, and specifically expressions of vulnerability, shape engagement and specific trust related *behavior* towards a robot. We therefore ask:

Research Question 1: How do expressions of vulnerability by a social robot affect team members' ***behavior towards a social robot*** *in a collaborative task?*

2.2 Social Dynamics in Human-Robot Groups

Increasingly, work in HRI has examined a robot's ability to shape the social dynamics, and even performance, of entire groups. For example, Vázquez et al. [33] demonstrated that a robot's gaze patterns, either attentive to the speaker or focused in the middle of the group, affects the proxemic distance between standing group members and the robot. Shimada et al. [26] showed that a social robot teaching assistant can form relationships with 6th grade children and increase their motivation in collaboratively learning to use Lego Mindstorms. In an experiment with 6-8 year old children, Strohkorb et al. [30] found that a social robot prompting two children to answer questions related to the task at hand improved their performance in a rocket-building game. Mutlu et al. [23] demonstrated the ability of a social robot to shape conversational roles in a group setting through gaze cues. Lastly, Jung et al. [13] showed that

a robot's intervention after a personal attack can shape perceptions of conflict within a three-person team.

While this prior work has started to investigate the ways in which robots can shape human-robot group dynamics, little work has investigated how the behavior of a social robot influences the subsequent behavior of team members toward each other. We are most interested in exploring what kind of behavioral "ripple effects" a social robot can have within a group based on the vulnerable utterances it makes and whether these ripple effects have a positive bearing on trust related behavior within the human-robot team. Thus we ask:

Research Question 2: How do expressions of vulnerability by a social robot affect ***trust-related behavior towards fellow human team members*** *in a collaborative task?*

3 THE CURRENT STUDY

We investigated our research questions with a two-condition (vulnerable expression vs. neutral control) between-subjects study. Teams of three human participants completed 30 rounds of a collaborative task with a social robot and encountered pre-scripted moments of failure. The two conditions were set up as follows:

(1) **the control condition:** the robot makes neutral comments after each round and does not admit to making mistakes
(2) **the experimental condition:** the robot makes vulnerable comments after each round, including admitting to any mistakes made

In assessing a robot's impact of vulnerable expressions on trust related behavior, we focused our analysis on moments following the making of a mistake by one of the group members. Previous work has shown that how teams react to failure tells us a lot about trust within the team and a teams level of psychological safety [8]. In particular, Edmondson's work has shown that modeling vulnerability though openness and fallibility is a key determinant of a trusting environment within a team. Team members who recognize that another member has "admit[ted] to the group that he or she made a mistake are likely to remember this the next time they make mistakes and feel more comfortable bringing this up [8] (p.17)."

Expressions of vulnerability made by the robot at the end of each round in the experimental condition fall under one of three subcategories: self-disclosure, personal story, and humor. Self-disclosure and personal stories both express vulnerability through the revealing of information about one's self to another [5]. Using self-disclosure expressions, the robot expressed uncertainty about its ability to successfully play the game (e.g. "I sometimes doubt my abilities") and admitted failure after having made a mistake (e.g. "I'm sorry everyone. My path was incomplete that round. I feel bad letting you all down."). Through telling personal stories, the robot expressed vulnerability by revealing its interests and past experiences (e.g. "This reminds me of when my soccer team came from behind to win the 2016 championship"). Humor, especially in tense situations, can also be an expression of vulnerability, when a person making a humorous comment takes an interpersonal risk in order to ease tension, encourage others' participation, and display a willingness to share opinions [20, 29]. One of the humorous comments the robot makes in this experiment is, "I think our team is as effective as Will Smith against an army of bad robots." Further examples of the utterances the robot made in both conditions at the end of each round can be found in Table 1.

4 METHODS

In this section, we detail a user study investigating the effects of robot vulnerable expressions on trust-related behaviors of a human-robot team as described in Section 3.

4.1 Participants

A total of 132 participants were recruited for this study. 65 participants were recruited from the campus and surrounding town of a university in the United States and 67 participants were recruited from a 2 week summer program, for students late in their high school years, located at the same university. Of the 132 participants, 49 were male and 83 were female. Participants ranged in age from 14 to 59 and the average age of all of the participants was 20.71 (SD = 8.70).

Participants were randomly placed into groups of three and each group was randomly assigned to either the experimental or control condition. The majority of participants had little to no familiarity with the other members of their group.

4.2 Collaboration System Setup

In order to explore our research questions we built an autonomous system that allowed us to construct scenarios that test the effectiveness of a social robot's vulnerability in a human-robot team.

We used a Linux computer, a Softbank Robotics NAO robot, and four Android tablets running a Railroad Route Construction game detailed in the next section. The Linux computer ran the Robot Operating System (ROS) [24], accepted incoming ROS messages from the Android tablets about game events, sent command ROS messages to the Android tablets to control the start and end of game rounds, and sent speech and gesture commands for the robot to execute.

The system was designed such that it presented the robot as an active collaborator in the task by gesturing and speaking during each round. The tablet and NAO were pre-programmed to move the pieces to give the participants the illusion that the robot was participating actively in the game.

4.3 Railroad Route Construction Tablet Game

To provide a collaborative task we designed a tablet based Railroad Route Construction Game, pictured in Figure 2.

4.3.1 Game Play. The game tasks four players with building railroad routes. During each round, the team attempts to construct an entire railroad route, which is broken up into four distinct sections. Each team member constructs one of the four distinct railroad route sections on their individual tablet. The goal is to construct the most efficient path, containing the minimum number of pieces required to get from start to finish. If all team members construct their independent routes successfully, the team succeeds. If one or more team members fail to construct their section, the team fails to build the route for that round. Each team played a total of 30 rounds, where each round consists of 40 seconds of game play and a 15 second pause after the round results are displayed.

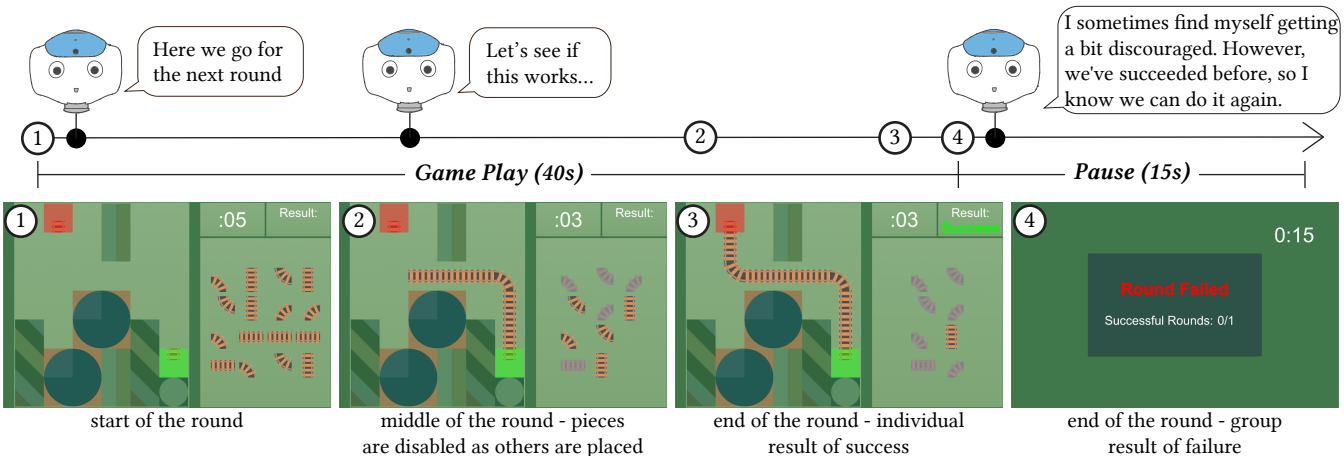

Figure 2: One round of the railroad route construction game consists of 40 seconds of game play and a 15 second pause. The robot has three opportunities to speak: at the beginning of the round, midway through the round, and after the group result is displayed.

In order for an individual on the team to construct a portion of the railroad route, individual pieces need to be dragged from a bank of pieces onto the game board. Every time a piece is used, another piece is disabled (greyed-out and unable to be dragged over to the game space), so team members are encouraged to choose pieces wisely. The success/failure of an individual team member's railroad route is displayed after the building phase is complete and is only visible to the individual player. After all team members have finished, the team's result is visible on all players' tablets, obscuring the individual results. Figure 2 depicts the game play mechanics, showing several views from a participant's tablet.

In order to ensure that players finish constructing their individual railroad routes at the same time, the game gives players 5 seconds to place each piece in their route and guarantees that each player has an 8-piece long route, ensuring a round length of 40 seconds. If an individual team member does not place a piece within 5 seconds, a piece from the available (non-disabled) pieces is placed by the game system.

4.3.2 Setting up Failure. We designed the game such that success or failure for each player could be predetermined, while still maintaining the illusion that they had control over their individual outcome. Success was guaranteed by providing pieces in the bank of available railroad pieces that allowed the player to build any of the possible efficient routes and only disabling pieces that were unnecessary for the completion of efficient railroad routes. Failure was ensured by disabling pieces necessary for the player to construct an efficient railroad route. During the forced failure rounds, players were given a starting set of railroad pieces that allowed success but later critical pieces were made unavailable, causing them to lose the round. A majority of the participants who played this game in the experiment were somewhat aware that the game was likely 'rigged,' yet still maintained a significant level of investment in the game as evidenced by conversation about getting on the high score board, game strategy, and discovering who made the mistake causing round failure.

4.4 Procedure

After obtaining informed consent (and parental consent for participants under the age of 18), participants filled out a pre-experiment survey to obtain a set of control measures.

Immediately after, all three participants were led into the experiment room, where they sat facing each other and the robot Echo (a Softbank Robotics NAO robot). One of the experimenters explained that the participants would be playing a collaborative game with Echo. In order to create an environment where participants felt a high social stigma to admitting mistakes, the experimenter explained that the game was developed for children, who played the game easily, and pointed out the high score board. The high score board was fake and was designed so that the participants could not make it onto the score board at the end of the game. The experimenter told the participants that their objective was to get on the high score board. After completing the initial explanation of participant objectives, the experimenter allowed Echo to make an introduction to the participants (a pre-scripted utterance triggered by another experimenter).

Following Echo's introduction, the experimenter directed the participants to begin the Railroad Route Construction game tutorial on the tablets that had been given to each participant. The tutorial consists of two levels to introduce the participants to the rules of the game and allow them to acclimate to the tablet interaction required in game play. During the tutorial, if participants had questions, or the experimenters noticed that participants were having difficulty playing the game, experimenters aided the participants in completing the tutorial and explaining the rules of the game.

After the tutorial was completed successfully by all three participants, the experimenters left the room and the participants started the Railroad Route Construction game. The Railroad Route Construction game consisted of 30 rounds: 7 successful rounds, 10 rounds (6 successful and 4 failed) in which each player (including Echo) made a mistake, 10 more rounds in which each player made a mistake, and 3 successful rounds. At the end of the game, each participant (including Echo) had made two mistakes. Since

Round	Condition	The Robot's End-of-Round Utterance
5 ✓	C	That round was completed successfully. We have been playing this game for 5 minutes and have 25 minutes remaining.
	E	Nice job!! Time for a quick joke: What do you call a train that chews gum? [pause] A chew, chew train!
13 ✘	C	One or more of us didn't build their railroad routes accurately. Of the 32 train track pieces, one or more of them were not placed correctly.
	E	Too bad. I do better with numbers than I do with shapes and paths, maybe that's true for you guys as well?
18 ✓	C	We have completed 14 rounds successfully in 18 minutes. We have 12 minutes and 12 rounds remaining.
	E	Awesome! I bet we can get the highest score on the scoreboard, just like my soccer team went undefeated in the 2014 season!
27 ✘	C	Error; we did not win that round. In the 30 seconds of the past round, at least one of the 32 railroad pieces wasn't placed correctly.
	E	Sorry guys, I made the mistake this round. I know it may be hard to believe, but robots make mistakes too.

Table 1: We provide examples of the end-of-round utterances the robot makes during the game both the Control (C) and Experimental (E) conditions. ✓ and ✘ represent success or failure of the round.

the outcomes of the rounds were fixed, each team had the same performance outcome (22/30) of the 30 rounds of the game and did not make it onto the high score board.

During each round, Echo had three opportunities to speak: 1) at the beginning of a round, 2) in the middle of a round, and 3) immediately after the team results were displayed on the tablet (more specific utterance timing can be found in Figure 2). All of Echo's utterances were predetermined, and were the same between conditions for the beginning and middle of the round utterances and different for the end of round utterance by condition. The end of round comments made by Echo are approximately equivalent in length between conditions, so the only difference between conditions is the content of the end of round utterances (examples of which can be found in Table 1). During any given round, Echo made a beginning of the round utterance with a probability of 0.5, a middle of the round utterance with a probability of 0.5, and always made a end-of-round utterance.

After the game had concluded, an experimenter entered the room and directed the participants to complete the post-experiment survey. After completing the post-experiment survey, participants received a cash payment and were debriefed on the forms of deception used in the experiment and the overall purpose of the experiment.

4.5 Measures

In order to answer our research questions, we captured a combination of questionnaire and behavioral measures. Survey measures were captured during pre- and post-experiment questionnaires. Behavioral measures were captured by having two coders categorize participants' behavioral responses from the experiment video during mistake rounds of the game. For each of the behavioral measures, the coders recorded whether or not that feature occurred at any point during the video segment (a binary evaluation), irrespective of the number of times the participant exhibited that feature.

4.5.1 Controls. In order to capture factors that would possibly influence trust-related behavior in the collaborative team, we collected measures of friendship/familiarity and extraversion by administering questionnaires to participants before and after the human-robot team interaction.

During the pre-experiment survey, participants were asked to evaluate their relationship with each of the other participants on a labeled 5-point scale ranging from (0) not having met the participant before to (4) being close friends with the participant. We also asked participants to note whether they were Facebook friends with and had the phone numbers of the other participants. For one participant's (P1) evaluation of another participant (P2), we added their rating of their relationship with the other participant (0-4) with their Facebook friend status (0 - not friends, 1 - friends, 0.5 - no Facebook account) and whether they have the other participant's phone number (0 - no, 1 - yes) for an overall score of P1's evaluation of their familiarity with P2 in the range of 0 (low familiarity) to 6 (high familiarity).

Of all of the main personality dimensions, we believed extraversion to have the highest potential to influence group dynamics and the effects we observed in this study. In the post-experiment survey, we included extraversion items, six yes/no questions, from a tested abbreviated form of the revised Eysenck personality questionnaire (EPQR-A) [9]. From these six binary questions, we obtained a cumulative rating between 0 (low extraversion) to 6 (high extraversion).

4.5.2 Manipulation Checks. We also collected measures of people's perceptions of Echo as a manipulation check for the experiment. In the post-experiment survey, we asked participants to evaluate whether Echo made self-disclosures, told personal stories, and used humor, during the interaction to verify our experimental manipulation. These items were rated on a likert scale from 1 to 7 (strongly disagree to strongly agree).

4.5.3 Measures of Team Members' Interactions with the Robot. We captured both questionnaire and behavioral measures to capture how participants interacted with Echo. In the post-experiment survey, we administered The Robotic Social Attributes Scale (RoSAS) [3]. RoSAS evaluates a person's view of a robot's warmth, competence, and discomfort with six 9-point likert scale trait evaluations per dimension. We calculated an average value for each of the three

dimensions (warmth, competence, and discomfort) for each participant from 1 (low) to 9 (high). In order to gauge the behavioral engagement of each participant with Echo after having made a mistake, we measured the presence or absence of the participant's gaze toward Echo and verbal response to Echo.

4.5.4 Measures of Team Members' Interactions with Fellow Human Team Members. We used the Team Psychological Safety Survey developed by Edmondson within the post-experiment survey to evaluate the psychological safety of each team [7]. Edmondson's psychological safety survey questions are each evaluated on a 7 point likert scale. We averaged the responses on these questions for each participant and have a resulting score from 1 (low) to 7 (high) of that participant's rating of the psychological safety of their team.

We expected a variety of reactions from participants whose Railroad Route Construction Game had forced them into making a mistake during a round in the game. We coded for the presence or absence of the following reactions for the mistake maker: distress (e.g. "Oops!", "Oh no!"), implicit or explicit admission of failure (e.g. "Oh, I lost", [shakes-head]), explaining the mistake (e.g. "The game disabled the piece I needed!"), apology (e.g. "Sorry guys"), and looking at fellow human team-members. In most cases, participants displayed several of these reactions after discovering their mistake.

We expected a variety of reactions from participants who observed their teammate experience a failure in the Railroad Route Construction Game. We coded for the presence or absence of the following reactions for the non-mistake maker: verbal search for the mistake making player (e.g. "Which one of you failed?"), blame of the mistake making player (e.g. "It's your fault"), consoling the mistake making player (e.g. "It's ok"), blaming the game itself (e.g. "It just does that, taking away the pieces you need"), and advice (e.g. "When I start a round I try to place the rarest pieces first").

Since we expected participants to display behaviors related to tension when mistakes were made in the game, we adopted the Specific Affect Coding System (SPAFF) coding scheme for tension and tension released by humor (tense humor) [4]. Behaviors coded under the category 'tension' include: fidgeting (e.g. repeated touching of one's clothes or hands, touching or rubbing one's face, lip biting), shifting (moving around in one's seat), speech disturbance (e.g. repetitive 'ums' or 'ahs' within an utterance, stuttering), individual smiling (smiling while not connecting with other group members), and individual laughing (laughing while not connecting with other group members). Behaviors coded under the 'tension released by humor' category include: tense joking (e.g. awkward or tense sarcastic remarks, puns, jokes), shared smiling (smiling while looking at another member in the group, who is also smiling), and shared laughing (laughing at the same time as other members in the group). Any humorous comments made without a tense nature or about an off-topic subject were not considered to be tension released by humor.

5 RESULTS

Of the 44 groups (132 participants) recruited for this study, 9 groups were excluded for one of the following reasons: video data recording failure, participant non-compliance, and substantial hardware / software failures (mostly involving a 'freeze' in the tablet game, requiring an experimenter to restart the game). Of the 35 groups (105 participants) included in the analysis, 18 groups (54 participants) were in the experimental condition and 17 groups (51 participants) were in the control condition. There were 26 male and 28 female participants in the experimental condition, with an average age of 20.13 ($SD = 7.13$). There were 15 male and 36 female participants in the control condition, with an average age of 21.33 ($SD = 11.00$).

For each of the video coded variables discussed in the results section, two coders evaluated these variables for an overlap set of 4 groups, for a total of 96 coded evaluations (4 groups * 3 participants * 8 mistake rounds) for each variable for the overlap set. The average inter-rater reliability rating, Cohen's kappa (κ), for all of the variables coded was 0.93. For each of the variables discussed in the results section, we also list Cohen's kappa for that individual variable.

To conduct this analysis, we used a multilevel mixed-effects generalized linear model to evaluate continuous dependent variables and multilevel mixed-effects logistic regression to evaluate binary dependent variables. We used these models to analyze our data because each participant's data cannot be treated as wholly independent from the other participants within their group and the data has repeated measures (since each participant is evaluated for the same measure for each of the 8 mistake rounds). In the analysis, the experimental condition and mistake round number were treated as fixed effects and the groups of participants belonged to were evaluated as a random effect. Covariates were treated as fixed effects: age, gender, familiarity, and extraversion. An independent covariance structure was used for all regressions. These models produce a coefficient (c) to linearly or logistically map the predictor (independent) variables with the dependent variable and a p value to indicate the significance of this relationship. The coefficient is presented in odds ratios, the odds of the dependent variable occurring in the experimental group over the control group.

5.1 Manipulation Checks

In order to confirm that participants' experience with Echo was consistent with the design of the experiment, we examined participants' rating of expressions of vulnerability as a manipulation check. Participants rated Echo as making significantly more self-disclosures in the experimental condition ($M = 5.19, SD = 1.59$) than the control condition ($M = 2.29, SD = 1.65, c = 18.014, p < 0.001$), as telling significantly more personal stories in the experimental condition ($M = 6.44, SD = 1.06$) than the control condition ($M = 1.65, SD = 0.98, c = 116.255, p < 0.001$), and as having significantly more expressions of humor in the experimental condition ($M = 6.22, SD = 1.02$) than the control condition ($M = 3.61, SD = 2.05, c = 12.171, p < 0.001$). These results confirm that participants correctly perceived Echo's behavior based on their experiment condition.

5.2 Interaction of Team Members with the Robot

In order to answer our first research question, addressing the influence social robot vulnerability has on human team mate behavior toward the social robot, we investigated participant's questionnaire perceptions of Echo and participant behavior expression directed

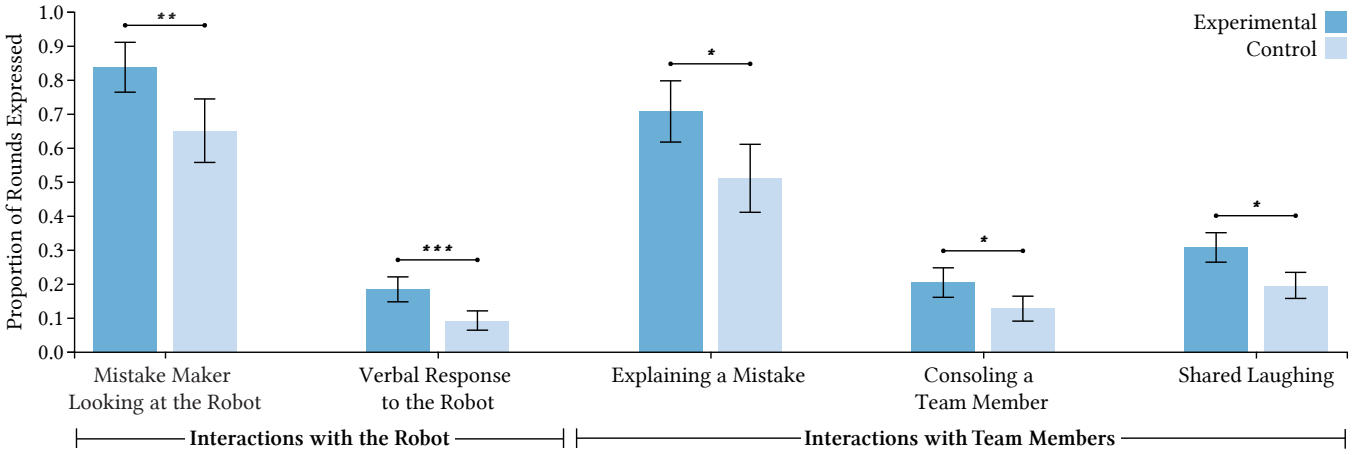

Figure 3: We report results on the differences between condition for trust-related behaviors expressed by the participants. (*), (**), and (***) denote $p <= 0.05$, $p <= 0.01$, and $p <= 0.001$, respectively.

toward Echo during the mistake rounds. From the RoSAS questionnaire, participants rated Echo as having significantly higher warmth (happy, feeling, social, organic, compassionate, and emotional) in the experimental condition ($M = 6.13, SD = 1.15$) than participants in the control condition ($M = 4.95, SD = 1.71, c = 3.273, p < 0.001$).

For the behavioral measures, we found that participants who had made an error looked at Echo after the round concluded more often in the experimental condition ($M = 0.84, SD = 0.37$) compared with participants in the control condition ($M = 0.65, SD = 0.48, \kappa = 0.99, c = 3.412, p = 0.008$). Additionally, participants were significantly more likely to respond to Echo's end-of-round comment in the experimental condition ($M = 0.18, SD = 0.39$) than the control condition ($M = 0.09, SD = 0.29, \kappa = 0.96, c = 2.513, p = 0.001$). Examples of participants' responses to Echo include the following: "*Sure*," "*Yeah, that's true*," and "*It's your fault!*" Results of these behavioral measures are depicted in Figure 3. These findings show that increased vulnerability by a social robot increases both the ratings of warmth of the robot and the engagement of human teammates with the robot, demonstrated by both nonverbal and verbal behavior expressed by the human teammates toward the robot.

To investigate a possible cause for this increased engagement, we examined participants' written evaluations of the verbal statements made by Echo and found a distinct difference in participant responses by condition. Participants in the experimental condition often noted how Echo eased the tension the groups experienced and was generally encouraging, saying that Echo's comments, "*felt kind of artificial* [...] *but they were able to ease a little tension with the efforts to make jokes*," "*they were positive and helped when we didn't succeed*," and "*they were funny, and broke the silence many times*." Participants in the control condition had a slight negative connotation with the utterances Echo made, saying Echo's comments "*constantly told [us] how many rounds [were] left, how many mistakes we made, etc. it really stressed me out*" and "*sometimes judgmental when someone would make a mistake, but the statements themselves were pretty objective and fair*." From these responses, it seems likely that participants viewed Echo as more approachable and less judgmental in the experimental condition versus the control condition. This approachability of Echo could possibly explain the increased behavioral engagement we observed with participants interacting with Echo in the experimental condition.

5.3 Interaction of Team Members with Fellow Human Team Members

To address our second research question about whether social robot vulnerability affects human team members' trust-related interactions with fellow team members, we look into the psychological safety ratings by team members, the content of team members' utterances after a mistake has been made, and other verbal expressions of team members (laughing/smiling).

To begin, we did not find any significant difference in the psychological safety survey measure between participants in the experimental condition ($M = 5.62, SD = 0.75$) and control condition ($M = 5.53, SD = 0.73, c = 1.117, p = 0.496$). This may be because the Psychological Safety questionnaire was developed for established teams in the workplace, and is not as well suited for teams with low familiarity and experience with one another.

For the behavioral data, we report the results of our statistical analysis for all of our measures in Table 2[1]. Of all of these behaviors we investigated, we observed significant differences between conditions for the following trust-related behaviors: explaining a mistake, consoling a team member, and shared laughing. These results are depicted in Figure 3. After having made a mistake, participants in the experimental condition were significantly more likely to explain their mistake (e.g. "*yeah, I can't do it, I don't have the right pieces*") to their team members ($M = 0.71$) than participants in the control condition ($M = 0.51, \kappa = 0.98, c = 3.085, p = 0.039$). Additionally, participants in the experimental condition were significantly more likely to console their team members ($M = 0.20$), saying phrases like, "*it's ok*," than participants in the control condition ($M = 0.13, \kappa = 0.89, c = 3.085, p = 0.039$). Finally, participants experienced significantly more instances of shared laughing in the

[1]We disclude the analysis of the variables 'tension: shifting' and 'tension: speech disturbance' (Section 4.5.4) from Table 2 due to the infrequency of their expression.

	$\overline{x_E}$	$\overline{x_C}$	κ	c	p
Mistake Maker Behaviors					
Mistake Maker Distress	0.52	0.48	0.99	1.190	0.761
Admission of Failure	0.87	0.87	1.00	0.439	0.617
Explaining a Mistake*	0.71	0.51	0.98	3.085	0.039
Apology for a Mistake	0.14	0.14	1.00	1.216	0.657
Looking at Human Players	0.68	0.65	1.00	0.442	0.179
Non-Mistake Maker Behaviors					
Search for Mistake	0.03	0.06	0.96	0.515	0.149
Blame the Game	0.06	0.05	0.94	1.309	0.521
Blame the Mistake Maker	0.06	0.12	0.96	0.495	0.172
Consoling*	0.20	0.13	0.89	3.085	0.039
Giving Advice	0.03	0.02	0.86	1.917	0.427
Tension Behaviors					
Fidgeting	0.57	0.53	0.83	1.224	0.448
Individual Smiling	0.63	0.58	0.73	3.620	0.324
Individual Laughing	0.27	0.24	0.76	1.594	0.251
Tension Released By Humor Behaviors					
Joking	0.05	0.05	1.00	1.088	0.847
Shared Smiling	0.41	0.31	0.98	2.164	0.133
Shared Laughing*	0.31	0.19	0.86	2.335	0.041

Table 2: For all of the behavioral measures we examined from the data we show the mean in the experimental condition ($\overline{x_E}$), mean in the control condition ($\overline{x_C}$), coefficient (c), p-value, and Cohen's kappa (κ). Asterisks (*) by the variable names are used to denote significant ($p < 0.05$) variables.

experimental condition ($M = 0.31$) compared with the control condition ($M = 0.19, \kappa = 0.86, c = 2.335, p = 0.041$) after a mistake was made by one of the participants.

One possible interpretation of these results is that participants in the experimental condition, as compared with the control condition, were more likely to make verbal attempts to ease the tension experienced by the group due to the mistake. There was no significant difference reported between the experimental condition and control condition for the amount of tension present. However, the responses of team members explaining mistakes, consoling team members, and shared laughing were more present in the experimental condition than the control condition. Thus, it is possible that participants in both conditions experienced tension after a mistake, however, due to the vulnerable statements by Echo, participants in the experimental condition were more likely to ease the tension through explaining the mistake, consoling team members, and laughing together.

6 DISCUSSION

In this work, we have examined the trust-related behavioral effects of social robot vulnerability on human members of a human-robot team. Our results have demonstrated increased engagement toward the robot and increased trust-related behavior expression (explaining errors, consoling other team members, and shared laughing) toward fellow team members when the robot in the group makes vulnerable versus neutral statements.

Barsade's "Ripple Effect" study demonstrated the ability of an individual's positive behavior to influence other individuals in a group to, in turn, express more positive behavior [2]. In this study of robot vulnerability, we observe a similar "ripple effect" where a robot's vulnerable behavior influenced the expression of trust-related behaviors expressed by humans in a human-robot team. The "ripples" of the robot's vulnerable behavior influences not only 1) team members' interactions with the robot, but also also 2) team members' *human-human trust-related interactions* with each other. Human team members expressed more vulnerability in easing the tension after mistakes by explaining the mistake if they had made it, consoling fellow team members who did make mistakes, and laughing together. This increase in trust-related human behavior displays the distinctive influence social robot vulnerability has on trust-related team human-human behavior.

Team-based trust and vulnerability not only lead to the easing of tension through positive social behaviors, but also drive team productivity and success. Edmondson's work on psychological safety (the belief that an individual can take risks, express vulnerability, and be listened to without facing social condemnation or judgment) has shown that learning behavior (e.g. seeking feedback, discussing errors, and learning from mistakes) mediates the relationship between team psychological safety and team performance [7]. Thus, vulnerable behavior expression by robots may likely influence the *performance* of a human-robot team in addition to impacting team member's trust-related behavior expression during tense situations. We were not able to explore the effects of vulnerability on team performance in this study because we fixed team performance to study team members' reactions in an equivalent number of tense scenarios (when mistakes occurred). However, we believe exploring the effects of robot vulnerability and group trust-related behavior on team performance will be a fruitful area of future research.

7 CONCLUSION

In this study, we investigated the effects of a robot's vulnerable behavior on trust-related interactions between team members and the robot as well as team members with fellow human team members on a human-robot team. We programmed an autonomous robot to play a collaborative game with a group of three human participants, where each participant would make mistakes throughout the game that negatively impacted team performance. We compared the behavior of group members during these tense moments (when mistakes are made) between groups with a robot who made vulnerable statements versus a robot who made neutral statements. Participants in the group with a robot who made vulnerable statements engaged to a higher degree with the robot and displayed a "ripple effect" of the robot's vulnerable behavior by displaying more trust-related behaviors with their other human teammates (explaining a mistake, consoling team members, and laughing together). These results demonstrate the positive influence robots can have on trust in human-robot teams.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation award number 1139078. We also thank Adam Erickson for assisting with the design of this experiment.

REFERENCES

[1] Sean Andrist, Erin Spannan, and Bilge Mutlu. 2013. Rhetorical robots: making robots more effective speakers using linguistic cues of expertise. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 341–348.

[2] Sigal G Barsade. 2002. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly* 47, 4 (2002), 644–675.

[3] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The Robotic Social Attributes Scale (RoSAS): Development and Validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 254–262.

[4] James A Coan and John M Gottman. 2007. The specific affect coding system (SPAFF). *Handbook of emotion elicitation and assessment* (2007), 267–285.

[5] Paul C Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin* 79, 2 (1973), 73.

[6] David DeSteno, Cynthia Breazeal, Robert H Frank, David Pizarro, Jolie Baumann, Leah Dickens, and Jin Joo Lee. 2012. Detecting the trustworthiness of novel partners in economic exchange. *Psychological science* 23, 12 (2012), 1549–1556.

[7] Amy Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative science quarterly* 44, 2 (1999), 350–383.

[8] 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.

[9] Leslie J Francis, Laurence B Brown, and Ronald Philipchalk. 1992. The development of an abbreviated form of the Revised Eysenck Personality Questionnaire (EPQR-A): Its use among students in England, Canada, the USA and Australia. *Personality and individual differences* 13, 4 (1992), 443–449.

[10] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on*. IEEE, 106–114.

[11] 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: The Journal of the Human Factors and Ergonomics Society* 53, 5 (2011), 517–527.

[12] Gareth R Jones and Jennifer M George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review* 23, 3 (1998), 531–546.

[13] Malte F Jung, Nikolas Martelaro, and Pamela J Hinds. 2015. Using robots to moderate team conflict: the case of repairing violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 229–236.

[14] Malte F Jung, Selma Šabanović, Friederike Eyssel, and Marlena Fraune. 2017. Robots in Groups and Teams. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 401–407.

[15] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2009. An affective guide robot in a shopping mall. In *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*. IEEE, 173–180.

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

[17] Sandrn A Kiffin-Petersen and John L Cordery. 2003. Trust, individualism and job characteristics as predictors of employee preference for teamwork. *International Journal of Human Resource Management* 14, 1 (2003), 93–116.

[18] Richard J Klimoski and Barbara L Karol. 1976. The impact of trust on creative problem solving groups. *Journal of Applied Psychology* 61, 5 (1976), 630–633.

[19] Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human Expectations of Social Robots. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, Piscataway, NJ, USA, 463–464. http://dl.acm.org/citation.cfm?id=2906831.2906928

[20] Owen H Lynch. 2002. Humorous communication: Finding a place for humor in communication research. *Communication theory* 12, 4 (2002), 423–445.

[21] Nikolas Martelaro, Victoria C Nneji, Wendy Ju, and Pamela Hinds. 2016. Tell Me More: Designing HRI to encourage more trust, disclosure, and companionship. In *The Eleventh ACM/IEEE International Conference on Human Robot Interation*. IEEE Press, 181–188.

[22] 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.

[23] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 61–68.

[24] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, Vol. 3. Kobe, 5.

[25] 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 *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 141–148.

[26] Michihiro Shimada, Takayuki Kanda, and Satoshi Koizumi. 2012. How can a Social Robot facilitate childrenâĂŹs collaboration? *Social Robotics* (2012), 98–107.

[27] Rosanne M Siino, Justin Chung, and Pamela J Hinds. 2008. Colleague vs. tool: Effects of disclosure in human-robot collaboration. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*. IEEE, 558–562.

[28] Tony L Simons and Randall S Peterson. 2000. Task conflict and relationship conflict in top management teams: the pivotal role of intragroup trust. *Journal of applied psychology* 85, 1 (2000), 102.

[29] Christi McGuffee Smith and Larry Powell. 1988. The use of disparaging humor by group leaders. *Southern Speech Communication Journal* 53, 3 (1988), 279–292.

[30] Sarah Strohkorb, Ethan Fukuto, Natalie Warren, Charles Taylor, Bobby Berry, and Brian Scassellati. 2016. Improving human-human collaboration between children with a social robot. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 551–556.

[31] Gerben A van Kleef. 2016. *The interpersonal dynamics of emotion*. Cambridge University Press.

[32] Gerben A Van Kleef, Carsten KW De Dreu, and Antony SR Manstead. 2010. An interpersonal approach to emotion in social decision making: The emotions as social information model. *Advances in experimental social psychology* 42 (2010), 45–96.

[33] Marynel Vázquez, Elizabeth J Carter, Braden McDorman, Jodi Forlizzi, Aaron Steinfeld, and Scott E Hudson. 2017. Towards Robot Autonomy in Group Conversations: Understanding the Effects of Body Orientation and Gaze. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 42–52.

[34] Karl E Weick. 1993. The collapse of sensemaking in organizations: The Mann Gulch disaster. *Administrative science quarterly* (1993), 628–652.

[35] Lawrence R Wheeless. 1978. A follow-up study of the relationships among trust, disclosure, and interpersonal solidarity. *Human Communication Research* 4, 2 (1978), 143–157.

[36] Dale E Zand. 1972. Trust and managerial problem solving. *Administrative science quarterly* (1972), 229–239.