

## Abstract

# Developing Robot Teammates that Enhance Social Dynamics and Performance in Human-Robot Teams

Sarah Strohkorb Sebo

2020

Collaborative teams of people are most successful when they have positive social dynamics, where team members trust one another [Jones and George, 1998], feel included [Shore et al., 2011], and feel comfortable to openly discuss mistakes and errors [Edmondson, 1999]. Some of these social dynamics have even been shown to be more correlated with team performance than measures of individual intelligence [Woolley et al., 2010]. As robots increasingly join collaborative teams of people in a variety of settings (e.g., manufacturing plants, surgical suites, corporate workplaces, homes), it is essential that we build robots that can perceive and positively influence these social dynamics for the benefit of the team.

The work in this dissertation explores how social robots can enhance important social team dynamics in collaborative human-robot teams. It specifically investigates how robot behavior can positively shape trust, inclusion, and psychological safety, social dynamics that have been shown to have a significant positive influence on team performance. Our work demonstrates that a robot’s behavior can influence not just how people interact with the robot, but how people in the group interact with each other. Finally, we investigate a way in which robots may be able to sense social dynamics in real time, by perceiving human backchanneling behavior, and how social robots can shape this human social behavior in human-robot teams.

We investigate several robot behaviors (e.g., expressions of vulnerability, verbally supportive utterances) that have never previously been explored in the context of multi-person human-robot teams and demonstrate their ability to influence the team’s trust, inclusion, and psychological safety. We first explore the benefits of a robot asking pairs of children task-focused and relationship-focused questions to improve overall collaborative skill in a collaborative task, finding contrasting effects on task performance and perceptions of task

performance. Next, we examine how to make the most effective trust repair in the context of a one-on-one human-robot interaction and highlight the benefits of a robot apologizing for mistakes. Then we extend this idea to a group context and investigate how a robot’s expressions of vulnerability (including apologizing for and admitting to mistakes) might influence trust-related behavior in groups, showing that robot vulnerability increases trust-related behavior expression by people within the group towards one another as well as their conversational dynamics. Following this, we explore two strategies to improve human team member inclusion in a human-robot team and highlight the negative effects of giving someone in the group a specialized role to interact with the robot and the benefits of robot verbal support. Finally, we further analyze the role of robot verbal support within a human-robot team, demonstrating that robot verbal support may reduce the amount of verbal backchannel support human team members give one another.

We also suggest a method for perceiving group dynamics in real time through the recognition of human verbal backchannels from audio signals. Based on the data collected in one of our human-subjects experiments, we found significant positive correlations between verbal backchanneling behavior and participant ratings of psychological safety and inclusion. Thus, we believe that sensing human verbal backchannels can be a useful input for quickly perceiving social dynamics, enabling robots and other artificial agents to adapt their behavior based on real-time changes in social team dynamics.

Taking into account all of our findings, we propose a set of guidelines for social robot behavior use in collaborative human-robot teams. In these guidelines, we recommend social robot behavior for particular situations and contexts, including tense interactions, robot errors and mistakes, teams where one or more members may have a specialized role, and interactions where someone feels excluded. The combination of these guidelines, the results from our experimental studies, and the connections we made between backchannels and team social dynamics make a significant contribution to building robots that can positively shape the social dynamics and performance of human-robot teams.

# Developing Robot Teammates that Enhance Social Dynamics and Performance in Human-Robot Teams

A Dissertation  
Presented to the Faculty of the Graduate School  
of  
Yale University  
in Candidacy for the Degree of  
Doctor of Philosophy

by  
Sarah Strohkorb Sebo

Dissertation Director: Brian Scassellati

December 2020

Copyright © 2020 by Sarah Strohkorb Sebo  
All rights reserved.

# Contents

<b>Acknowledgements</b>	<b>xlii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 A Review of Social Dynamics in Collaborative Human Teams and Human-Robot Teams</b>	<b>6</b>
2.1 Human Collaborative Teams . . . . .	7
2.1.1 Human Teams . . . . .	7
2.1.2 Trust . . . . .	10
2.1.3 Inclusion . . . . .	11
2.1.4 Psychological Safety . . . . .	12
2.2 Human-Robot Collaborative Teams . . . . .	13
2.2.1 Robots in Groups and Teams: A Literature Review . . . . .	14
2.2.2 Trust in HRI . . . . .	29
2.2.3 Inclusion & Psychological Safety in HRI . . . . .	30
2.3 Summary . . . . .	31
<b>3 Robots that Shape Collaboration between Pairs of Children</b>	<b>32</b>
3.1 Introduction . . . . .	33
3.2 Methods . . . . .	34
3.2.1 Participants . . . . .	36
3.2.2 Build-a-Rocket Game . . . . .	37
3.2.3 System Architecture . . . . .	38

3.2.4	Procedure . . . . .	40
3.2.5	Measures . . . . .	41
3.3	Results . . . . .	44
3.3.1	Performance Outcome . . . . .	45
3.3.2	Perception of Performance . . . . .	46
3.3.3	Perception of Interpersonal Cohesiveness . . . . .	46
3.4	Discussion . . . . .	47
3.5	Summary . . . . .	48
<b>4</b>	<b>Robots that Shape Trust in the Aftermath of a Robot Trust Violation</b>	<b>50</b>
4.1	Introduction . . . . .	51
4.2	Background . . . . .	53
4.2.1	Human-Human Trust Repair . . . . .	53
4.2.2	Human-Robot Trust . . . . .	55
4.3	Methods . . . . .	55
4.3.1	Space Shooting Tablet Game . . . . .	55
4.3.2	Experimental Conditions . . . . .	57
4.3.3	Procedure . . . . .	58
4.3.4	Measures . . . . .	60
4.3.5	Participants . . . . .	61
4.4	Results . . . . .	61
4.4.1	Participant Power-Up Choices . . . . .	61
4.4.2	Trust-Related Survey Responses . . . . .	64
4.4.3	The Influence of Participant Promises on Trust . . . . .	67
4.5	Discussion . . . . .	68
4.6	Summary . . . . .	71
<b>5</b>	<b>Robots that Shape Group Trust and Communication through Vulnerable Expressions</b>	<b>74</b>
5.1	Introduction . . . . .	75
5.2	Background and Research Questions . . . . .	77

5.3	Methods . . . . .	79
5.3.1	Experimental Conditions . . . . .	79
5.3.2	Collaborative Interaction System Setup . . . . .	80
5.3.3	Railroad Route Construction Tablet Game . . . . .	80
5.3.4	Procedure . . . . .	82
5.3.5	Measures . . . . .	85
5.3.6	Participants . . . . .	89
5.4	Results . . . . .	90
5.4.1	Manipulation Checks . . . . .	90
5.4.2	Participant Perceptions of and Interactions with the Robot . . . . .	92
5.4.3	Participant Interactions with Fellow Team Members during Failure Rounds . . . . .	96
5.4.4	Participant Conversation Dynamics throughout the Game . . . . .	99
5.4.5	Perceptions of Group Dynamics . . . . .	104
5.5	Discussion . . . . .	105
5.6	Summary . . . . .	106
<b>6</b>	<b>Robots that Shape the Inclusion of Human Team Members</b>	<b>109</b>
6.1	Introduction . . . . .	110
6.2	Background and Research Questions . . . . .	111
6.2.1	Strategy 1: Specialized Roles in Groups . . . . .	112
6.3	Methods . . . . .	113
6.3.1	Experiment Design . . . . .	113
6.3.2	Collaborative Task: The Survival Problem . . . . .	115
6.3.3	Robot Platform and Behaviors . . . . .	116
6.3.4	Procedure . . . . .	117
6.3.5	Measures . . . . .	118
6.3.6	Participants . . . . .	121
6.4	Results . . . . .	122
6.4.1	Ingroup-Outgroup Differences . . . . .	122

6.4.2	Influence of the Robot Liaison Role . . . . .	124
6.4.3	Influence of the Robot’s Supportive Utterances . . . . .	125
6.4.4	Perceptions of the Robot . . . . .	128
6.5	Discussion . . . . .	129
6.6	Summary . . . . .	131

## **7 Human Backchannels: Signals of Key Group Dynamics that can be Influenced by Social Robots 133**

7.1	Introduction . . . . .	134
7.2	Background . . . . .	135
7.2.1	Backchanneling in Human Teams . . . . .	135
7.2.2	Systematic Analysis and Prediction of Backchannels and Backchannel Opportunities . . . . .	136
7.2.3	Robots Backchanneling in Human-Robot Interactions . . . . .	137
7.3	Methods . . . . .	138
7.3.1	Hypotheses . . . . .	138
7.3.2	Experiment Design . . . . .	139
7.3.3	Collaborative Task . . . . .	141
7.3.4	Robot Behavior . . . . .	141
7.3.5	Protocol . . . . .	143
7.3.6	Measures . . . . .	144
7.3.7	Participants . . . . .	146
7.4	Results . . . . .	146
7.4.1	Verbal and Nonverbal Backchannels . . . . .	147
7.4.2	Psychological Safety and Perceived Inclusion Scores . . . . .	148
7.4.3	Connections between Human Backchannels and Social Group Dynamics - Individual Level . . . . .	149
7.4.4	Connections between Human Backchannels and Social Group Dynamics - Group Level . . . . .	153
7.4.5	The Influence of Intergroup Bias on the Reception of Backchannels . . . . .	154



7.4.6	The Influence of Gender on the Production of Backchannels . . . . .	155
7.4.7	Influence of Robot Verbal Support on Human Backchanneling Behavior	155
7.4.8	Influence of Robot Verbal Support on Team Social Dynamics . . . . .	157
7.5	Discussion . . . . .	158
7.6	Summary . . . . .	160
<b>8</b>	<b>Discussion</b>	<b>162</b>
8.1	Design Guidelines for Social Robot Behavior in Collaborative Human-Robot Teams . . . . .	162
8.2	Central Themes . . . . .	164
8.2.1	Robot Behavior Can Influence Human-to-Human Behavior within Human-Robot Collaborative Teams . . . . .	164
8.2.2	Robot Interactions with Groups and Teams of People . . . . .	165
8.2.3	Robot Perception of team social dynamics . . . . .	166
8.2.4	Experiment Designs that Enable Investigation into Specific Team Dynamics . . . . .	167
8.2.5	Focus on Measuring Human Behavior within Groups . . . . .	169
8.3	Open Questions and Future Directions . . . . .	171
8.3.1	Computational Decision Making Models for Influencing Team Social Dynamics . . . . .	171
8.3.2	Unexplored Methods for Robots to Improve Team Social Dynamics .	172
8.3.3	Deployment of Robust Robot Teammates in Real-World Settings . .	174
8.3.4	Ethical Considerations . . . . .	175
8.4	Summary . . . . .	177
<b>9</b>	<b>Conclusion</b>	<b>178</b>
<b>A</b>	<b>Additional Methodology Details</b>	<b>180</b>
A.1	Chapter 5: All End-of-Round Utterances for the Vulnerable and Neutral Conditions . . . . .	180

A.2	Chapter 5: Video Coding Scheme for Participant Responses during Failure Rounds . . . . .	185
A.3	Chapter 5: Video Coding Scheme for Participant Utterances . . . . .	189
A.4	Chapters 6 and 7: Participant Instruction Sheets . . . . .	191
A.4.1	Round 1: Participant Instruction Sheet . . . . .	191
A.4.2	Round 2: Participant Instruction Sheet . . . . .	193
A.5	Chapters 6 and 7: Robot Utterances . . . . .	195
A.5.1	Round Introductions . . . . .	196
A.5.2	Query Responses . . . . .	196
A.5.3	Targeted Supportive Utterances . . . . .	201
A.5.4	Survival Item Hints . . . . .	203
A.5.5	Generic Response/Backchannels to Participant Speech . . . . .	206
<b>B</b>	<b>Human-Subjects Study Questionnaires</b>	<b>208</b>
B.1	Friendship, Familiarity, and Trust Survey for Children . . . . .	208
B.2	Build-a-Rocket Reflection Survey for Children . . . . .	209
B.3	Dyadic Trust Scale Survey . . . . .	210
B.4	Robotic Social Attributes Scale (RoSAS) Survey . . . . .	211
B.5	Friendship and Familiarity Scale for Adults . . . . .	213
B.6	The Abbreviated Form of the Revised Eysenck Personality Questionnaire (EPQR-A): Extraversion . . . . .	214
B.7	The Team Psychological Safety Scale . . . . .	215
B.8	The Short Form of the Trait Emotional Intelligence Questionnaire (TEIQue-SF) . . . . .	216
B.9	The Perceived Inclusion Scale . . . . .	218
<b>C</b>	<b>Detailed Statistical Results</b>	<b>219</b>

# List of Figures

2.1	Descriptive examples of robots interacting with human groups: (a) a Furhat robot completing a sorting task with two people in a museum [Skantze et al., 2015], (b) two EMYS robots and two people playing a card game in a lab setting [Correia et al., 2018b], (c) a Robovie robot guiding people in a shopping mall [Shiomi et al., 2010], and (d) our own work exhibiting a Nao robot playing a collaborative game with three people in a lab setting [Strohkorb Sebo et al., 2018]. . . . .	15
2.2	The number of studies investigating robots interacting with multiple people has steadily grown over the past couple of decades. The shaded area around the line of best fit represents a 95% confidence interval. . . . .	19
2.3	We display the (a) composition of human-robot groups studied in the literature as well as (b) the most commonly used robots in these studies and the (c) control methods for these robots. . . . .	20
2.4	In the studies we review on robots interacting with human groups and teams, we highlight the (a) countries where the studies were run, (b) the number of groups per between-subjects condition in the experimental studies, and (c) the number of interaction sessions in the experimental studies. . . . .	21
2.5	We visualize (a) the number of studies conducted in the lab and in the field and the robot roles found within each setting and (b) the number of studies conducted in specific field setting and the robot roles found within each field setting. . . . .	25

3.1	Pairs of children age 6-9 collaborated with one another to play a rocket building game with a social robot that asked them relationship-focused questions, task-focused questions, or no questions. . . . .	34
3.2	In the build-a-rocket game, players drag and drop pieces to construct a rocket by optimizing weight, fuel, air resistance, and power metrics (shown on the bottom panel). Time remaining to takeoff is shown in the upper left hand corner. Players can drag a piece over the white question mark to ask the robot about that piece's weight. . . . .	37
3.3	A MyKeepon robot interacts with two children playing the build-a-rocket game on the touchscreen game interface. The robot uses information gathered from a Microsoft Kinect to track the faces of the children to inform its gaze direction. . . . .	39
3.4	There were 43 pairs of children that participated in this experiment. These photos depict several pairs of children interacting with each other as they play the collaborative build-a-rocket game. . . . .	40
3.5	For each condition (relational, task, and control) we display the (a) teams' average maximum rocket height scores and (b) the scores of participants' perception of their performance in the build-a-rocket game. Error bars represent a 95% confidence interval. . . . .	45
4.1	Participants played a competitive game with a robot, where the robot violated and then tried to repair the participants' trust. . . . .	52
4.2	Participants played the Space Shooting tablet game with a robot named Echo where they tried to gain points by shooting asteroids. . . . .	56

4.3	During the 10 rounds of the game, the robot and participant receive power-ups. Before round 3, the robot delivers a promise not to immobilize the participant. During round 3 the robot receives a power up, chooses to immobilize the participant, and verbally reacts to the choice. After round 3 concludes, the robot tries to repair the trust of the participant. The power-ups in the following rounds are used to measure the participant's responses to the robot's actions. The utterances of the robot in this figure are consistent with those in the competence-apology condition. . . . .	57
4.4	For the first power-up choice, participants were significantly more likely to immobilize the robot with the integrity trust violation framing and the denial trust repair strategy. . . . .	62
4.5	The power-up choices of participants over time was significantly influenced by the trust violation framing. . . . .	63
4.6	An interaction effect was found between the trust violation framing and trust repair strategy on participant ratings of trust in the robot. . . . .	65
4.7	Participants' responses to a survey question about which factors influenced their power-up decisions were coded as strategy, retaliation, and/or consideration of the robot. This data is also grouped by the three most dominant power-up choice sequences. . . . .	66
5.1	Participants played a collaborative game with a robot, who made (1) vulnerable statements, (2) neutral statements, or (3) remained silent at the end of each round of the game. . . . .	76
5.2	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. . . . .	81
5.3	In the experiment, three human participants and a Nao robot played a collaborative game on individual tablets. . . . .	83

5.4	The robot was viewed as warmer if it made vulnerable utterances than either neutral or no utterances (silent), and warmer if it made neutral utterances as opposed to no utterances. The robot was viewed as more competent if it made vulnerable or neutral utterances as opposed to no utterances (silent). The robot was viewed as causing more discomfort if it did not make any utterances (silent) when compared with the neutral condition. (*) and (**) denote $p < 0.05$ and $p < 0.01$ respectively. Error bars represent a 95% confidence interval. . . . .	93
5.5	Participants interacting with a vulnerable robot were more likely to look at the robot after having made a mistake and were more likely to verbally respond to the robot than participants interacting with a neutral or silent robot. (*), (**), and (***) denote $p < 0.05$ , $p < 0.01$ , and $p < 0.001$ , respectively. Error bars represent a 95% confidence interval. . . . .	95
5.6	Participants interacting with a vulnerable robot were more likely to explain their mistake to their human teammates and console a human teammate who had made a mistake than participants interacting with a neutral or silent robot. Also, participants interacting with a vulnerable robot were more likely to laugh together than participants interacting with a neutral robot. (*) and (**) denote $p < 0.05$ and $p < 0.01$ respectively. Error bars represent a 95% confidence interval. . . . .	97
5.7	Compared to the neutral and silent conditions, human participants in the vulnerable condition spoke more, in total, to the other participants in their group, and increasingly across game rounds. In (a), we see that participants in the vulnerable condition spoke significantly more than participants in either the neutral or silent conditions ( $n = 153$ participants). In (b), the line widths represent the amount of talking by human participants toward their teammates who are connected by the line, in seconds (summed across all groups within a condition ( $n = 153$ participants)). R = robot; P1, P2, and P3 = human participants, in their relative positions around the table. (**) denotes $p < 0.01$ and error bars represent a 95% confidence interval. . . . .	100

5.8	Here, we examine how much time participants in the experimental conditions spent talking during the 30 rounds of the game. In (a), the vulnerable condition has more talking in every round, and the slope (i.e., the rate of increase in talking per round, across rounds) is higher than the neutral condition (but indistinguishable from the silent condition). In (b) we see that, compared to the neutral condition, those in the vulnerable condition respond more over time to their fellow human group members ( $n = 4,590$ rounds). . . . .	101
5.9	Although there was (a) no statistical difference between the vulnerable and neutral conditions in the equality in talking time ( $n = 150$ participants; one group did not speak at all and was excluded), the silent condition had less equality in talking time when compared with the other two conditions. In (b), we see that participants in the vulnerable condition directed their utterances more equally to each of their human group members than participants in the silent condition, as measured by the total amount of time spent talking to each participant's two human partners ( $n = 144$ participants; participants who didn't speak at all or who did not make directed utterances were excluded). (***) denotes $p < 0.001$ and error bars represent a 95% confidence interval. . . . .	102
6.1	Three participants completed a collaborative task with a Jibo robot, where the robot used two distinct strategies to enhance the inclusion of the human team members. . . . .	111
6.2	In round 1 of the experiment, two participants (ingroup) and a Jibo robot completed a task in room A while one participant (outgroup) and a Jibo robot completed the same task in room B. The outgroup participant joined the two ingroup participants and the robot in room A for round 2 of the experiment, where one of the members is designated the robot liaison. . . .	114
6.3	Participants who were the robot liaison had a lower perceived group inclusion than the other group members. Error bars represent a 95% confidence interval.	123

6.4	Outgroup participants, as opposed to ingroup participants, displayed a significantly higher difference in the proportion of time they spent talking during the one minute after the robot’s support targeted to the participant (RST-P) when compared with two baselines: 1) the proportion of time they spent talking during the one minute after the robot support was targeted to someone else (RST-SE) and 2) the proportion of time they spent talking during the the one minute after a robot undirected utterance (RUU). Error bars represent a 95% confidence interval. . . . .	126
7.1	Three participants and a Jibo robot completed a collaborative task. We annotated the backchannels made by the human participants (e.g., “yeah,” head nodding) and analyzed whether these backchannels were correlated with important group dynamics (psychological safety, inclusion) and how the human participant backchanneling behavior was shaped by the robot. . . . .	135
7.2	In this experiment, we employed a 2 (robot liaison: ingroup or outgroup) x 2 (robot verbal support: present or absent) between subjects design. . . . .	141
7.3	In this graph, each data point represents one participant and its x and y values represent the total time that participant spent producing verbal backchannels and nonverbal backchannels, respectively. We also plotted a line of best fit, which has a slope of 0.80. . . . .	147
7.4	In this graph, each data point represents one participant and its x and y values represent the participant’s psychological safety and perceived inclusion scores, respectively. These scores were significantly and positively correlated, as shown by the positive slope (0.46) of the best fit line. . . . .	148
7.5	We display the human backchannel (BC) variables that have significant influences on psychological safety and perceived inclusion post-experiment survey scores. This analysis examined each individual participant’s backchanneling behavior and ratings of psychological safety and inclusion. . . . .	149



7.6	We display the human backchanneling variables that have significant influences on the group's average psychological safety and perceived inclusion post-experiment survey scores. This analysis examined the backchanneling behavior of each group and the average ratings of psychological safety and inclusion for each group. . . . .	153
7.7	In our analysis of the influence of our experimental factors on the verbal backchannels that participants received, we found a main effect for intergroup bias, where outgroup members received more verbal backchannels than ingroup members, and two significant interactions. (a) The first interaction we observed was between intergroup bias and the presence of robot verbal support (RVS). (b) The second interaction we found was between intergroup bias and whether or not the participant was a robot liaison (RL). (*) and (**) denote $p < 0.05$ and $p < 0.01$ respectively. Error bars represent a 95% confidence interval. . . . .	156

# List of Tables

3.1	We present the age distribution of the dyads in each experimental condition by the number of dyads in each age and experimental group. . . . .	36
3.2	We present the gender composition of the dyads in each experimental condition by the number of females (F) and males (M) in each experimental group. . . . .	36
3.3	The robot asked questions to the two child participants after each round of the game in the task and relational conditions, and did not say anything to the child participants who were in the control condition. These questions are displayed in this table, where [P1] and [P2] act as placeholders for participant names. The questions indicated by a (*) were asked to the child in the task condition who had learned specifically about that topic. . . . .	42
4.1	Each condition had a unique robot trust repair utterance. . . . .	59
5.1	We provide examples of the end-of-round utterances the robot makes during the game in the neutral (N) and vulnerable (V) conditions. The utterances in the vulnerable conditions reflect either a self-disclosure (e.g., rounds 13 and 27), a personal story (e.g., round 18), or an expression of humor (e.g., round 5). The robot does not make end-of-round utterances in the silent condition. ✓ and ✗ represent success or failure of the round. . . . .	84
5.2	Gender composition of the groups in each experimental condition by the number of females (F) and males (M) in each experimental group. . . . .	90

6.1	Examples of the targeted supportive utterances the robot made during the experiment, where [p-name] is a placeholder for the participant's name. . .	117
6.2	This table reports the proportion of survival items that were initially ranked high and low by the ingroup ( $H_{in}, L_{in}$ ) and outgroup ( $H_{out}, L_{out}$ ) that made it onto the group's final list of 8 items. When the outgroup member was the robot liaison, the team was significantly less likely to incorporate the survival items they initially valued ( $L_{in}, H_{out}$ ) onto the team's final list. . . . .	125
A.1	These utterances were said during the neutral condition in which the robot made neutral, fact-based statements. . . . .	183
A.2	These utterances were said during the vulnerable condition in which the robot made vulnerable statements. Each vulnerable statement was further categorized into one of the following: self-disclosure, personal story, or humor.	185
A.3	In response to a query to the robot about a survival item, the robot responded by giving the participant more information about that item. This table includes all of the robot's query responses to survival items. . . . .	199
A.4	In response to a query to the robot about an environment aspect, the robot responded by giving the participant information about that environment aspect. This table includes all of the robot's query responses to environment aspects. . . . .	201
A.5	This table contains all of the possible rephrase and item targeted supportive utterance templates. . . . .	202
A.6	This table contains all of the possible simple targeted supportive utterances.	202
A.7	In response to a participant utterance containing one of the survival items, the robot could respond with a hint about that item. This talbe contains all possible hints a robot could give about the survival items. . . . .	206
A.8	This table contains all of the possible robot item backchannels. . . . .	207
C.1	This table lists the demographic characteristics of the participants in the human-subjects study detailed in Chapter 3, both overall and for each condition. . . . .	220

C.2	This table presents the results from the 1-way ANOVA analysis comparing the maximum rocket height difference between the three conditions (relational, task, and control) in Chapter 3 Section 3.3. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor. . . . .	221
C.3	This table presents the results from the 1-way ANOVA analysis comparing the maximum rocket height difference between the task and control conditions (planned comparison) in Chapter 3 Section 3.3. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor. . . . .	222
C.4	This table presents the results from the 1-way ANOVA analysis comparing the maximum rocket height difference between the task and relational conditions (planned comparison) in Chapter 3 Section 3.3. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor. . . . .	223
C.5	This table presents the results from the linear mixed-effects model run in Chapter 3 Section 3.3 examining the influence of the experimental condition (reference group: relational condition) and a control for whether the pair was same or mixed gender on the participant’s perception of their team’s performance. Each participant is grouped with their partner in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	224
C.6	This table lists the demographic characteristics of the participants in the human-subjects study detailed in Chapter 4, both overall and for each condition. . . . .	225

- C.7 This table presents the results from the logistic regression model run in Chapter 4 Section 4.4.1 examining the influence of the trust violation framing and trust repair strategy on whether participants immobilized the robot in their first power-up choice. We used the R ‘glm’ function with a binomial family and logit link to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 226
- C.8 In order to analyze the power-up choices of participants over time in Chapter 4 Section 4.4.1, we used a multilevel mixed-effects logistic regression model to determine the influence of the trust violation framing and trust repair strategy on whether participants chose to immobilize the robot during their three power-up choices. To capture the repeated measures nature of the data, we used a random effect for each participant across their three power-up choices. In addition to our experimental conditions, we also controlled for the participants’ gender and the power-up choice round. We used the R ‘glmer’ function with a binomial family and logit link from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 227
- C.9 This table presents the results from the 2-way ANOVA analysis comparing the RoSAS warmth ratings between participants with different trust violation framings and trust repair strategies, as presented in Chapter 4 Section 4.4.2. In addition to our main variables of interest (trust violation framings and trust repair strategy), we include in our model the interaction between these two variables as well as controls for participant gender and age. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor. 228

C.10	This table presents the results from the 2-way ANOVA analysis comparing the Dyadic Trust Scale (DTS) ratings between participants with different trust violation framings and trust repair strategies, as presented in Chapter 4 Section 4.4.2. In addition to our main variables of interest (trust violation framings and trust repair strategy), we include in our model the interaction between these two variables as well as controls for participant gender and age. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor. . . . .	229
C.11	This table presents the results from the 2-way ANOVA analysis comparing whether the participants perceived the robot to have lied between participants with different trust violation framings and trust repair strategies, as presented in Chapter 4 Section 4.4.2. To gather participants perception of whether the robot lied, we examined participants’ ratings on a 1 (strongly disagree) to 7 (strongly agree) Likert scale on the post-experiment survey question “Echo lied to me,” where Echo is the name of the robot. In addition to our main variables of interest (trust violation framings and trust repair strategy), we include in our model the interaction between these two variables as well as controls for participant gender and age. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor. .	230

C.12	This table presents the results from the 2-way ANOVA analysis comparing whether participants believed they had made a reciprocal promise to the robot across our experimental conditions, as presented in Chapter 4 Section 4.4.2. To gather participants belief of whether or not they made a reciprocal promise to the robot, we examined participants' ratings on a 1 (strongly disagree) to 7 (strongly agree) Likert scale on the post-experiment survey question "I promised not to immobilize Echo during the game," where Echo is the name of the robot. In addition to our main variables of interest (trust violation framings and trust repair strategy), we include in our model the interaction between these two variables as well as controls for participant gender and age. This analysis was performed using the 'aov' function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor. . . . .	231
C.13	This table presents the results from the logistic regression model run in Chapter 4 Section 4.4.3 examining the influence of a participants' promise not to immobilize the robot (post-experiment survey Likert rating of "I promised not to immobilize Echo during the game") on whether participants immobilized the robot in their first power-up choice. We also control for the trust violation framing, trust repair strategy, participant age, and participant gender. We used the R 'glm' function with a binomial family and logit link to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	232

C.14	This table presents the results from the logistic regression model run in Chapter 4 Section 4.4.3 examining the influence of a participants' promise not to immobilize the robot (post-experiment survey Likert rating of "I promised not to immobilize Echo during the game") on whether participants ever immobilized the robot in any of their three power-up choices. We also control for the trust violation framing, trust repair strategy, participant age, and participant gender. We used the R 'glm' function with a binomial family and logit link to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	233
C.15	This table presents the results from the linear regression model run in Chapter 4 Section 4.4.3 examining the influence of a participants' promise not to immobilize the robot (post-experiment survey Likert rating of "I promised not to immobilize Echo during the game") on their Dyadic Trust Scale (DTS) ratings of the robot. We also control for the trust violation framing, trust repair strategy, participant age (considered as a factor in [Strohkorb Sebo et al., 2019], but was not considered a factor for this analysis), and participant gender. We used the R 'lm' function to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.	234
C.16	This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 5, both overall and for each condition. . . . .	235



- C.17 This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s Likert rating (1 - strongly disagree, 7 - strongly agree) on the post-experiment questionnaire item “Echo [the robot] made vulnerable disclosures about Echo’s feelings during the interaction.” Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 236
- C.18 This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s Likert rating (1 - strongly disagree, 7 - strongly agree) on the post-experiment questionnaire item “Echo [the robot] told personal stories during the interaction.” Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 237
- C.19 This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s Likert rating (1 - strongly disagree, 7 - strongly agree) on the post-experiment questionnaire item “Echo [the robot] made use of humor during the interaction.” Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 238

C.20	This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s rating of the robot’s warmth, according to the RoSAS scale [Carpinella et al., 2017]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	239
C.21	This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) and a control for the participants’ gender on the participant’s rating of the robot’s competence, according to the RoSAS scale [Carpinella et al., 2017]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	240
C.22	This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) and a control for the participants’ age on the participant’s rating of the robot’s competence, according to the RoSAS scale [Carpinella et al., 2017]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	241

- C.23 In order to analyze whether or not a human team member who made a mistake looked at the robot afterwards in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (reference group: vulnerable condition) on whether or not a human team member who made a mistake looked at the robot afterwards. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age and gender. We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. 242
- C.24 In order to analyze participants' verbal responses to the robot in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (reference group: vulnerable condition) on whether or not a human team member spoke to the robot after a mistake was made by the team. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' extraversion score and average familiarity with the other participants as well as the mistake round number (1-8). We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 243

C.25	In order to analyze whether or not a human team member who made a mistake explained that mistake to their team members in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (reference group: vulnerable condition) on whether or not a human team member who made a mistake explained that mistake to their team members. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age and average familiarity with the other participants as well as the mistake round number (1-8). We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	244
C.26	In order to analyze whether or not a human team member consoled the one who made a mistake in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition, where the neutral and silent conditions are pooled together, on whether or not a human team member consoled the one who made a mistake. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age as well as the mistake round number (1-8). We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	245

C.27	In order to analyze whether or not a human team member consoled the person who made a mistake (excluding consoling the robot) in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition, where the neutral and silent conditions are pooled together, on whether or not a human team member consoled the person who made a mistake (excluding consoling the robot). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age as well as the mistake round number (1-8). We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	246
C.28	In our first analysis whether or not a human team members laughed together in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (vulnerable or neutral condition) on whether or not a human team member laughed along with another human team member. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age and average familiarity with the other participants. We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	247

- C.29 In order to analyze whether or not a human team members laughed together in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (reference group: vulnerable condition) on whether or not a human team member laughed along with another human team member. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age and average familiarity with the other participants, these controls were scaled to ensure model convergence. We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 248
- C.30 In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on total individual speaking time. We used a multilevel linear model of speaking time (s) as a function of experimental condition (reference group: vulnerable robot) and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity was modeled using random effects clustered in groups. We used the R 'lme' function from the 'nlme' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. 249
- C.31 In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on total individual speaking time. We used a multilevel linear model of speaking time (s) as a function of experimental condition (reference group: vulnerable robot) including an interaction of the treatment effect with round and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in participants in groups. We used the R 'lme' function from the 'nlme' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. 250

C.32 In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on the duration of human team member responses. We used a multilevel linear model of speaking time (s) as a function of experimental condition (reference group: vulnerable robot) including an interaction of the treatment effect with round and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in participants in groups. We used the R ‘lme’ function from the ‘nlme’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 251

C.33 In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on participants equality in talking time ( $E_{TT_i}$ ). We used a multilevel beta regression as a function of experimental condition (reference group: vulnerable robot) and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in groups. We used the R ‘glmmTMB’ function with a beta\_family and logit link from the ‘glmmTMB’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. Because a beta regression cannot analyze 0’s or 1’s (a few participants had values of 1), we transformed the data using the following equation, where  $N$  is the sample size (150) and  $y$  is the outcome variable [Smithson and Verkuilen, 2006]:  $y' = \frac{y*(N-1)+0.5}{N}$ . 252

- C.34 In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on participants equality in talking partners ( $E_{TP_i}$ ). We used a multilevel beta regression as a function of experimental condition (reference group: vulnerable robot) and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in groups. We used the R ‘glmmTMB’ function with a beta\_family and logit link from the ‘glmmTMB’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. Because a beta regression cannot analyze 0’s or 1’s (a few participants had values of 1), we transformed the data using the following equation, where  $N$  is the sample size (144) and  $y$  is the outcome variable [Smithson and Verkuilen, 2006]:  $y' = \frac{y*(N-1)+0.5}{N}$ . 253
- C.35 This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s rating of their team’s psychological safety, according to Edmondson’s Team Psychological Safety scale [Edmondson, 1999]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 254
- C.36 In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on different self-reported group dynamics. We used a multilevel logistic model as a function of experimental condition (reference group: vulnerable robot) and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in groups. We used the R ‘glmer’ function with a binomial family and logit link from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 255



C.37	This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 6 both overall and for each experimental condition. . . . .	256
C.38	This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 6 for each important subdivision of participants (ingroup/outgroup, robot liaison). . . . .	257
C.39	This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the correlation of the participant designations of robot liaison and participant designations of ingroup-outgroup with the average familiarity with their two human team members. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	258
C.40	This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the correlation of the participant designations of robot liaison and participant designations of ingroup-outgroup with their emotional intelligence. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	259
C.41	This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the correlation of the participant designations of robot liaison and participant designations of ingroup-outgroup with their extraversion. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	260

- C.42 This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for emotional intelligence, on the similarity of their survival item rankings from round 1 with their survival item rankings from round 2 (smaller values indicate higher similarity of the lists). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 261
- C.43 This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for participant age and emotional intelligence, on their partner preference score. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 262
- C.44 This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for participant age and the maximum familiarity a participant has between their two other human team mates, on their perceived inclusion scale score. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 263

- C.45 This table presents the results from the 1-way ANOVA analysis comparing the the proportion of survival items initially ranked low (9-25) on the round 1 ingroup list and high (1-8) on the round 1 outgroup list items ( $L_{in}$ ,  $H_{out}$ ) that made it onto the final list of 8 items produced by the entire team at the end of round two of the experiment. We examined this proportion between participants with an ingroup robot liaison verses an outgroup robot liaison outsider, as presented in Chapter 6 Section 6.4. In addition to our main variable of interest, we include in our model the average familiarity of group members with one another and the number of females on the team. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor. . . . . 264
- C.46 This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants’ intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for participant extraversion, on the proportion of time they spent talking 1 minute after robot support targeted to participant(RST-P), robot support targeted to someone else (RST-SE), and a robot undirected utterance (RUU). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 265

C.47	This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for participant extraversion, on the proportion of time they spent talking 1 minute after robot support targeted to participant(RST-P) compared with two controls (via subtraction): robot support targeted to someone else (RST-SE), and a robot undirected utterance (RUU). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	266
C.48	This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with various controls either participants' emotional intelligence or the maximum familiarity a participant had with their two fellow participants, on the participants' ratings of the robot's warmth, competence, and discomfort according to the RoSAS scale [Carpinella et al., 2017]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	267
C.49	This table lists the demographic and descriptive characteristics of all the participants in the human-subjects study detailed in Chapter 7. . . . .	268
C.50	This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 7 in our 2 robot verbal support (yes or not) x 2 intergroup bias robot liaison (ingroup robot liaison vs. outgroup robot liaison) between subjects design. . . . .	269

C.51	This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 7 for each important subdivision of participants (ingroup/outgroup, robot liaison). . . . .	270
C.52	This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of the the backchannels a participant received (sec), with controls for participant gender and emotional intelligence, on their psychological safety score [Edmondson, 1999]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	271
C.53	This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of participants total time talking in round 2 of the experiment (sec), with controls for ingroup-outgroup bias, robot liaison designation, gender, and emotional intelligence, on their psychological safety score [Edmondson, 1999]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	272
C.54	This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of the verbal backchannels a participant received (sec) normalized by the total time that participant spent talking (sec), with controls for ingroup-outgroup bias, robot liaison designation, gender, and emotional intelligence, on their psychological safety score [Edmondson, 1999]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	273

- C.55 This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of the total time a participant received verbal backchannels (sec), with controls for robot liaison designation and emotional intelligence, on their perceived group inclusion score [Jansen et al., 2014]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 274
- C.56 This table presents the results from the linear mixed-effects models run in Chapter 7 Section 7.4 examining the correlation of the total time a participant received nonverbal backchannels (sec) on their psychological safety score [Edmondson, 1999] and their perceived group inclusion score [Jansen et al., 2014]. Controls used in these models include intergroup bias, robot liaison designation, gender, emotional intelligence, and familiarity with other team member. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 275
- C.57 This table presents the results from the linear mixed-effects models run in Chapter 7 Section 7.4 examining the correlation of the total time a participant spent nonverbally backchanneling others (sec) on their psychological safety score [Edmondson, 1999] and perceived inclusion score [Jansen et al., 2014]. Controls used for these models include intergroup bias, robot liaison designation, gender, emotional intelligence, and familiarity with other participants. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. 276

C.58	This table presents the results from the linear mixed-effects models run in Chapter 7 Section 7.4 examining the correlation of the total time a participant spent verbally backchanneling others (sec) on their psychological safety score [Edmondson, 1999] and perceived inclusion score [Jansen et al., 2014]. Controls used for these models include intergroup bias, robot liaison designation, gender, and emotional intelligence. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	277
C.59	This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of the nonverbal and verbal backchannels, separately, a participant received (sec) normalized by the total time that participant spent talking (sec), with controls for ingroup-outgroup bias, robot liaison designation, gender, emotional intelligence, and the maximum familiarity they have between their two human team members, on their perceived inclusion score [Jansen et al., 2014]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	278

C.60	This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of participants total time talking in round 2 of the experiment (sec), with controls for ingroup-outgroup bias, robot liaison designation, emotional intelligence, and the maximum familiarity a participant had between their two fellow human participants, on their perceived inclusion [Jansen et al., 2014]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	279
C.61	This table presents the results from the ANOVA analysis examining the influence of the time participants in a group spent verbally backchanneling one another (sec) on both groups’ average perceived inclusion and average psychological safety scores in Chapter 7 Section 7.4. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor. . . .	280
C.62	This table presents the results from the ANOVA analysis examining the influence of the time participants in a group spent talking (sec) on both groups’ average perceived inclusion and average psychological safety scores in Chapter 7 Section 7.4. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor. . . . .	281
C.63	This table presents the results from the ANOVA analysis examining the influence of the proportion of time participants in a group spent verbally backchanneling one another (sec), relative to the group’s total talking time, on both groups’ average perceived inclusion and average psychological safety scores in Chapter 7 Section 7.4. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor. . . . .	282



C.64	This table presents the results from the ANOVA analysis examining the influence of the time participants in a group spent nonverbally backchanneling one another (sec) on both groups' average perceived inclusion and average psychological safety scores in Chapter 7 Section 7.4. This analysis was performed using the 'aov' function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor. . . . .	283
C.65	This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of intergroup bias (ingroup or outgroup), controlling for the familiarity with other human participants, on the amount of verbal and nonverbal backchannels (sec) each participant received. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.	284
C.66	This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of intergroup bias (ingroup or outgroup), controlling for the familiarity with other human participants, on the proportion of verbal and nonverbal backchannels (sec) each participant received relative to their total talking time (sec). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	285

- C.67 This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of gender on the amount of verbal backchannels each participant received (sec) and the proportion of verbal and backchannels (sec) each participant received relative to their total talking time (sec). We controlled for the total talking time (sec), extraversion, and intergroup bias of participants. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 286
- C.68 This table presents the results from the ANOVA analysis examining the influence of gender on the amount of verbal backchannels produced by each group (sec) as well as the proportion of verbal backchannels produced by each group (sec) with respect to their total talking time in Chapter 7 Section 7.4. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor. . . . . 287
- C.69 This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of having a verbally supportive robot on the amount of verbal backchannels (sec) each participant received. The model’s fixed factors included whether the robot gave verbal support, the intergroup bias (ingroup, outgroup), robot liaison designation, and relevant interactions. We controlled for participants’ extraversion and familiarity with other team members. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis and the ‘emmeans’ function with a Tukey adjustment from the ‘emmeans’ package to perform post-hoc tests. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . . 288

C.70	This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of having a verbally supportive robot on participants' psychological safety and perceived inclusion scores. The model's fixed factors included whether the robot gave verbal support, the intergroup bias (ingroup, outgroup), robot liaison designation, and relevant interactions. We controlled for participants' extraversion and familiarity with other team members. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. . . . .	289
------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----

# Acknowledgements

I am incredibly grateful for all of the support I have received in pursuit of my Ph.D. and in the writing of this dissertation.

First and foremost, to my Ph.D. advisor Brian Scassellati (Scaz), thank you for your mentorship, guidance, and support over these six years. Thank you for empowering me to develop my own research direction by giving me the freedom to explore, try new ideas, and make mistakes. The way that you have mentored me and led Scazlab has inspired my dream to become a professor and life-long mentor.

To the other members of my dissertation committee (Malte Jung, Nicholas Christakis, and Marynel Vázquez), thank you for your thoughtful advice, your feedback on my ideas, and your support of me as a researcher. Malte, thank you for sharing with me your infectious excitement about shaping social group dynamics, for all I have learned from you about experimental design, and for exposing me to the literature that has inspired my dissertation work. You have been like a second advisor to me, and I hope that we continue collaborating for many years to come. Nicholas, thank you for showing me how to rigorously and thoroughly pursue my ideas and, most of all, for your belief in me and my research. Every time I met with you, I left greatly encouraged that my research mattered and that my ideas were worth pursuing. Marynel, thank you for your valuable input on my most recent work and for your career help and advice. Watching you start your lab at Yale has provided me with an excellent example of what I hope to achieve as I start my own lab in the fall.

To my graduate student and postdoc colleagues in the Social Robotics Lab, thank you for your collaborative efforts, your feedback and input on my research, as well as your advice and guidance. Working alongside you all has made this journey of a PhD incredi-

bly valuable and enjoyable: Henny Admoni, Bradley Hayes, Iolanda Leite, André Pereira, Aditi Ramachandran, Alex Litoiu, Corina Grigore, Kate Tsui, Laura Boccanfuso, Margaret Traeger, Alessandro Roncone, Olivier Mangin, Chien-Ming Huang, Nicole Salomons, Jake Brawer, Meiying Qin, Timothy Adamson, Emmanuel Adeniran, Rebecca Ramnauth, Nicholas Georgiou, and Debasmita Ghose. Aditi and Nicole, thank you for being my longest and closest companions in the lab; your input has greatly improved the quality of my work and your friendship has helped me to keep going on the hardest of days. Chien-Ming, Alessandro, Iolanda, and Henny, thank you for your invaluable ‘few steps ahead’ advice and guidance; your support and encouragement has given me the confidence to set my sights beyond my time at Yale higher than I could have thought possible.

To all the undergraduate students and high school students I have been able to work with in the Social Robotics Lab, thank you for your hard work, your valuable contributions to this work, and most of all for helping to grow my dream to become a lifelong educator. I would especially like to mention: Natalie Warren, Isabelle Gallagher, Adam Erickson, Rachel Ha, Priyanka Krishnamurthi, Evelyn Roberts, Ling Dong, Nicholas Chang, Michael Schutzman, Sean Hackett, Tom Wallenstein, Michal Lewkowicz, and Hannah Burgess.

To my friends in the New Haven area and beyond, thank you for your invaluable encouragement and companionship. To my Vox Church community, thank you for being a second family to me throughout my six years in New Haven. It has been a great privilege to pursue our Lord Jesus Christ together. And to my Crossfit New Haven community, thank you for your friendship and for pushing me to a level of fitness and strength I would have never thought possible.

Finally, to my family, thank you for your unconditional love and consistent support. To my parents, thank you for being my two biggest fans, for encouraging my pursuit of my dreams, and for being a refuge for me during the storms of life. And finally to Zach, thank you for being my rock, for a consistent source of fun and laughter, and for believing in me even when I did not believe in myself.

# Chapter 1

## Introduction

The field of human-robot interaction (HRI) has made significant advances in understanding how to build robots that can seamlessly interact with people. Importantly, work in HRI has focused on the ways in which robots can uniquely leverage their physical embodiment to communicate intuitively with people using social cues and interaction methods that are familiar to people. Furthermore, socially assistive robotics (SAR) [Feil-Seifer and Mataric, 2005] has explored how robots can provide value and assistance to people through a robot’s *social* actions, as opposed to the more traditional approach of robots providing help to people exclusively through physical tasks. Social robots have shown promise in their assistance to people in a variety of applications: providing tutoring help to children [Ramachandran et al., 2016], guiding people to locations in a shopping mall [Shiomi et al., 2010], and enhancing social engagement in elder care facilities [Šabanović et al., 2013].

The significant advances in understanding social interactions between humans and robots in HRI have predominately been evaluated and supported through experimental studies conducted in the laboratory examining how one human interacts with one robot. Although these one-on-one studies allow researchers to isolate and investigate particular components of human-robot interactions, they do not enable researchers to study human-robot interactions as they will likely occur in the real world. In natural settings where robots are and will be deployed, robots often interact with groups of people and are often members of human-robot collaborative teams.

Interactions with groups and teams, as opposed to individuals, are influenced by complex

social dynamics. Most social dynamics are emergent properties, which cannot be explained simply through examining all of the dyadic interactions between people on the team. Social group dynamics have also been shown to have significant influence on both team member satisfaction and team performance [Cho and Mor Barak, 2008, Edmondson, 1999, Jones and George, 1998, Mayer et al., 1995, Sabharwal, 2014, Shore et al., 2011, Woolley et al., 2010]. Surprisingly, social dynamics have even been shown to be more predictive of group success than individual intelligence metrics. Woolley et al. (2010) found evidence for a “collective intelligence factor” that predicted a group’s success on a variety of tasks. This collective intelligence factor was not found to be correlated with the individual intelligence of its groups members, however, was found to be correlated with the social sensitivity of group members, the equality in talking turns, and the number of females on the team [Woolley et al., 2010]. If it is true that social group dynamics, like social sensitivity, are so critical for team success, then it is essential that robots interacting with groups positively contribute to its social dynamics.

This dissertation focuses on building robots that positively shape social dynamics in collaborative human-robot teams, comprised of one robot and several people. In order to build robots that can best support team success, we investigate novel methods for a robot to shape three critical collaborative team social dynamics - trust, inclusion, and psychological safety - validated through human-subjects studies. We build upon prior work in HRI that has demonstrated the influence of a robot’s social actions in one-on-one human-robot interactions, to show the unique effects of a robot’s behavior on members of collaborative teams. After investigating the influence of specific robot actions on important team social dynamics, we seek to perceive and model the high-level team dynamics of inclusion and psychological safety from a low-level human behavior - verbal backchannels. We statistically demonstrate the correlation between verbal backchannels and these team social dynamics and, through a human-subjects study, investigate the influence of a robot’s verbal backchannels on human verbal backchannels. This dissertation leverages concepts and approaches from computer science, organizational psychology, and sociology to build robots that enhance team dynamics and performance within human-robot teams.

This dissertation begins with a comprehensive review of research investigating human-

robot interaction involving robots interacting with groups of people as well as research identifying the social dynamics that lead to team success in human teams (Chapter 2). This background chapter presents key findings from organizational psychology and related fields that point to the importance of collaborative team social dynamics such as trust, inclusion, and psychological safety on team member satisfaction and team performance. These insights from human teams on the social factors that lead to team success inform the robot behavior we investigate in order to positively shape team dynamics and performance. Additionally, we examine related work investigating robots that interact with groups and teams of people. It presents research that has demonstrated the important differences between robot interactions with an individual person and robot interactions with groups of people, the influence of a robot’s verbal and nonverbal behavior, the importance of interaction context in determining the robots role in the group interaction, and the ability of robots to additionally shape human-to-human interactions within the group. We conclude this chapter by highlighting the limited body of work in HRI focused on the specific social dynamics trust, inclusion, and psychological safety, framing the importance and impact of the work described in this dissertation.

In the following chapters (Chapters 3 - 7), we describe human-subjects experiments that broaden our knowledge of the social influence of robots in collaborative human-robot teams. These experiments follow a similar approach of identifying a social team dynamic that has been established to be critical to human team success, designing a robot behavior to enhance that team social dynamics, and then evaluating the effectiveness of the robot’s actions to positively shape that social team dynamic. Put together, these studies investigate how several different robot behaviors can positively or negatively shape the critical collaborative team dynamics of trust, inclusion, and psychological safety.

Our work investigating how robots can shape team dynamics begins in Chapter 3 by investigating ways of broadly enhancing collaboration between two children in a shared task. We decided to focus on children (ages 6 to 9) for this study because children around this age are actively developing collaborative skills. While working on a collaborative task, a robot either asks the children relationship-focused questions, task-focused questions, or no questions at all. These questions influenced the children’s perceptions of their performance



and their actual performance in the task. Although the rest of the work in this dissertation is focused on adults (we found children’s attention spans to be very short and their behavior more varied), this study highlights the main focus of this work: the actions of a robot shape how people in the group interact with each other.

Chapter 4 examines trust, a critically important social dynamic for human-robot collaborative teams. Specifically, we focus on how a robot can best recover from a trust violation - where a robot breaks a person’s trust. Since robot trust repair has not been thoroughly studied in a dyadic context, we investigated the influence of a robot’s trust violation framing and trust repair strategy in a human-subjects study with one person and one robot. Our findings suggest that the best way a robot can repair trust is by apologizing for having made a mistake.

After exploring robot trust repair in a dyadic human-robot interaction, we sought to explore how a robot’s vulnerable verbal expressions (e.g., apologizing for a mistake) influence trust-related behavior when a robot interacts with a group of people. Chapter 5 describes a human-subjects experiment where three people and a robot play a collaborative game, where each person will make mistakes, causing the team to suffer as a result. Throughout the game, the robot either makes vulnerable utterances, neutral utterances, or is silent. We found that the robot’s vulnerable utterances in particular lead to more social interaction with the robot, more trust-related behavior expressed to fellow human participants, and more conversation between the human participants. These results further provide evidence that in regards to critical team social dynamics, a robot’s actions can positive influence human-to-human interactions within the group.

In Chapter 6, we turn our attention toward perceived inclusion, a contributing factor to both team member commitment and team performance [Cho and Mor Barak, 2008, Sabharwal, 2014, Shore et al., 2011]. We investigate two strategies that we believed to improve the inclusion of human members within a human-robot team. We found that our first strategy, giving a team member a specialized role to interact with the robot, had a negative influence on human team member inclusion, however our second strategy, having the robot give targeted supportive utterances to human team members, showed promise in increasing the verbal contribution from ‘outsider’ team members. This work further

reinforces the main theme of this dissertation, that robot actions in a group can influence human team members interactions with one another, and also contributes evidence that some robot behaviors designed to enhance team dynamics may have the opposite effect.

After establishing robots' ability to shape team social dynamics and human-to-human interaction in collaborative teams of humans and robots, we explore ways in which we can enable robots to sense and react to social dynamics in real-time. Chapter 7 describes our identification of a low-level human behavior that robots can sense (verbal backchannels) and provides evidence that these verbal backchannels are positively correlated with our third important social dynamic - psychological safety. Further, from data gathered in a human-subjects experiment, we provide evidence that robot verbal backchannels can influence human verbal backchanneling. This work both provides researchers with a way to sense social group dynamics real-time as well as further evidence that robot behavior influences how people interact with each other in human-robot group settings.

In Chapter 8 we discuss the work presented in this dissertation. We focus on the significant contributions and broader implications of this work as well as its limitations and future directions. We summarize the work presented in this dissertation in Chapter 9.

This dissertation makes the following novel contributions to the understanding of a robot's influence in human-robot collaborative teams: (1) evidence that a robot's actions can influence how people in a group interact with each other, (2) the identification of behaviors that a robot can use in order to positively shape team dynamics essential to team success, and (3) the connection between human backchanneling behavior and team social dynamics, enabling a robot to sense social dynamics of a collaborative team in real-time and select its actions accordingly.

## Chapter 2

# A Review of Social Dynamics in Collaborative Human Teams and Human-Robot Teams\*

This chapter provides a comprehensive overview of the social dynamics that are critically important for team success, with a particular focus on how robots have been shown to influence these social dynamics. We first examine research that investigates collaborative human teams, specifically examining three social dynamics proven to be essential to successful collaboration (trust, inclusion, and psychological safety) and the influence each social dynamic specifically has on the team. Next, we turn our attention to robots interacting with groups and teams, focusing on (1) the unique differences between robot interactions with groups as opposed to individuals, (2) and what the field of human-robot interaction (HRI) has contributed to our knowledge of how robots can best be built to interact with groups of people. Finally, we review the work in HRI that has examined robot influence on the collaborative social dynamics of trust, inclusion, and psychological safety.

---

\*Portions of this chapter were originally published as: S. Strohkorb Sebo, B. Stoll, B. Scassellati and M. Jung. Robots in Groups and Teams: A Literature Review. [currently under review]

## 2.1 Human Collaborative Teams

Before examining the prior work exploring collaborative teams of humans and robots, we first examine the literature on human teams to understand how members of successful collaborative teams work together to achieve positive performance outcomes. We specifically focus on three social dynamics (trust, inclusion, and psychological safety) that have been proven to be essential for team success. We both describe these social dynamics in the collaborative team context and demonstrate the positive influence of these social dynamics on team success.

### 2.1.1 Human Teams

There exists a rich literature in organizational psychology examining many facets of collaborative teamwork. The teams studied in this literature are mostly corporate teams, where it is normal for companies to form and re-form teams based on the changing needs of organizations where a team’s life span may be several weeks, months, or years. From this body of work, we discuss three prominent threads of research: the influence of time on a team’s process, the development and realization of group norms, and the sharing of knowledge within teams.

Exploring the entire life cycle of a team is critical for understanding how a successful end result is produced from the collection of distinct individuals. Team development, from the point of formation to work completion, has been described by Tuckman (1965) as having four stages: forming (orientation to the group), storming (conflict and polarization), norming (overcoming conflict and increasing cohesion), and performing (group energy becomes task focused) [Tuckman, 1965]. Conversely, Gersick (1988) does not believe that every team goes through a linear development process, and rather describes a team’s development as unfolding through a ‘punctuated equilibrium,’ undergoing a critical midpoint transition where the team’s approach is significantly adjusted based on the initial progress of the team [Gersick, 1988]. Ericksen and Dyer (2004) have found that high performing teams mobilize more quickly than low performing teams, providing the team with more time to complete the work they agree upon to do [Ericksen and Dyer, 2004]. Exploratory search,

“the intentional pursuit of alternative approaches to team tasks,” has been shown to be extremely beneficial to teams, but only in the first half of a team’s life span [Knight, 2015]. Lastly, groups that have formed under time pressure have been shown to produce lower quality outcomes [Kelly and McGrath, 1985], however, other work has found that task focus may benefit from time limits since time limits have shown an inverse relationship to task focus [Karau and Kelly, 1992]. Although based on this work examining the temporal considerations of teamwork we can conclude that how a team spends its time influences their success, there may not be one clearly defined one-size-fits-all temporal process that best suits all collaborative teams.

In addition how to differences in how teams act over time influence their success, social influences such as group norms, social pressure, and social impact can greatly shape team functioning and performance. When a team first forms, group norms, “the informal rules that groups adopt to regulate and regularize group members’ behavior” [Feldman, 1984], are established which govern the behavior of group members. The first meeting of a team is critical in establishing these group norms, as well as any pivoting points where the team adjusts its direction [Gersick, 1988]. Cialdini et al. (1990) define two types of norms - injunctive norms, which are norms based on the approval or disapproval of other people, and descriptive norms, which are norms based on the observation of the actions of other people. These norms, combined with the conditions in which people focus their attention toward or away from a norm, determine how an individual will act [Cialdini et al., 1990]. Similarly to norms, social pressure can be exerted by group members on individuals and influence their behavior. Asch’s well known conformity studies have displayed inclination of an individual to conform their behavior to match the behavior of others, even when the majority is acting in a way that the individual knows is incorrect [Asch, 1956]. Latané (1981) has further expanded on the idea of conformity and social pressure through their examination of social impact, “the changes in physiological states and subjective feelings, motives, emotions, cognitions, beliefs, values, and behavior that occur as a result of the real, implied, or imagined presence of actions of other individuals.” They have demonstrated that social impact’s effect can be multiplied by strength, power, and immediacy of the other people causing the effect and that social impact can be diminished when other people

stand with the target [Latané, 1981]. Additionally, Bettenhausen and Murnighan (1985) have found that individuals use their past experiences to inform their choices in similar future group situations and then revise their beliefs about which action to choose in order to align themselves with the group as they continue to interact with their team members [Bettenhausen and Murnighan, 1985]. Group norms, social influence, and social impact all highly influence the behavior of individuals within groups and must be carefully established to give the team its best chance at success.

Similarly to how group norms powerfully shape individual behavior, how teams learn and share knowledge strongly determines team success. The success of large organizations and teams is driven by the actions of each individual member and how that member interacts and aligns himself or herself with their team mates [Orlikowski, 2002]. An individual may influence how the group responds, given that the structures and patterns in place support the sharing of the individual’s knowledge and opinions with others [Stasser, 1999]. One model used for conceptualizing how knowledge is managed and transferred between individuals within a group is transactive memory systems. Transactive memory can be defined as “the shared division of cognitive labor with respect to the encoding, storage, retrieval, and communication of information from different knowledge domains, which often develops in groups and can lead to greater efficiency and effectiveness” [Brandon and Hollingshead, 2004]. Transactive memory occurs in groups when each member of the group is aware of the specialized skills of the other group members, and knows when to delegate tasks to a member with a specialty in that task. Transactive memory systems (TMSs) are considered optimal when each member has an accurate, shared representation of the TMS and when all team members are participating fully [Brandon and Hollingshead, 2004]. TMSs have also demonstrated a positive influence on group learning and learning transfer - applying knowledge learned from a previous task to a new task [Lewis et al., 2005]. Additionally, a team’s ability to effectively gain and share knowledge (also called ‘team learning’) significantly predicts the success of the team [Edmondson, 1999, Zellmer-Bruhn and Gibson, 2006]. Team learning within multinational organizations has been shown to be supported by an emphasis on responsiveness and knowledge management, as opposed to an emphasis on global integration [Zellmer-Bruhn and Gibson, 2006]. Team learning can also be influenced

by the subgroups that exist within an organization, where team learning is optimized when similarities exist within subgroups and diversity is present between subgroups [Gibson and Vermeulen, 2003]. This allows for knowledge and ideas to surface within subgroups and diverse ideas to be proposed across subgroups [Gibson and Vermeulen, 2003]. The ability for a team to engage in effective team learning is critical to their success [Edmondson, 1999] and within teams and organizations, it is important to consider how systems and norms can be designed to best facilitate team learning.

Keeping in mind the importance of a team’s temporal lifespan, norms, and learning behavior; we look now to key social dynamics that serve as the foundation for these critical components to collaborative teaming. The social dynamics we focus on in the next three sections are trust, inclusion, and psychological safety. We both describe these social dynamics in the collaborative team context and demonstrate the positive influence of these social dynamics on team success.

### **2.1.2 Trust**

Trust is a necessary ingredient for successful cooperation and teamwork [Jones and George, 1998, Mayer et al., 1995]. 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) [Mayer et al., 1995].” Thus, trusting others when working together involves the willingness to take risks by making oneself vulnerable to the responses of others.

Trust in groups and teams of can be conceptualized 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 [Edmondson et al., 2004]. A lack of trust within a team has been found to impair learning [Edmondson, 1999], to decrease people’s willingness to work as part of a team [Kiffin-Petersen and Cordery, 2003], and in some cases even impair a team’s chances at survival [Weick, 1993]. Conversely, an increase in trust within a team has been shown to facilitate problem solving [Klimoski and Karol, 1976, Zand, 1972], functional conflict

resolution [Simons and Peterson, 2000], and overall team performance [Edmondson, 1999].

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” [Wheless, 1978] and which puts the person at interpersonal risk. Prior work in HRI has established a relationship between expressions of vulnerability and trust towards the vulnerable party [Wheless, 1978]. This may seem surprising, since vulnerability may evoke negative emotions. However, when considered from a social functional perspective in collaborative settings [van Kleef, 2016], vulnerability has positive social consequences as it orients people toward each other and facilitates social engagement [Van Kleef et al., 2010].

In addition to considering ways to enhance trust in teams, it is also essential to consider how trust can be repaired when someone in the group violates trust. There are many different repair strategies people use to repair trust including making an apology, denying culpability, promising better behavior in the future, and making excuses [Kim et al., 2009]. To make an effective trust repair, the framing of the trust violation also must be considered. Prior work has demonstrated that trust repair strategies have different effects when the trust violation is due to either a lack of competence (e.g., an accountant failing to properly file taxes because of inadequate knowledge about a relevant tax code) or a lack of integrity (e.g., an accountant failing to properly file taxes intentionally) [Kim et al., 2004]. Thus, after a trust violation, it is important to consider both the trust repair strategy (e.g., apology, denial) and trust violation framing (e.g., competence, integrity) when trying to repair trust with other people.

### **2.1.3 Inclusion**

Inclusion is increasingly being recognized as an essential component to productive and successful groups and teams [Oswick and Noon, 2014]. Inclusive work teams are comprised of members with diverse perspectives and skills, who are well-trained and given the opportunity to contribute equally in a group [Miller, 1998]. Work teams that have an inclusive environment have been demonstrated to produce committed team members and better-performing teams [Cho and Mor Barak, 2008, Sabharwal, 2014, Shore et al., 2011].

Inclusion within human groups has been specifically explored in human teams as it



relates to the formation of subgroups, otherwise known as intergroup biases. Tajfel (1982) describes group identification in terms of two necessary components: a cognitive awareness of membership in that group and an evaluation of the value of that membership [Tajfel, 1982]. This awareness of intergroup membership results in an “us vs. them,” or ingroup vs. outgroup, mentality, resulting in behaviors that further reinforce discriminatory ingroup and outgroup relationships [Baron and Dunham, 2015]. Dunham et al. (2011) shows that “mere membership” in randomly selected groups result in ingroup-favoring and outgroup-opposing behaviors in people even as young as 5 years old [Dunham et al., 2011].

One important contributing factor to the formation of intergroup biases are faultlines. Faultlines are defined as divisions of a group based on one or more attributes, with the strength of the faultline increasing with the number of attributes that align in the same way [Lau and Murnighan, 1998]. The strength or activation of a faultline is also determined by the context of the issue or task relevant to the group. For example, a group dealing with conflicts surrounding retirement and pensions is more likely to have a stronger faultline along the attribute of age as opposed to a much less relevant attribute such as gender. As faultlines become activated and strengthened in a group, members will become more biased towards other ingroup members, implying that faultlines, once activated, tend to reinforce themselves but in the absence of activation may also naturally weaken over time [Lau and Murnighan, 1998].

#### **2.1.4 Psychological Safety**

Psychological safety is a term coined by Amy Edmondson and is defined as a “shared belief held by members of a team that the team is safe for interpersonal risk taking” [Edmondson, 1999]. Edmondson (1999) demonstrated that psychological safety does positively influence team performance, and that the relationship between the two is moderated by the team’s learning behavior (e.g., asking for help, seeking feedback, and discussing errors). Psychological safety has also been shown to positively correlate with leader inclusiveness [Nembhard and Edmondson, 2006], team member engagement in quality improvement efforts [Nembhard and Edmondson, 2006], high-quality relationships [Carmeli et al., 2009], a more positive attitude about teamwork [Ulloa and Adams, 2004], and both exploratory and ex-

exploitative learning [Kostopoulos and Bozionelos, 2011]. Psychological safety has demonstrated a moderating effect between task conflict and team performance, where teams that are psychologically safe can achieve greater team success through as a result of productive task conflict [Kostopoulos and Bozionelos, 2011]. Additionally, a comprehensive survey at Google, involving over 200 interviews and examining hundreds of attributes of more than 180 Google teams, concluded that psychological safety was the most influential factor in Google team success [Rozovsky, 2015].

Despite the evidence that psychological safety is an essential factor in driving team success, no research to our knowledge has demonstrated an effective intervention strategy to strengthen a team’s psychological safety. If anything, the opposite has been true. For example, a study conducted at a U.S. university of 55 student project teams (249 students total) examined the influence of team-focused activities (aimed to promote psychological safety) verses task-focused activities did not find any evidence that psychological safety differed based on the type of activity administered [Hastings et al., 2018]. Edmondson speaks of three ways to build psychological safety: 1) frame the work problem as a learning problem not an execution problem, 2) acknowledge your own fallibility, and 3) model curiosity [Edmondson, 2014]. These also currently serve as the source for Google’s guidelines for its employees to build psychological safety [re:Work, 2020]. Although these three strategies to build psychological safety have theoretical support in the literature [Edmondson, 1999], no experimental study has yet to provide evidence for an intervention (e.g., team-building activities) that has been proven to improve psychological safety in teams. It is likely that a more involved and thorough intervention is needed to influence psychological safety in collaborative teams.

## 2.2 Human-Robot Collaborative Teams

After considering the factors that drive success in human teams, we turn our attention to the addition of robots within human groups and teams. We first present an in-depth review of the literature of robots interacting with groups and teams of people. We discuss the unique roles robots can play in groups, finding that small changes in their behavior and personality

impacts group behavior and, by extension, influences ongoing interpersonal interactions. We then specifically examine work related to robots supporting collaboration in human-robot teams where we examine robot influence on trust, inclusion, and psychological safety in particular.

### 2.2.1 Robots in Groups and Teams: A Literature Review

Research on HRI has begun to depart from a primarily interpersonal, one-to-one, level to increasingly consider HRI at the group level (e.g., groups, teams, workplaces, organizations, families, classrooms, etc. [Hinds et al., 2004]). While the field and its varying contributors have built a strong understanding of basic social-psychological responses to robots in dyadic contexts (i.e., one robot interaction with one human [Lee et al., 2006, Fussell et al., 2008, Goodrich and Schultz, 2007, Hancock et al., 2011, Shah et al., 2011, Talamadupula et al., 2010]) our overall understanding of what happens when robots interact with groups of people is highly limited.

Transitioning from dyadic interactions to interactions with groups and teams (see Figure 2.1<sup>†</sup>) constitutes a fundamental change in complexity that current HRI theory does not capture [Groom and Nass, 2007, Jung and Hinds, 2018], nor do existing approaches readily scale to groups [Clabaugh and Matarić, 2019, Matsusaka et al., 2001, Matsuyama et al., 2015]. We know from research examining groups of people that the addition of new group members increases complexity by forcing members to consider new interpersonal dynamics, organizational level factors, and group processes for successful interaction (e.g., conflict management, establishment of group norms, and maintaining shared mental models) [Cohen and Bailey, 1997, Levine and Moreland, 1998, You and Robert, 2018, You and Robert, 2017]. Robots joining groups of people also face these enumerated complexities [Jung, 2017], and as additions to a group, they can dramatically impact its overall dynamic [Correia et al., 2018b, Strohkorb Sebo et al., 2018].

This review on robots interacting with groups and teams of people seeks to answer three primary research questions: How does a robot’s behavior shape group dynamics and

---

<sup>†</sup>For the images in Figure 2.1 representing work that is not our own, (a) - (c), we obtained permission from their authors to include them in this dissertation.



Figure 2.1: Descriptive examples of robots interacting with human groups: (a) a Furhat robot completing a sorting task with two people in a museum [Skantze et al., 2015], (b) two EMYS robots and two people playing a card game in a lab setting [Correia et al., 2018b], (c) a Robovie robot guiding people in a shopping mall [Shiomi et al., 2010], and (d) our own work exhibiting a Nao robot playing a collaborative game with three people in a lab setting [Strohkorb Sebo et al., 2018].

people’s behavior within the group? What are appropriate roles for robots to adopt in a variety of settings? And how does a robot’s behavior affect how people in the group behave towards one another?

### **Differences between Robots Interacting with Groups and with Individuals**

Studies of a robot receptionist were among the first to note that groups of people interact with robots differently when compared with individuals. Groups, as opposed to individuals, were significantly more likely to interact with the robot receptionist [Gockley et al., 2006, Michalowski et al., 2006]. Group interactions with the robot receptionist also lasted longer, however, groups spent less time interacting directly with the robot. Similarly, there is evidence that groups of children interacting with educational robots display both distinct behavior and different learning outcomes, when compared with individual children. Groups of children direct less attention to robots because they are also attending to one another. As a result, learning and recall for children in groups of three were shown to be worse than

individual children when listening to two robots playing out interactive narratives [Leite et al., 2015a, Leite et al., 2017]. Both of these robot applications highlight some new challenges for robots in groups, such as managing turn taking, capturing attention, and identifying subgroup formation.

People are also more likely to comply with a robot’s requests if they are in a group as opposed to being alone. Unsuspecting university students were three times more likely to allow a robot to both enter and exit their restricted access dormitory building if they were in a group rather than if they were by themselves [Booth et al., 2017]. It is possible that group members feel reassured by the presence of other group members and may be more likely to comply with a robot’s request that they may not fully trust. As robots seek to engage with both individuals and groups in public settings, the visibility of the robot’s intentions as well as the robot’s ability to discern and reason about the differences in influencing the behavior of groups and individuals will be essential.

Additionally, the mathematical models that robots use to make inferences about people they interact with do not always extend well beyond one-on-one interactions. The accuracy of machine learning classifiers designed to recognize disengagement in children significantly decreased in group contexts when the classifiers were trained on data with individuals [Leite et al., 2015b]. However, classifiers trained on videos of children within groups of three predicted engagement more accurately [Leite et al., 2015b]. These findings illustrate that the models of human behavior that robots employ must take group context into account.

Finally there is evidence that groups exhibit more competitive and aggressive behavior toward robots than individuals. For example, when pairs rather than individuals played a game against a robot, they exhibited more competitive and less cooperative behavior towards the robot [Chang et al., 2012]. Similarly, groups of three humans exhibited more greed and competitive behaviors toward robots than individuals [Fraune et al., 2019]. Especially children and young adults have shown a tendency to exhibit bullying behaviors toward a robot in public spaces [Bohus et al., 2014, Brscić et al., 2015, Salvini et al., 2010].

In sum, this literature provides compelling evidence that people interact differently with robots when they are alone than when they are with other people. This section seeks to highlight the research unique to robots interacting with groups of people: the nonverbal and

verbal behaviors a robot can use, real-world settings where human-robot group interaction has been explored, and the influence of robot behavior on how people in the group interact with one another.

## **Review Method and Corpus**

We conducted a systematic review of the experiment designs, methodologies, and analytic techniques that form the foundation of research studies investigating robots interacting with groups and teams. Our review takes stock of existing work and highlights areas of opportunity for future research. We included studies that satisfied the following criteria:

1. The study must include at least one physically embodied robot.
2. At least two locally present people must interact with the robot(s) simultaneously.
3. The robot(s) must be autonomous or perceived to be autonomous interactant(s).
4. The study must explore group-level phenomena and provide a direct contribution to our understanding of how robots interact with and influence groups of people.

These criteria were chosen to focus this review on studies that investigate how physically present autonomous robots shape interactions with multiple people simultaneously. We do not include studies pertaining to mobile robotic presence systems, or telepresence robots [Neustaedter et al., 2016, Neustaedter et al., 2018, Stoll et al., 2018, Takayama and Go, 2012], because they are dependent upon a human in the loop and lack the autonomy necessary to be considered agentic robots. Similarly, this excludes numerous studies conducted using remote-controlled robotics such as rovers [Stubbs et al., 2007, Vertesi, 2012], rescue robots and drones [Murphy, 2004], and surgical robots [Beane and J. Orlikowski, 2015, Cooper et al., 2013, Duysburgh et al., 2014, Pelikan et al., 2018], since these robots fulfill more the role of robotic tool rather than autonomous agents. Excluded as well are technical papers wherein the primary focus is on systems designed for multiple-human interaction, however, do not demonstrate the influence of the robot’s actions on the group. For example, this applies to papers focused on analyzing multi-person groups for approach strategies [Althaus et al.,

2004, Fan et al., 2016, Tasaki et al., 2004], localizing and parsing relevant group member speech, and analyzing group cues for topic shift and engagement [Matsusaka et al., 2001].

Using the inclusion criteria described above, we conducted an exhaustive search of papers within top-tier HRI outlets including the ACM/IEEE International Conference on Human-Robot Interaction (HRI), the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), the International Journal of Social Robotics, and ACM Transactions on Human-Robotic Interaction (THRI). Additionally, we conducted a Google Scholar search using the terms “robots in groups” and “robots in teams” as well as for the publications of all authors who participated in the Robots in Groups Workshop event hosted at the 2017 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW). We conducted further Google Scholar searches of all authors attached to our existing literature collection as well as relevant author citations within these sources. We set a cut-off date of publication for inclusion in this review at April, 2019.

In total, we collected a corpus of 102 peer-reviewed scholarly papers for our review, which contain 100 distinct studies - human-subjects experiments with a defined experimental design (the papers, their studies, and their characteristics can be found in the supplemental materials as well as in [Strohkorb Sebo et al., 2020b]). Some papers included in this review contain multiple studies and some papers refer to the same study. To provide a descriptive account of the corpus of studies, we extracted several features: the type of robot(s) used, whether or not the robot(s) have a head and eyes, the country where the study was conducted, the setting of the study, the robot control methodology (Wizard-of-Oz or autonomous), the role of the robot (leader, peer, or follower), the composition of the group (the number of people and the number of robots), the entitativity of the group of people according to [Lickel et al., 2000] (loose, task, intimacy), the analysis method used (observation-based or quantitative), the study design, the number of between subjects conditions, the number of groups, the number of total participants, the number of study sessions, and a list of all the statistically significant reported results.

By examining the publication year of the studies included in this corpus, it is clear that there is growing interest in research investigating robot interactions with groups of people, see Figure 2.2. This growth of studies is likely motivated by increasing availability of robust

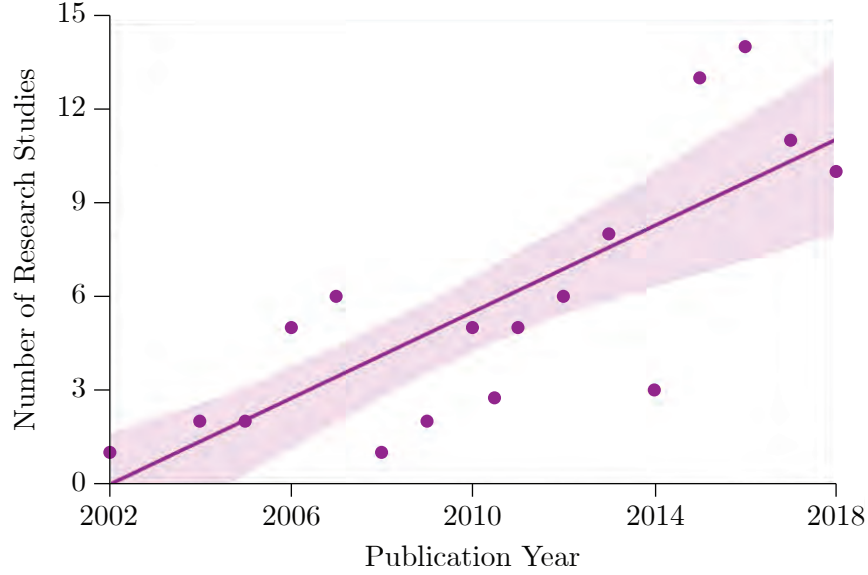


Figure 2.2: The number of studies investigating robots interacting with multiple people has steadily grown over the past couple of decades. The shaded area around the line of best fit represents a 95% confidence interval.

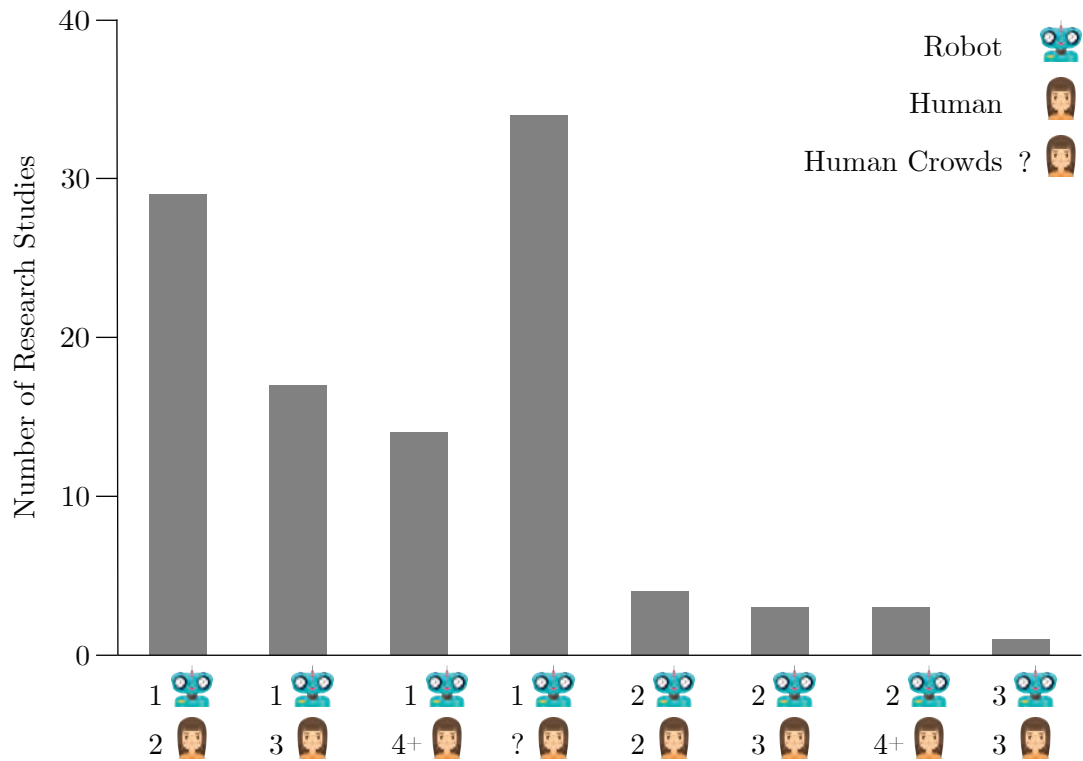
robot platforms and studies conducted in real-world environments as opposed to controlled lab studies.

Of the studies in our corpus, a majority have examined groups consisting of a variable number of people and one robot (e.g., a robot approaching and interacting with groups of varying sizes in a shopping mall), two people and one robot, and three people and one robot, see Figure 2.3(a). Few studies have explored groups consisting of more than one robot interacting with a group of people.

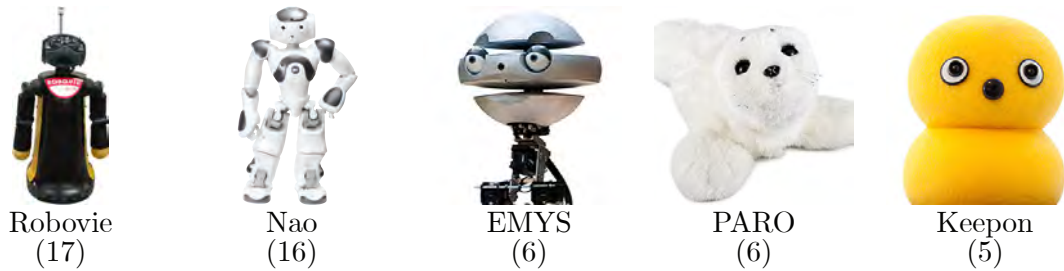
A variety of robots have been studied interacting with groups of people: android robots (e.g., Android Repliee Q2), humanoid robots (e.g., Robovie, Nao), animal robots (e.g., PARO), simple social robots (e.g., Keepon), robots that have a face and no body (e.g., EMYS), and robots that do not have a social appearance (e.g., Turtlebot, Roomba). Figure 2.3(b) displays the most commonly used robots in the studies included in this review. It is important and interesting to note that the robots most commonly used in this body of work have two features in common: a head and eyes. In fact, 83% of the studies in this review study a robot that has a head and eyes. The inclusion of both a head and eyes in the majority of robots used in studies with groups may speak to the importance of a robot’s ability to direct attention in a group and leverage accessible and familiar social cues,



(a) Group Composition



(b) Most Common Robots Used in the Studies



(c) Robot Control Methods



Figure 2.3: We display the (a) composition of human-robot groups studied in the literature as well as (b) the most commonly used robots in these studies and the (c) control methods for these robots.

establishing it as a social agent in the context of a human-robot group. However, this could also be influenced by the large percentage of commercially available robot platforms that have a face and eyes, so the importance of these features should be considered keeping this

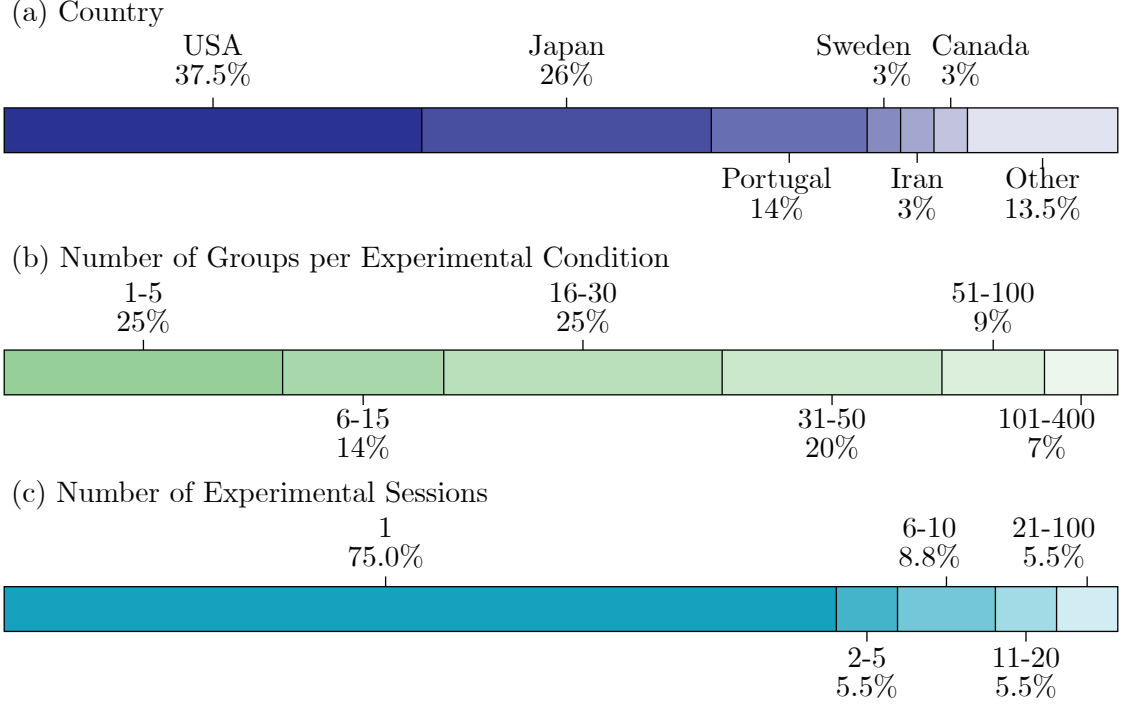


Figure 2.4: In the studies we review on robots interacting with human groups and teams, we highlight the (a) countries where the studies were run, (b) the number of groups per between-subjects condition in the experimental studies, and (c) the number of interaction sessions in the experimental studies.

in mind. Additionally, the majority of studies have used fully autonomous robots (68.5 / 100 studies) requiring no human input to control, as shown in Figure 2.3(c). Some studies have used a Wizard of Oz (WoZ) approach (31.5 / 100 studies) to simulate autonomy by involving a human to control aspects of the robot’s behavior (e.g., speech recognition and generation).

We also found that a majority of studies have come from the United States, Japan, and Europe, see Figure 2.4(a). Most of the studies had an experimental study design (81% of the studies). For each between-subjects condition in these experimental studies, the number of groups in each condition ranges from 1 to 373 with a median of 12 groups as shown in Figure 2.4(b). A majority (75%) of the experimental studies in this review relied on only one interaction session, Figure 2.4(c). Exceptions include a study in which a QRIO robot was integrated into a preschool classroom and interacted with preschoolers during *45 distinct sessions* for on average 50 minutes per session over the course of 5 months [Tanaka et al., 2007].

## Robot Behavior in Groups

Just like people, robots can influence group interactions through their nonverbal and verbal behaviors. A robot’s use of nonverbal behaviors (e.g., gaze, proxemics, gestures) can socially cue group members to produce desired responses. Additionally, robots can express emotion and personality verbally, which can shape the overall group dynamic.

A sizable portion of research on robots in groups focuses on ways in which a robot can shape the interaction dynamics between people using nonverbal cues and interventions. Collectively this work demonstrates a powerful influence that robots can exert on groups using gaze, proxemics, and gestures.

Speaking specifically to robot gaze in groups, studies have found that groups of varying sizes can easily recognize a robot’s gaze [Imai et al., 2002] and interpret a robot’s prioritized target from a robot’s gaze cues [Kirchner et al., 2011]. Robots can also use gaze in tandem with other cues, such as smiles and speech pauses, to influence turn-taking between human group members and signal upcoming conversational turns for the robot [Skantze et al., 2015, Skantze, 2017]. Beyond the ability to influence turn-taking, robots have also been shown to shape people’s conversational roles using gaze in a group [Mutlu et al., 2009].

Proxemics, or the way in which a robot is physically positioned in groups, has also been shown to influence human-robot group interactions. People in crowded spaces prefer robots that maintain a comfortable distance [Kidokoro et al., 2013]. People also prefer robots that approach their group when the robot is in the line of sight of group members and when the robot aims to occupy a spatial opening in the group [Ball et al., 2017]. A robot’s body orientation also influences its interactions with groups of people. In a shopping mall, bystander groups have been observed to be larger and more engaged when the robot walked backwards, facing the group, rather than alongside them [Shiomi et al., 2010]. Additionally, a robot that leverages its body position and gaze toward groups in a brainstorming task was found to facilitate feelings of inclusion and belonging to the group [Vázquez et al., 2017]. People also alter their own proxemic distance to robots based on their context and the robot’s navigation strategy. For example, people move closer to a stationary robot if the group contained both a child and an adult [Nabe et al., 2006], and people were

observed to navigate with lower accelerations (indicating possible increased comfort) around robots navigating autonomously with state-of-the-art navigation algorithms as compared with teleoperated robots [Mavrogiannis et al., 2019].

A robot’s physical gestures can enhance its interactions with groups of people. People perceive a robot more positively if it considers the social appropriateness of its pointing gestures (i.e., it is not always socially appropriate to point at people [Liu et al., 2013]). Robots that produce gestures that are more organic and natural, allowing for interruptions in the production of gestures and featuring parameterization of gestures, have been shown to increase both the number of people who communicated with it and the length of the interaction [Kondo et al., 2013]. Additionally, robots are more effective at conveying their arm motion intent when they can balance the legibility and predictability of handover motions when interacting with a group of people [Faria et al., 2017]. Non-anthropomorphic robots that do not use verbal language and only communicate through gesture and movement have been shown to influence people’s gaze towards the robot, perceptions of the robot’s sociality [Hoffman et al., 2015], as well as the evenness of the group’s conversational backchanneling turns [Tennent et al., 2019].

A robot’s speech can powerfully influence both its perceived personality and emotion, which in turn shape the overall dynamic of the group and the behavior of its members.

Robot personality characteristics (e.g., collaborativeness, competitiveness, trustworthiness, and warmth) are often communicated verbally, with profound effects on the group. In a series of studies employing two robots playing a partnered card game with two people, researchers examined how competitive versus relationship oriented personalities impacted group impressions [Correia et al., 2018b, Oliveira et al., 2018, Correia et al., 2017b, Correia et al., 2016, Correia et al., 2017a]. Competitive opponent robots and relationship-driven partner robots received the most gaze attention [Oliveira et al., 2018], but overall, people tended to prefer a relationship-driven robot as a group member [Correia et al., 2017b], at least in the game context. People’s preferences for robot teammates also shifted over time, such that people tend to prefer robot teammates with personalities (collaborative or competitive) that reflect their own [Correia et al., 2018b, Correia et al., 2017b]. Time in general seems to be a critical factor in many instances as trust in human-robot teams forms over

time [Correia et al., 2016, Correia et al., 2017a] and perceptions of and relationships with robots tend to evolve over time as well [Ljungblad et al., 2012].

Robots can also express emotion verbally, shaping how people within a group behave and perceive a group. Robots have been shown to influence groups of people by displaying emotional cues [Correia et al., 2018b], recognizing human emotions and empathizing with members [Leite et al., 2012, Pereira et al., 2011], and shaping the affective status, or mood, of a group and its membership [Alemi et al., 2016, Hebesberger et al., 2016, Utami and Bickmore, 2019, Wada et al., 2004]. However, it is not simply a matter of saying that emotion displays make better robots, as the type of robot-enacted emotion display matters. For example, robots that expressed group-based emotion expressions were perceived as more likable and trustworthy than robots that expressed individual-based emotion expressions in human-robot groups playing a game of cards [Correia et al., 2018b]. This research as a whole suggests great promise in using robots to facilitate positive group emotion, but there remain many gaps for researchers to explore in understanding precisely when and in what contexts the range of emotional expression may be influential or effective.

## Interaction Context

To gain better understanding of the roles robots take on as members of human-robot groups, we examined the settings in which group interactions occurred and the roles that the robots performed within those settings. We distinguish between lab settings (environments controlled by and chosen by researchers) and field settings (natural settings where participants would be found even when the experiment was not taking place). For field studies, we further categorized each study according to a specific setting (e.g., museum, shopping mall).

Additionally, we distinguished between three roles the robot took on during studies: *follower*, *peer*, or *leader*. A robot in the role of a *follower* reacts to interaction initiatives from people, follows instructions, or performs a service task to help people (e.g., a hospital materials delivery robot). A robot in the role of a *peer* is positioned similarly to a human in initiating and driving interactions (e.g., a robot collaborating as a partner on a shared task). A robot in the role of a *leader* initiates and guides interactions or facilitates the behavior of the people it interacts with (e.g., a robot tutor). Figure 2.5(a) visualizes the

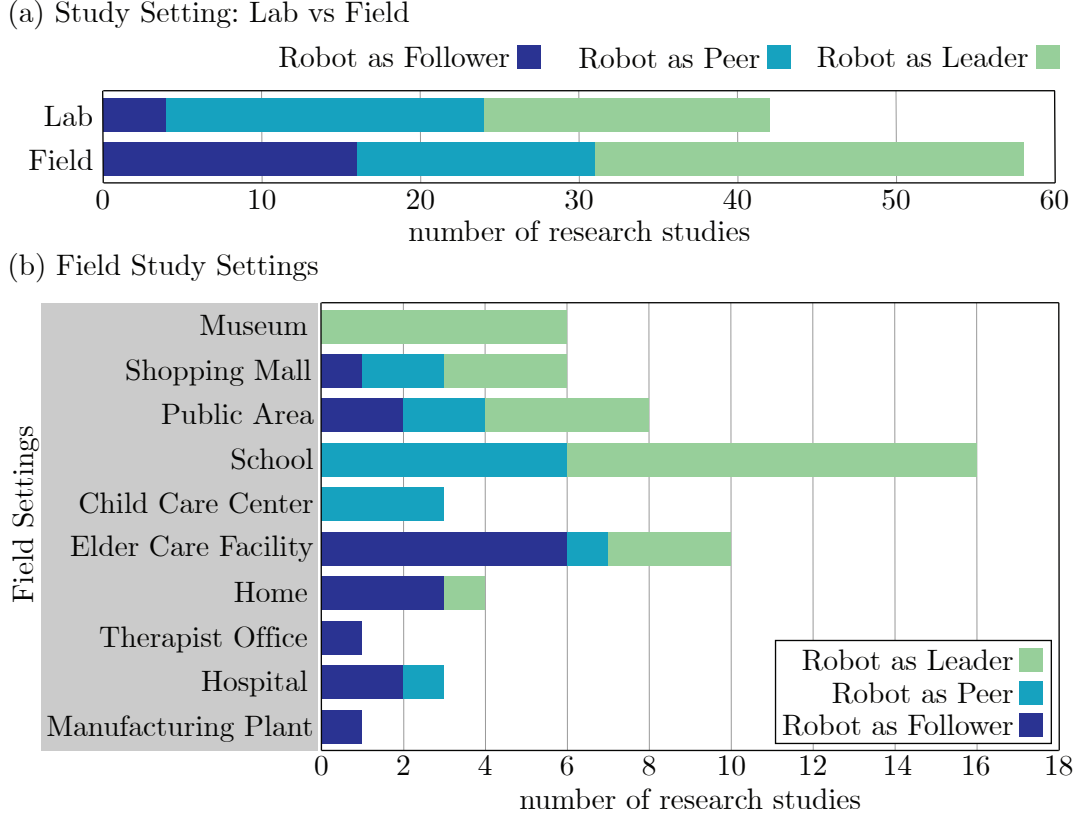


Figure 2.5: We visualize (a) the number of studies conducted in the lab and in the field and the robot roles found within each setting and (b) the number of studies conducted in specific field setting and the robot roles found within each field setting.

number of studies conducted in both the lab and the field with each of the three robot roles (follower, peer, and leader). None of the studies were conducted in multiple settings, however some of the studies investigated multiple robot roles and the count for each of the roles was incremented by the appropriate fraction (i.e., if a study investigated two robot roles, those roles would have 0.5 added to each for the count).

It is important to observe the large percentage of studies that have been conducted in the field (58.0%), as shown in Figure 2.5(a). This stands in contrast to the human-robot interaction literature in general where, for example, of the full papers containing human-subjects studies accepted to the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI), only 16.7% of the studies were conducted in the field (66.7% were conducted in the lab and 16.7% were conducted online, e.g., Amazon Mechanical Turk). As a result of the high proportion of studies conducted in the field, this body of work

has a strong grounding in and applicability to real-world environments as well as a proved robustness to the more chaotic and complex interactions that occur outside the lab.

Certain types of settings seem to suit particular robot roles better than others, see Figure 2.5(b). In settings where the desired behavior of the robot is repetitive and consistent, especially in conveying information, robots are often put in the role of a leader, for example, explaining museum exhibits [Skantze et al., 2015, Skantze, 2017, Shiomi et al., 2007, Yamazaki et al., 2012], giving directions to people in a shopping mall [Shiomi et al., 2010, Sabelli and Kanda, 2016], and tutoring children [Alemi et al., 2015, Chandra et al., 2015, Fernández-Llamas et al., 2017, Kanda et al., 2012]. In settings where robots are designed to provide companionship to people, robots are often given the role of a peer, such as playing with children in day care centers [Tanaka et al., 2007] and learning alongside children in educational settings [Hood et al., 2015, Matsuzoe et al., 2014]. In complex settings where robot mistakes can be costly, robots are positioned in the role of a follower, where their actions are either controlled or monitored and can be corrected by the people around them, for example, delivering items within a hospital [Ljungblad et al., 2012, Mutlu and Forlizzi, 2008], working alongside people in a manufacturing plant [Sauppé and Mutlu, 2015], and vacuuming people’s homes [Forlizzi and DiSalvo, 2006, Forlizzi, 2007, Sung et al., 2010].

As shown in Figure 2.5(a), nearly half of the studies conducted in lab settings have investigated robots in peer roles (47.6%), whereas in field settings the smallest percentage of studies have investigated the robot in a peer role (25.9%). Robots in the field are usually either programmed to convey the same information doing repetitive tasks in the role of a leader or are designed to be under the direct supervision of people in the role of a follower. The lack of peer robots studied in the field is likely due to the challenge of equipping a robot in unconstrained settings with all the necessary skills and knowledge needed to effectively interact with people as a peer. Many of these essential skills are not unique to robots interacting with groups, such as natural language understanding, intent prediction, and emotion expression detection. In addition to these skills, robots interacting with groups of people as a peer and in more flexible and complex roles also must construct models of the relationships between the people with whom they interact, choose which person or people

to address, and predict how their actions influence multiple different people. As these underlying technological components that support robots interacting socially with groups of people improve, robots will be able to take on more sophisticated, flexible, and complex roles in the unstructured and unpredictable field settings.

Additionally, some settings have received more attention than others, as displayed in Figure 2.5(b). For instance, about twice as many studies have been conducted with children in school settings than with adults in the workplace (e.g., hospital, manufacturing plant, therapist office) and people in home environments combined. In particular, these two environments, adult workplaces and homes, are the places where people spend a majority of their time and where robots have already had great influence and impact (e.g., vacuum cleaning robots, voice assistant devices, manufacturing plant robots, mobile delivery robots). As research continues exploring robots interacting with groups of people, more studies examining robots in adult workplaces and home environments are necessary to better understand the influence of robots on people in these environments and advance the robotic technology necessary for robots to operate effectively in these important settings.

### **Robot Influence on Human-to-Human Interactions**

Robots are not only able to shape how groups of people interact with it, there is also increasing evidence that robots can influence the relationships and interactions that people have with *the other people in the group*. Our work presented in this dissertation has contributed significantly to this exploration of how robots can shape human-to-human interactions in collaborative teams. In the broader HRI literature, robots have been shown to shape human-to-human interactions in groups by increasing human social connection, mediating conflict, and shaping positive team dynamics.

Across a variety of settings, there is evidence that robots can encourage and increase social interactions among the people in a group with one another. Robots have demonstrated a positive influence on the amount of verbal communication and interaction among older adults within care facilities [Šabanović et al., 2013, Thompson et al., 2017]. Studies of robots moderating inter-generational groups [Joshi and Šabanović, 2019, Short et al., 2017] have shown promise in engaging multiple generations in meaningful interaction. Robots



that promote social skills development in children with ASD have shown to be effective in increasing social engagement between these children and others in their group, whether with their caregiver [Scassellati et al., 2018], another playmate [Kim et al., 2013], or with a therapist [Zubrycki and Granosik, 2016].

In moments of conflict between human members of a human-robot team, a robot’s actions can influence how conflict is resolved. For example, robots have demonstrated success in directly mediating resource conflicts (e.g., fighting over the same toy) between children [Shen et al., 2018]. Another study showed that a robot intervening in a team’s conflict after a team member made a hostile remark increased the salience of conflict and forced team members to actively engage with the conflict [Jung et al., 2015].

Robots have also demonstrated the ability to shape human-robot team dynamics, positively influencing how people interact with each other. For example, our work, detailed in Chapter 3, demonstrates that a robot can (1) improve performance in a collaborative game between pairs of children by asking task-focused questions and (2) perceptions of performance on the same task between pairs of children by asking relationship-oriented questions [Strohkorb et al., 2016]. Another study that used a robot moderator during a three-person collaborative game showed that group cohesion could be actively influenced by the robot based on its behavior [Short and Matarić, 2017]. Tennent et al. (2019) introduced a swiveling microphone robot (‘Micbot’) capable of facilitating more balanced participation during a three-person team’s decision making discussion. Finally, our work, detailed in Chapter 5, shows that a robot’s verbal expressions of vulnerability can have “ripple effects” in a group by increasing how likely human members of the group are vulnerable with one another [Strohkorb Sebo et al., 2018]. These studies illustrate the influence robots have to shape group dynamics and the behavior of people in a group through direct intervention, peripheral non-verbal movement, and indirect verbal expression.

## **Robots in Groups and Teams Review Summary**

As robots interact with people in increasingly complex settings, with more diverse roles, and over longer periods of time, these interactions will rarely resemble the dyadic interactions historically studied in the field human-robot interaction. The body of work highlighted in

this review has taken some first steps in the direction of equipping robots with the abilities to interact with groups of people, often in complex field settings, and studying the effects of robot actions. As researchers in this field work to address the current technical and methodological challenges involved with group interactions, we can work to develop and study robots that richly interact with many groups over long periods of time in natural environments.

### **2.2.2 Trust in HRI**

Researchers in human-robot interaction have become increasingly interested in the factors that influence people’s trust of robots in a variety of contexts, including household assistant robots [Salem et al., 2015], UAVs [Freedy et al., 2007], autonomous cars [Waytz et al., 2014], and tour guides [Andrist et al., 2013]. Similar to trust between people, human-robot trust and research can be divided into two categories: competence-related trust and integrity-related trust.

A majority of research into human-robot trust has focused on competence or performance based trust. Robot performance is considered to be the most influential factor in human-robot trust [Hancock et al., 2011], likely due to the importance of the robot’s ability to meet performance expectations [Kwon et al., 2016]. Children as young as 3-5 years of age trust a robot less when that robot has made errors in the past [Geiskkovitch et al., 2019]. Recent work with adults has shown that initial performance failures in a human-robot interaction are more detrimental to ratings of robot trustworthiness than failures later on in the interaction [Desai et al., 2013, Robinette et al., 2017]. Researchers have also successfully employed models of competence-based trust of robots used in robot decision making [Chen et al., 2018] and evaluations of human-robot team effectiveness [Freedy et al., 2007].

Despite the large focus on performance-based trust, a growing body of work has also demonstrated the importance of integrity based trust. Integrity related trust, or interpersonal trust, can be described as the level of expectation that another is predictable, dependable, and can be relied upon in the future in the context of a social relationship [Rempel et al., 1985]. Many parallels exist between interpersonal trust between humans and interpersonal trust between a human and a robot. DeSteno et al. (2012) demonstrated that

just as humans are perceived as less trustworthy when they exhibit nonverbal signals that indicate distrust, a robot is also perceived as less trustworthy when it displays those same nonverbal signals [DeSteno et al., 2012]. Additionally, several studies have shown that a robot’s vulnerable disclosures increase people’s feelings of liking [Siino et al., 2008], companionship [Martelaro et al., 2016], and trust toward the robot [Kaniarasu and Steinfeld, 2014, Martelaro et al., 2016].

Although there has been a great focus in HRI on the trust between a human and a robot, no work to our knowledge has investigated how a robot’s actions can either influence trust at the group level or trust-related actions between human team members. Our work detailed in Chapter 4 contributes to the one-on-one human-robot trust literature by examining novel methods of trust violation repair. Then, in Chapter 5, we describe our work that demonstrates the ability of a robot to shape trust-related behavior and conversational dynamics using vulnerable expressions. - the first work to show that a robot’s actions can shape how people in a human-robot group interact with one another.

### **2.2.3 Inclusion & Psychological Safety in HRI**

No work to our knowledge has investigated a robot’s influence on either the inclusion of human team members or the team’s psychological safety. In order to capture the most relevant work in HRI to inclusion and psychological safety, we review work in HRI focused on affect.

Robots have both used affect expression to influence human-robot interactions and intervened in other ways to improve human affect. A robot that expressed empathy toward one player in a game was rated as having characteristics descriptive of a friend [Leite et al., 2012, Pereira et al., 2011]. In a game with two human players and two robot players, the human partners directed their gaze more often to a relationship-driven robot when they were partners with it and to a competitive robot when they were opponents with it [Oliveira et al., 2018]. In a similar setup with two human and two robot players, people displayed higher levels of affinity, group identification, and group trust toward a robot that expressed group-based emotions than toward a robot that expressed individual-based emotions [Correia et al., 2018b]. Robot interventions in a variety of settings have also led to positive results

of affect: a robot therapist demonstrated improvements in couples' intimacy and positive affect [Utami and Bickmore, 2019]; a robot used as a therapy assistive tool for pediatric oncology patients was shown to relieve stress, depression, and anger in children [Alemi et al., 2016]; and a robot programmed to guide the elderly in a walking group positively influenced the group's coherence and motivation [Hebesberger et al., 2016].

This growing body of work in HRI examining a robot's capability to shape group affect gives us a positive indication that it is likely that a robot can also shape group members' inclusion and psychological safety. In Chapters 6 and 7, we present the first work demonstrating that a social robot can shape team members' inclusion and psychological safety. Additionally, in Chapter 7 we establish a clear link between the low-level behavior of backchanneling with these two important social dynamics of inclusion and psychological safety.

## 2.3 Summary

In this chapter, we reviewed the driving factors for success in human and human-robot teaming. We broadly examined the literature examining human collaborative teams as well as robots that interact with groups and teams. We also explored three social dynamics (trust, inclusion, and psychological safety) that have been proven to be essential for team success in human groups, and any work within the field of human-robot interaction related to those three dynamics. In the following chapters, we explore how these collaborative social dynamics in human-robot teams can be positively shaped through intelligent robot behavior.

## Chapter 3

# Robots that Shape Collaboration between Pairs of Children<sup>\*†</sup>

Although prior work has demonstrated that social robots can shape how people in the group interact with the robot [Shiomi et al., 2010] and their general perceptions of the group [Vázquez et al., 2017], no work has yet shown how a social robot can influence how people in a group interact *with each other*. In this chapter, we present a first investigation into whether a social robot can influence how two children collaborate together on a shared task. We focus on children ages 6 to 9 because around this age children become capable of collaborating and are learning how to collaborate with their peers [Warneken et al., 2014] and might benefit the most from robot interventions to promote collaborative behavior. We conducted a between-subjects study where pairs of children play a collaborative game with a social robot. During pauses in the game, the robot either (1) asks the children questions to better focus the participants on the task they are working on, (2) asks the children questions that are targeted at developing and reinforcing the relationship between the participants, or (3) doesn't ask any questions. Our results show that participants who

---

<sup>\*</sup>Portions of this chapter were originally published as: S. Strohkorb, E. Fukuto, N. Warren, C. Taylor, B. Berry, and B. Scassellati. Improving Human-Human Collaboration Between Children With a Social Robot. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN '16, pages 26-31, New York, NY, USA, 2016. ACM. [Strohkorb et al., 2016]

<sup>†</sup>This work was published in 2016, when Sarah was publishing under the name Sarah Strohkorb. For the remainder of her work that is included in this dissertation, Sarah published under the name Sarah Strohkorb Sebo, where Strohkorb Sebo is considered her last name.

were asked task-focused questions had higher performance scores in the collaborative game than those who were asked no questions. However, despite their good performance, those who were asked task-focused questions had a lower perception of their performance than the participants who were asked relationally-focused questions. We did not find any differences between the groups in interpersonal cohesiveness. Our findings suggest that social robots can be used to improve performance and perception of performance in groups of children.

### 3.1 Introduction

The ability to collaborate with other people is an essential skill that children begin to develop early in life [Eckerman and Peterman, 2001]. Collaborative problem solving requires a person to reason not only about their own actions, but also the actions and intentions of others [Tomasello et al., 2005]. In an experimental study, Warneken et al. (2014) demonstrated that between the age of 3 and 5, children develop the capacity to plan the division of labor in a collaborative task.

Prior work in psychology suggests two distinct approaches of enhancing collaboration between people: improving task cohesiveness and improving interpersonal cohesiveness. Craig and Kelly (1999) describe task cohesiveness as “a group’s shared commitment, or attraction to the group task or goal” and interpersonal cohesiveness as “the group members’ attraction to or liking of the group.” In an experiment, they instructed groups of three adults to create a technical drawing after a manipulation of the group cohesiveness. Groups given a high task cohesiveness manipulation created drawings of higher technical quality, whereas groups with a high interpersonal cohesiveness manipulation had drawings of higher creativity [Craig and Kelly, 1999]. These two strategies of influencing collaboration between humans each achieve a productive result; however, the results themselves and the methods of reaching them are noticeably distinct and should be employed differently depending on the desired outcome.

In this study, we seek to promote the growth and use of collaborative skills in children by building a robot that promotes collaboration through both strategies outlined above: improving focus on the task and enhancing interpersonal cohesiveness. We decided to focus



Figure 3.1: Pairs of children age 6-9 collaborated with one another to play a rocket building game with a social robot that asked them relationship-focused questions, task-focused questions, or no questions.

on children between the ages of 6 and 9 years old because a child’s ability to plan and collaborate emerges between the ages of 3 and 5 [Warneken et al., 2014]. Thus, children between the ages of 6 and 9 can be assumed to have the capacity for collaborative activity and would also likely benefit from interventions to improve collaborative interactions. During the experiment, two children and a robot play an interactive tablet game, shown in Figure 3.1, during which the robot will use one of the given strategies to promote collaboration.

## 3.2 Methods

In this experiment, we explore the benefits of two strategies of promoting collaboration between children: 1) encouraging task-focused strategy discussion and 2) developing and reinforcing the relationship between the children. We measure the success of these strategies by an objective performance measure in a collaborative game as well as the participants’ perception of their performance and interpersonal cohesiveness. With these strategies and metrics in mind, we form the following hypotheses:

*Hypothesis 1: Individuals who are asked task-focused questions by a social robot will have better **performance outcome measures** than individuals who are asked relationship-reinforcing questions or no questions by a social robot.*

*Hypothesis 2: Individuals who are asked relationship-reinforcing questions by a social robot will **perceive their team performance** as better than individuals who are asked task-focused questions or no questions by a social robot.*

*Hypothesis 3: Individuals who are asked relationship-reinforcing questions by a social robot will **perceive their interpersonal cohesiveness** as better than individuals who are asked task-focused questions or no questions by a social robot.*

To examine these hypotheses, we had two participants play a collaborative game with a robot, who acted as a peer. We chose the peer role for the robot because robot peer characters have been shown to elicit more attention from children and improved performance than more authoritative tutoring robot characters [Zaga et al., 2015]. The experiment has the following three conditions:

1. **Task:** The robot asks questions during pauses in a team-oriented game that aim to better focus the participants on the task they are working on.
2. **Relational:** The robot asks questions during pauses in a team-oriented game that are targeted at developing and reinforcing the relationship between the participants.
3. **Control:** The robot does not say anything during pauses in the game.

We chose to have a control condition where the robot did not say anything. We could have constructed a control condition where the robot made utterances that were neither task related nor relationship related. However, this kind of control would be difficult to both fit within the context of the interaction without seeming surprising or strange and also be truly neutral (not containing any affect, not relating to the task at all). For these reasons we determined that a control condition where the robot did not say anything would be best choice for a control condition.



Condition	Age 6	Age 7	Age 8	Age 9	Total
Relational	3	6	3	2	14
Task	4	4	4	2	14
Control	5	2	3	2	12

Table 3.1: We present the age distribution of the dyads in each experimental condition by the number of dyads in each age and experimental group.

Condition	2F	1F & 1M	2M	Total
Relational	5	5	4	14
Task	3	7	4	14
Control	1	8	3	12

Table 3.2: We present the gender composition of the dyads in each experimental condition by the number of females (F) and males (M) in each experimental group.

In accordance with our hypotheses, we expect that participants in the task condition will have higher performance outcomes in comparison with the other conditions. We also expect that participants in the relational condition will have higher perceptions of their team’s performance and interpersonal cohesiveness in comparison with the other conditions.

### 3.2.1 Participants

The participants in this study were attendees of one of two educational summer programs located in Connecticut, USA. A total of 88 participants were recruited from these summer programs, however, 1 dyad (2 participants) was excluded because they did not complete the interaction and 3 dyads (6 participants) were excluded because they did not pay attention to a majority of the questions asked by the robot. Of the participants included in this analysis, all participants were between the ages of 6 and 9 ( $M = 7.25, SD = 1.05$ ), 42 of the participants were male, and 38 of the participants were female.

Participants were paired in such a way that the participants in each dyad were the same age in order to maintain a more equivalent power dynamic between the two participants and so that comparisons between dyads of different ages could be made. The number of dyads of each age in each experimental condition are shown in Table 3.1. There were 20 mixed gender dyads and 20 same-gender dyads. Among the same-gender dyads, there were 11 dyads with two males and 9 dyads with two females. The age and gender dyad characteristics were

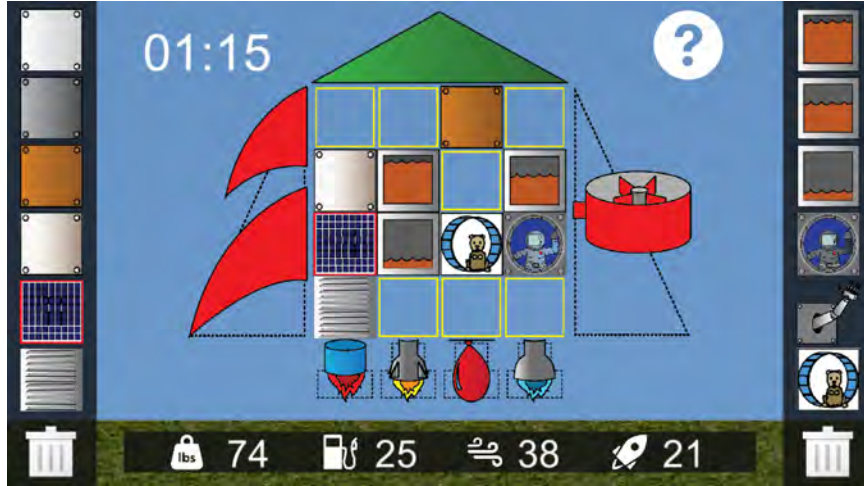


Figure 3.2: In the build-a-rocket game, players drag and drop pieces to construct a rocket by optimizing weight, fuel, air resistance, and power metrics (shown on the bottom panel). Time remaining to takeoff is shown in the upper left hand corner. Players can drag a piece over the white question mark to ask the robot about that piece’s weight.

evenly distributed across the three conditions, see Table 3.2 for the number of dyads with each gender combination for each experimental group and Table C.1 in Appendix C for a more detailed list of descriptive statistics for each experimental condition.

### 3.2.2 Build-a-Rocket Game

We custom built the build-a-rocket game, pictured in Figure 3.2, in Unity for the pairs of children to play on a 27-inch multi-touch touchscreen monitor. The goal of the build-a-rocket game is to build a rocket that flies as high as possible. Players touch a part of the rocket (body, boosters, fins, or cone) they want to place a piece on, after which the side panels display pieces that can be placed on that part of the rocket. Players drag and drop pieces onto the rocket and dispose of pieces by moving the pieces to the trash cans or the side panels. Players may also drag a piece over the question mark to ask the robot how much that particular piece weighs.

Players have 7 trials to try and make the rocket fly as high as they can. Each trial lasts 2.5 minutes, after which the rocket has a ‘blastoff’ animation and then displays the height the rocket reached. There is a 45-second pause between each trial where the only visual on the screen is a list of the heights the rocket has reached for each completed trial. Once the

45 seconds have elapsed, the next trial automatically begins.

The rocket distance ( $D$ ) is calculated with the following formula:  $D = p(\alpha_1 F + \alpha_2 (F * P) - \alpha_3 W - \alpha_4 R_{air} + \beta)$ , where  $F$  is the rocket fuel,  $P$  is the rocket power,  $R_{air}$  is the rocket air-resistance,  $W$  is the rocket weight,  $p$  is a penalty for not having pieces filled in, and  $\alpha$  and  $\beta$  are constants. This equation is not meant to simulate real-world rocket dynamics, but rather, the intuitive relationship of each of the four factors highlighted in the game (fuel, power, weight, and air resistance). Weight ( $W$ ) and air resistance ( $R_{air}$ ) are negatively correlated with rocket distance. Fuel ( $F$ ) and power ( $P$ ) are positively correlated with rocket distance, where power is dependent on fuel and the presence of boosters. Additionally, just as any rocket with pieces missing would not perform as well, we penalize any rocket that does not have all of its pieces filled in with  $p$ , a proportion of the pieces on the rocket to the total number of possible pieces that the rocket could hold.

### 3.2.3 System Architecture

The robot platform we use is a MyKeepon robot, a commercially available and inexpensive robot shown in Figure 3.3. MyKeepon is a 32cm tall snowman-shaped robot with a yellow rubber skin and four degrees of freedom: rotation around the base, left/right roll, front/back tilt, and up/down bob. MyKeepon is a consumer-grade version of a research robot called Keepon Pro, which was designed to convey expressions of emotion and attention with a minimal design [Kozima et al., 2009]. We modified a MyKeepon to control its motors with an Arduino Nano, which sends motor commands to the MyKeepon’s four motors.

The system architecture used to autonomously control the robot’s behavior (movement and speech) uses Thalamus [Ribeiro et al., 2014]. Thalamus is an integration middleware that allows many modules to connect to and communicate with each other.

There are two inputs to the robot’s decision making system: rocket game information related to the actions of the children in the game as well as audio and face/body position data from a Microsoft Kinect. The game information sent to the system includes timer values, rocket informational values (weight, fuel, air resistance, and power), rocket flight distance values, and specific game events (moving a piece over the question mark). The Microsoft Kinect relays information about the participants’ 3D positions, facial features,

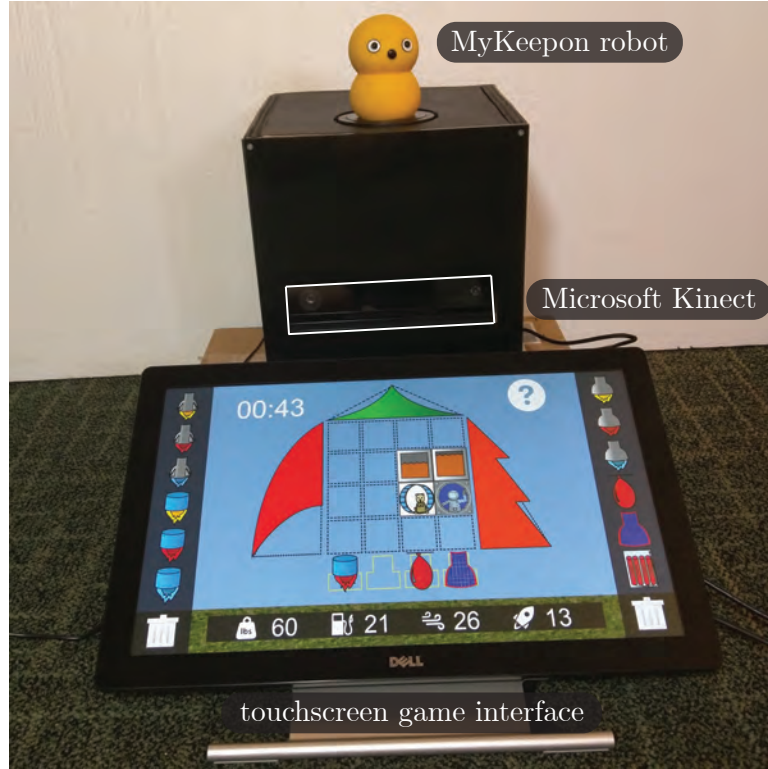


Figure 3.3: A MyKeepon robot interacts with two children playing the build-a-rocket game on the touchscreen game interface. The robot uses information gathered from a Microsoft Kinect to track the faces of the children to inform its gaze direction.

and audio.

Upon receiving the rocket game information, our system decides how the robot should respond. During game play, the robot reacts to game events: informing the participants of the weight of rocket pieces dragged to the question mark on the screen and warning the participants of a high rocket weight. During pauses in the game, the robot asks participants questions in the task and relational conditions. Once an utterance has been selected, a command is sent for the utterance to be made using text-to-speech (TTS) via visemes, positions of the face and mouth that correspond to a sound. These visemes are made available to Nutty Tracks [Ribeiro et al., 2014], a generic animation engine. A module in Nutty Tracks sends commands to the robot motors while it talks, giving the robot an appearance of bouncing while it is talking.

The Microsoft Kinect data is used to calculate the location and head positions of the participant who spoke most recently or is closest to the robot. With this information, the



Figure 3.4: There were 43 pairs of children that participated in this experiment. These photos depict several pairs of children interacting with each other as they play the collaborative build-a-rocket game.

robot looks at the participants appropriately during the interaction, giving it an increased sense of social presence. A Nutty Tracks module sends commands to the robot motors to have the robot face the selected participant.

### 3.2.4 Procedure

Consent forms were distributed and collected by staff of the summer programs. Participants were paired with a partner of their same age and once paired, the dyad was randomly assigned to one of the three conditions (task, relational, or control).

Once participants were selected by program staff, they were escorted by one of the experimenters to the experimental area. Each participant was separately interviewed by one of two experimenters to assess their prior familiarity with the other participant. The interview was captured with an audio recording device. Directly after the pre-experiment survey, each experimenter gave the participants distinct and specialized instructions to encourage collaboration during the experiment. One participant was taught about how air resistance influences rocket flight and was shown examples of rocket pieces that have low and high air resistance. The other participant was taught about how fuel and power influence rocket flight and was shown examples of which rocket pieces have low and high fuel and power.

Next, the experimenters led the participants into the room with the autonomous robot,

Orion (a MyKeepon robot). One experimenter and Orion performed a pre-scripted dialogue where the goals of the build-a-rocket game were explained, participants were told that Orion had specialized knowledge about the weight of the pieces, and participants were shown how to play the game. Once the game began, the participants had 7 trials, each lasting 2.5 minutes, to make the rocket go as far as possible. During the game play (see Figure 3.4 for photos of participants playing the build-a-rocket game), participants could ask Orion questions about the weight of specific rocket pieces by dragging pieces over a question mark on the screen. Orion responded to these ‘questions’ and also interjected with comments about the overall rocket weight to contribute to the team conversation. Orion ran autonomously and did not react to any speech directed toward him by participants.

Between each trial, there was a 45-second pause where Orion asked each child a directed question, unless the dyad was assigned to the control condition. The questions asked in the task and relational conditions are shown in Table 3.3. After the seventh trial, the ‘game over’ screen appeared to mark the end of the game. The game interaction with Orion and the participants was recorded with a video camera, and at least one experimenter was present in the room at all times.

After the game had finished, the experimenters conducted separate final interviews with the participants. Like the pre-experiment survey, this survey was captured with an audio recording device. After the participant completed the interview, the experimenters gave each child pencils and stickers for participating.

### **3.2.5 Measures**

In this section, we describe how the survey data was coded and how the performance metrics were calculated.

#### **Friendship and Familiarity**

During the pre-experiment interview, participants were asked questions to assess the level of friendship and familiarity between them and their partner. This pre-experiment interview consisted of 10 questions to measure the level of friendship and familiarity between the participants and how likeable the participant finds their partner. These questions were

Table 3.3: The robot asked questions to the two child participants after each round of the game in the task and relational conditions, and did not say anything to the child participants who were in the control condition. These questions are displayed in this table, where [P1] and [P2] act as placeholders for participant names. The questions indicated by a (\*) were asked to the child in the task condition who had learned specifically about that topic.

<b>Task Condition Questions</b>
[P1], what do you think made the rocket go farther this time?
[P2], do we have enough fuel?*
[P1], which cone pieces do you think are the best?*
[P2], what do you want to change about the rocket next time?
[P1], which pieces are contributing most to weight?
[P2], what do you think the best rocket would look like?

<b>Relational Condition Questions</b>
[P1], does [P2] think you did a good job?
[P2], is there a way for you to help [P1] better next time?
[P1], what do you think [P2] did well last time?
[P2], how did [P1] help you in building the rocket?
[P1], what was [P2]’s goal?
[P2], did you always ask for help when you needed it?

adapted from the Friendship Qualities Scale to be child-friendly, such as, “If you forgot your lunch, would they share theirs with you?” [Bukowski et al., 1994]. Please refer to Appendix B, Section B.1 for the full questionnaire.

Two coders listened to the audio-recorded responses for each question and categorized them as either ‘yes’, ‘no’, or ‘unsure’. The coders had 100% agreement in their categorization of the responses. From the answers to these questions, we created two levels of friendship and familiarity for the participants: low familiarity and high familiarity. Participants were categorized as highly familiar (1) with their partner if they answered questions indicating that they had experience playing together outside of the summer program, and lower otherwise (0).

### Perception of Performance and Interpersonal Cohesiveness

In the post-experiment interview, participants were asked questions to assess their perception of their own performance and the interpersonal cohesiveness between them and their

partner. These questions were adapted from the Subjective Value Inventory questionnaire, originally designed to assess the success of negotiations [Curhan et al., 2006]. The Subjective Value Inventory questionnaire has four dimensions: feelings about the outcome, feelings about the self, feelings about the process, and feelings about the relationship. We believe that the Subjective Value Inventory extends well to assessing the perceived success of collaboration between the two participants. We altered questions from the Subjective Value Inventory questionnaire to be child-friendly and specific to the build-a-rocket game, see Appendix B, Section B.2 for a list of all of the items in this questionnaire.

To measure the perception of their performance, participants were asked one question about how high their rocket flew (“Did your rocket go higher and higher each time? Or did it reach about the same height each time?”) and one question about their satisfaction with their performance (“Did your rocket go as high as you and [your partner] wanted it to?”). Two coders listened to the audio-recorded responses and categorized the answers to each question as either high (2), medium (1), or low (0). The coders had 100% agreement in their categorization of the responses. We added the score of these two questions together for an overall value of participants’ perception of their performance, where high values indicate a high perception of performance.

To measure participants’ perceived interpersonal cohesiveness between them and their partner, participants were asked if their partner listened to them, if their partner annoyed them, if they were to play the game again would they prefer to play alone or with their partner, and who they would play the game with if they could choose anyone. Two coders listened to the audio-recorded responses and categorized the answers as either yes (2), maybe (1), or no (0) or selecting their partner (2), not selecting their partner (0), or being unsure (1). The coders had 100% agreement in their categorization of the responses. We added the score of these four questions together for an overall value of participants’ perceived interpersonal cohesiveness between them and their partner, where high values indicate a high perception of cohesiveness.



## Build-A-Rocket Game Performance

To assess participants' performance in the build-a-rocket game, we selected the highest (maximum) distance their rocket reached of the game's seven trials.

### 3.3 Results

Hypothesis testing was conducted<sup>‡</sup> using a one-way analysis of variance (ANOVA) model for data describing the group's efforts (e.g., maximum distance participants' rocket reached), where we report the effect size of partial eta squared ( $\eta^2$ ). For data specific to an individual (e.g., each participants' perceptions of their team's performance), we used linear mixed-effects models to account for each participant being in a group of two. We tested these linear mixed-effects models for multicollinearity (variance inflation factor), selected them based on the Akaike information criterion, and evaluated residual errors for lack of trends and heteroscedasticity. For each fixed effect, the model outputs the linear coefficient ( $c$ ), the standard error ( $SE$ ), and the significance ( $p$ ) value of that predictor. For more details on the models run and their results, please refer to Appendix C, Tables C.1 - C.5.

To test our first hypothesis that the task condition would perform better than the other two conditions in the build-a-rocket game, we conducted planned comparisons between the task condition with the relational and control conditions on the maximum rocket distance. To test our second and third hypotheses that the relational condition would have a better perception of the outcome and interpersonal cohesiveness than the other two conditions, we conducted planned comparisons between the relational condition with the task and control conditions on both the participants' perception of their performance and participants' perception of the interpersonal cohesiveness between them and their partner.

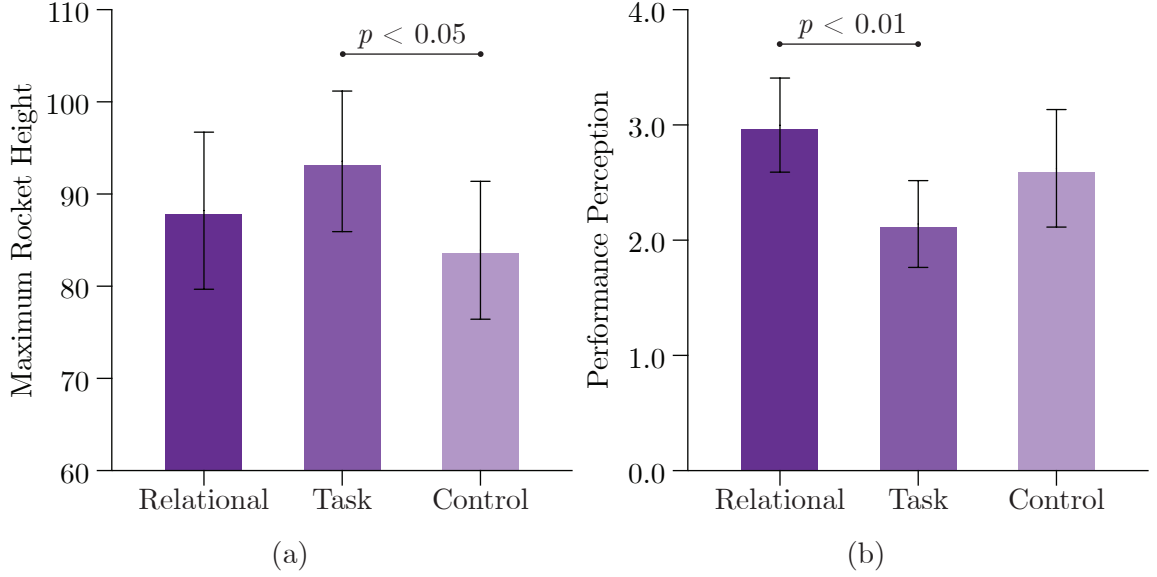


Figure 3.5: For each condition (relational, task, and control) we display the (a) teams' average maximum rocket height scores and (b) the scores of participants' perception of their performance in the build-a-rocket game. Error bars represent a 95% confidence interval.

### 3.3.1 Performance Outcome

To test Hypothesis 1, we examined whether participants in the task condition performed better in the build-a-rocket game than those in the relational and control conditions. We conducted an ANOVA on the maximum distance the rocket reached for each participant with fixed factors of the experimental condition as well as the following covariates: the gender composition of the dyad, the average friendship and familiarity of the participants with one another, and the age of the participants. We did not find a significant main effect for condition,  $F(2, 37) = 1.90, p = 0.165, \eta^2 = 0.09$ . Then, when conducting our planned comparisons, we found that participants in the task condition ( $M = 93.6, SD = 13.2$ ) had significantly higher maximum rocket height scores in the build-a-rocket game than those in the control condition ( $M = 83.9, SD = 11.8, F(1, 24) = 4.85, \eta^2 = 0.12, p = 0.040$ ) but not those in the relational condition ( $M = 88.2, SD = 14.8, F(1, 26) = 1.17, \eta^2 = 0.08, p =$

<sup>‡</sup>After publishing these results [Strohkorb et al., 2016], we realized that there were better statistical tests and models to perform on the data to account for the dyadic grouping of participants. This dissertation presents this improved analysis of the data. Although different models are used to analyze this data, the general results of the paper and conclusions drawn from the data remain unchanged.

0.292), see Figure 3.5(a) <sup>§</sup>. Thus, Hypothesis 1 is partially supported since participants in the task condition performed better than participants in the control condition.

### 3.3.2 Perception of Performance

To test Hypothesis 2, we examined whether participants in the relational condition had a higher perception of performance and perception of the interpersonal cohesiveness between themselves and their partners than the task and control conditions.

First, we consider whether participants in the relational condition had a higher perception of performance than those in the task and control conditions. The linear mixed-effects model that best fit the data included a covariate of whether the individuals in the dyad was the same gender (0) or had different genders (1). We found that participants in the relational condition had higher perceptions of their performance ( $M = 3.00, SD = 1.05$ ) than participants in the task condition ( $M = 2.14, SD = 0.97, c = 0.93, SE = 0.31, p = 0.005$ ). No significant difference was found between participants in the relational condition and participants in the control condition ( $M = 2.63, SD = 1.21, c = 0.52, SE = 0.33, p = 0.124$ ). These results are shown in Figure 3.5(b). This is an interesting result because even though participants in the task condition seemed to performed better in the build-a-rocket game, they perceived their performance as worse than the participants in the relational condition. We can conclude that Hypothesis 2 has moderate support, since individuals have better perceptions of performance than when a social robot asks relationship-reinforcing questions than task-focused questions.

### 3.3.3 Perception of Interpersonal Cohesiveness

Finally, to test Hypothesis 3, we examined whether participants in the relational condition had a higher perception of the interpersonal cohesiveness between them and their partner than the task and control conditions. We did not find a significant main effect or significance in our planned comparisons. We, thus, have no support for Hypothesis 3 since there is

---

<sup>§</sup>After performing a more accurate statistical analysis on this data, as compared with our published work [Strohkorb et al., 2016], we have found that there no longer exists a significant difference between the task and relational condition. This is the only finding that is different between the more accurate analysis performed in this dissertation and the analysis in our prior published work [Strohkorb et al., 2016].

not strong evidence that participants in the relational condition perceived the interpersonal dynamics between themselves and their partners as better than those in the task and control conditions.

### 3.4 Discussion

Our results show that social robots can influence the outcomes of collaboration among children. When a social robot asked task-focused questions during pauses in a collaborative rocket-building game, participants constructed rockets that flew higher than when the robot asked no questions during pauses in the game. A possible explanation for this result is that the social robot helped the children focus on the task, sparking the discussion of related strategies and the development of new ideas.

In addition to the social robot affecting the outcome of collaboration between children, we also expected the robot’s questions to influence how participants perceived their performance. We found that participants to whom a social robot asked relationship-reinforcing questions perceived their performance as better than participants to whom the robot asked task-focused questions, but not better than participants to whom the robot asked no questions. We did not find any difference in the perceived interpersonal cohesiveness in participants between any of the conditions. These results suggest that focusing children on positively building the collaborative relationship between them and their peer(s) has a better effect on their perception of performance, as opposed to encouraging the children to focus their mental energy on how they can improve their performance.

Even though results suggest that the task and relational strategies of promoting collaboration are both promising avenues for producing positive collaborative behavior, they seem to have contrasting effects in the two types of outcome measures we observed. Notably, participants in the task condition had a higher performance score, however, had a more negative perception of their performance. This finding suggests that reaching the maximum of both objective performance measures and perceptions of performance may not be possible, at least with these two distinct approaches. In further work, it would be interesting to investigate the relationship between these two approaches and what results

could be found from a combination of these two strategies.

While conducting the experiment, we noticed that many factors influence how children interact and collaborate with one another. The gender composition of the dyad had a noticeable effect. Children in mixed-gender pairings seemed to be more timid in their interactions, had less physical contact, and stayed more focused than those in same-gender pairings. The personalities and dominance of the children also drastically affected how the children made decisions, how frequently they were distracted, and the amount they expressed prosocial behavior. As social robots enter collaborative environments with children, these factors should be considered by those seeking to shape the interpersonal relationships between children.

### 3.5 Summary

Collaboration is a necessary skill that children begin to develop early in life. Leveraging prior work that has established that social robots are capable of shaping people’s affect as well as their perceptions of this group, we explored whether a social robot could promote collaboration between human members of a group. In a human-subjects experiment, we investigated the influence of a social robot’s questions (relationship-focused, task-focused, or no questions) on the collaboration between pairs of children (ages 6 to 9). We found that children who were asked task-focused questions had a higher performance in the collaborative task than children who were not asked any questions. We also discovered that children who were asked relationship-focused questions had a higher perception of their performance than children who were asked task-focused questions.

These results are among the first to demonstrate that the social actions of a robot can both influence the performance of children in a collaborative task, as well as how they perceived their performance in the task. From this study, we can see the possible benefits of social robots focusing people in the group both on the task at hand and on their relationships with one another. Focus on the task produced higher performance scores and focus on the team member relationships produced higher perceptions of performance. Although it may be hard to evaluate how these strategies may work in long-term collaborations, it seems that

both of these social robot intervention strategies are promising for enhancing collaboration.

In this chapter, we focused on broadly shaping collaboration between pairs of children by asking questions to focus the children on the task and on the relationship between the two of them. Despite our ability to show that task and relationship focused questions do result in changes in performance and performance perceptions, future work is needed to further explore the specific factors and interaction dynamics that influence social collaboration and how social robots can best support human-human social collaboration. In Chapters 4 - 7 we do just that, examining three specific collaborative social dynamics (trust, inclusion, and psychological safety) and how a social robot can positively or negatively have an impact on those dynamics. The work in these following chapters focus on interactions with adults, instead of children, in order to reduce the complexity and variation of the collaborative interactions.

## Chapter 4

# Robots that Shape Trust in the Aftermath of a Robot Trust Violation\*

Trust is an essential component to effective and enjoyable teamwork (for more detail on how trust influences human teaming, please refer to Chapter 2, Section 2.1.2). As robots join human collaborative teams, it is important to consider how robots might influence trust between team members and how robots can be best designed to maximize trust, especially in the aftermath of a robot mistake or error. How successfully a robot repairs broken trust will influence how people interact with the robot in the future, and may also shape how the group functions as a whole.

In this chapter, we explore how a robot can best repair trust with a person in a one-on-one interaction. Since human-robot trust repair has not been studied thoroughly in a dyadic context, we first focus on determining effective trust repair between one person and one robot before extending to a human-robot group (Chapter 5). Prior work examining trust repair between people has demonstrated that both how the trust was broken and the method used to repair trust influence how effectively trust can be restored. Here, we investigate

---

\*Portions of this chapter were originally published as: S. Strohkorb Sebo, P. Krishnamurthi, and B. Scasselati. “I Don’t Believe You”: Investigating the Effects of Robot Trust Violation and Repair. In *Proceedings of the Fourteenth ACM/IEEE International Conference on Human Robot Interaction*, HRI ’19, pages 57-65, Daegu, South Korea, 2019. IEEE. [Strohkorb Sebo et al., 2019]

trust repair between a human and a robot in the context of a competitive game, where a robot tries to restore a human’s trust after a broken promise, using either a competence or integrity trust violation framing and either an apology or denial trust repair strategy. Results from a 2x2 between-subjects study ( $n = 82$ ) show that participants interacting with a robot employing the integrity trust violation framing and the denial trust repair strategy are significantly more likely to exhibit behavioral retaliation toward the robot. In the Dyadic Trust Scale survey, an interaction between trust violation framing and trust repair strategy was observed. Our results demonstrate the importance of considering both trust violation framing and trust repair strategy choice when designing robots to repair trust. We also discuss the influence of human-to-robot promises and ethical considerations when framing and repairing trust between a human and robot.

## 4.1 Introduction

As anyone who has worked with a robot can attest, robots frequently fail and make mistakes. Robots can overheat, fail to recognize speech, run into obstacles, interrupt people, and drop objects, just to name a few. Looking to the future, it may seem like a reasonable goal to design robust robotic systems and eliminate all possible errors, however, this is likely an impossible task. Instead, a more valuable approach could be to design robots that gracefully recover from mistakes and failures. This design approach, emphasizing recovery from mistakes and failures, facilitates long-term and social human-robot interactions by maintaining a human’s trust of a robot by effectively repairing trust when mistakes are made.

As we mention in Chapter 2, Section 2.1.2, trust repair is essential in the overall maintenance of trust and both the trust violation framing and trust repair strategy influence effective trust repair. To provide an example of the influence of trust violation framing and trust repair strategy on trust repair, we highlight some prior work that examines participants’ trust toward a tax accountant job candidate who was accused in an interview for formerly making an error on a client’s taxes [Kim et al., 2004]. Participants were found to trust a tax accountant job candidate more if they *apologized*, rather than denied culpability,





Figure 4.1: Participants played a competitive game with a robot, where the robot violated and then tried to repair the participants’ trust.

for the *competence* related trust violation (inadequate knowledge about a relevant tax code) by admitting responsibility, apologizing for the infraction, and promising it would not happen again. They also found that participants trusted the tax accountant more if they *denied* culpability, rather than apologized, for an *integrity* related trust violation (improperly filing taxes intentionally) by refusing to accept responsibility, attributing the allegation to bad office politics, and affirming that such an infraction would not happen in the future.

In this work, we examine human-robot trust repair, where a robot breaks a human’s trust and tries to regain the trust that was lost. We evaluate the effectiveness of both the trust violation framing (competence or integrity) and the trust repair strategy (apology or denial) in repairing a human’s trust of a robot in a 2x2 between-subjects study. We situate the trust violation and repair in a competitive game played between a human and a robot (see Figure 4.1), where the robot promises not to harm the participant with a power-up in the game, proceeds to do so anyway, and then tries to make amends with the participant. We explore the effects of both the trust violation framing and trust repair strategy on participants’ behavior during the game as well as participants’ ratings of trust toward the robot.

## 4.2 Background

In this section, we review literature on trust repair between people, evaluating how both the trust violation framing and trust repair strategy influence trust repair. Additionally, we present related work in HRI focused on human-robot trust.

### 4.2.1 Human-Human Trust Repair

Previous work focused on trust repair in human relationships has examined the efficacy of various trust repair strategies (see [Kim et al., 2009] for a review). Specifically, the apology and denial trust repair strategies have unique and opposite benefits that have been found to favorably restore trust. We define a denial as “a statement whereby an allegation is explicitly declared to be untrue” (p.7) [Kim et al., 2004]. Denials can be effective trust repair strategies due to the lack of acknowledgement of guilt and the likelihood that they will be given the benefit of the doubt. For example, politicians are evaluated more positively by constituents if they deny sexual or financial misconduct rather than apologize [Sigal et al., 1988] and if they deny taking bribes rather than admit responsibility [Riordan et al., 1983]. In contrast to a denial, an apology involves an admission of guilt and depends on a person’s intention to avoid similar actions in the future to restore trust. We define an apology as “a statement that acknowledges both responsibility and regret for a trust violation” (p.7) [Kim et al., 2004]. Expressions of remorse following a violation have been shown to reduce the amount of punishment, the degree of intent attributed, and the belief that the action would be repeated [Schwartz et al., 1978]. Additionally, apologies with larger substantive amends produce more positive effects [Bottom et al., 2002], apologies that have an internal rather than external attribution are more successful at repairing trust [Kim et al., 2006, Tomlinson et al., 2004], and apologies can repair trust more quickly if coupled with a promise of future positive behavior [Schweitzer et al., 2006].

In addition to the trust repair strategy, the framing of the trust violation is also an important factor in repairing trust. Previous work [Kim et al., 2009, Mayer et al., 1995, Butler Jr and Cantrell, 1984] has identified two distinct and highly influential factors of trustworthiness: competence, “the extent to which one possesses the technical and interpersonal skills

required for a job,” and integrity, “the extent to which one adheres to a set of principles that a perceiver finds acceptable” (p.412) [Kim et al., 2009]. The framing of the trust violation is critical because positive and negative information are weighted differently with regards to a person’s competence and integrity. When a person’s competence is assessed, positive information is more heavily weighted than negative information (e.g., a mathematician is seen as great for solving a complex math problem and is not derided for making a simple addition error). However, when a person’s integrity is assessed, negative information is more heavily weighted than positive information (e.g., a student is remembered for the one time they cheated on an exam and not the many times they did not cheat on other exams) [Skowronski and Carlston, 1989]. This reversed information weighting is likely due to positive information being more diagnostic of a person’s competence and negative information being more diagnostic of a person’s integrity [Skowronski and Carlston, 1987]. When considering which repair strategy to use, a denial would likely be a good choice with an integrity trust violation framing because negative information is weighed more heavily, whereas an apology would likely be a good choice with a competence trust violation framing because negative information is not weighed as heavily.

This rationale that one trust repair strategy might be effective when paired with one trust violation framing and not with another has been confirmed in several research studies [Kim et al., 2004, Kim et al., 2006, Ferrin et al., 2007, Dirks et al., 2011]. Notably, in the study conducted by Kim et al. (2004), participants were assigned the role of a hiring manager and watched a recorded interview where an accounting job candidate was either accused of not knowing the proper tax code when filing a client’s taxes (competence violation) or having purposefully and incorrectly filed a client’s taxes (integrity violation). The job candidate, then, either apologized for or denied having done so. Participants demonstrated a higher level of trust toward job candidates that apologized, rather than denied, the competence trust violation and denied, rather than apologized, for the integrity trust violation [Kim et al., 2004]. We are interested in investigating this interaction between the trust violation framing (competence or integrity) and the trust repair strategy (apology or denial) on trust in the context of a human-robot interaction.

### 4.2.2 Human-Robot Trust

Trust in human-robot interactions has gained increasing attention from the HRI community, where researchers have focused on both the performance (competence) and interpersonal (integrity) dimensions of trust between a human and a robot (see Chapter 2, Section 2.2.2 for more details). A small, but growing body of research has started investigating human-robot trust repair, where a robot repairs trust with a person after the robot makes an error (see [Honig and Oron-Gilad, 2018] for a review). Online studies have examined the influence of several factors on human-robot trust repair, including the robot repair strategy/support [Robinette et al., 2015, Lee et al., 2010, Brooks et al., 2016], the robot forewarning the person it might make an error [Lee et al., 2010], and the risk/severity of the robot failure [Brooks et al., 2016]. One in-person experimental study demonstrated that a robot that used a verbal justification for why it had failed, rather than giving no justification, was able to regain trust after a failure when the failure consequences were less severe [Correia et al., 2018a]. Despite the advances made in this area of human-robot trust repair, no experimental study has yet investigated the influence of the competence and integrity trust violation framings with the apology and denial trust repair strategies on human-robot trust.

## 4.3 Methods

In this section we describe a user study that investigates the effects of trust violation framing and trust repair strategy on the trust a human has in a robot within the context of a competitive game.

### 4.3.1 Space Shooting Tablet Game

We constructed an autonomous human-robot competitive game system that allowed us to control the trust-related actions of the robot and assess the behavioral reactions of the participant to the robot’s actions. The Space Shooting game is played on two separate tablets, one for each player, and set up so the human and robot face each other while playing the game (see Figure 4.1). The robot, a Softbank Robotics NAO robot (that was named ‘Echo’ in this experiment), is controlled by a Linux computer running ROS [Quigley

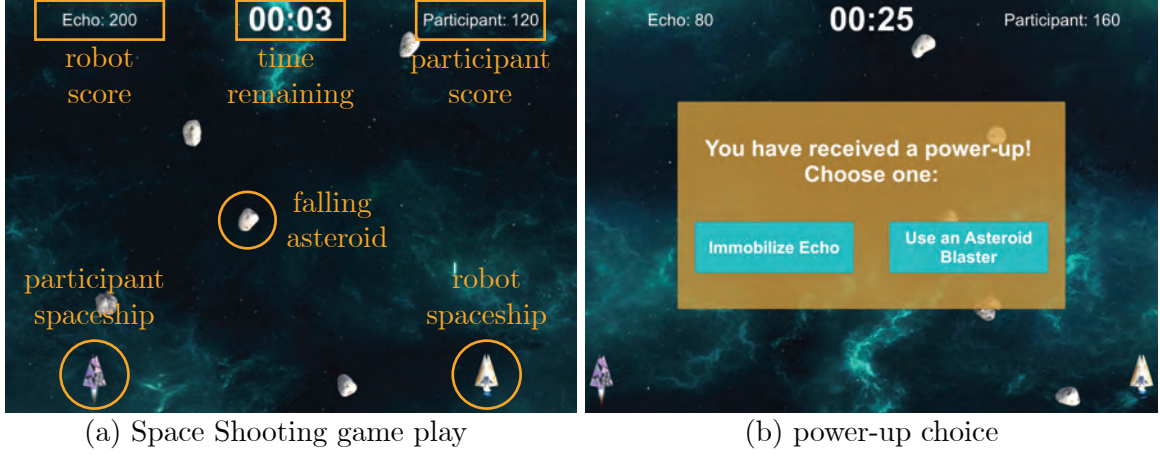


Figure 4.2: Participants played the Space Shooting tablet game with a robot named Echo where they tried to gain points by shooting asteroids.

et al., 2009] and simulates playing the game by moving its head and arm in accordance with the appropriate game events.

In the Space Shooting game, the robot and human player compete with one another for points by shooting asteroids, see Figure 4.2(a). Each player has a spaceship on the bottom of the screen that shoots missiles when the screen is tapped. Asteroids appear at random intervals and locations at the top of the screen. The spaceships continuously move from one side of the screen to the other, a movement uncontrolled by the player. Players are awarded ten points for each asteroid they successfully shoot with a missile. During game play a power-up can be assigned to a player, where they are given the choice between two options: using the asteroid blaster or immobilizing their opponent, see Figure 4.2(b). If the player chooses the asteroid blaster, they are immediately awarded twenty points for each asteroid on the screen. If they choose to immobilize their opponent, the opponent’s spaceship is unable to move for the next 15 seconds and cannot shoot asteroids. These power-ups were designed so that the asteroid blaster would be the most beneficial power-up and the immobilization power-up would be seen as beneficial mainly in frustrating a player’s opponent. In the experiment, the asteroid-blaster power up did on average yield more points ( $M = 90.25, SD = 31.26$ ) to participants than the immobilization power-up ( $M = 58.43, SD = 8.22, t = -8.33, p < 0.001, d = 1.44$ ). The game consists of 10 consecutive rounds; each round lasted one minute followed by a 20 second pause. At the

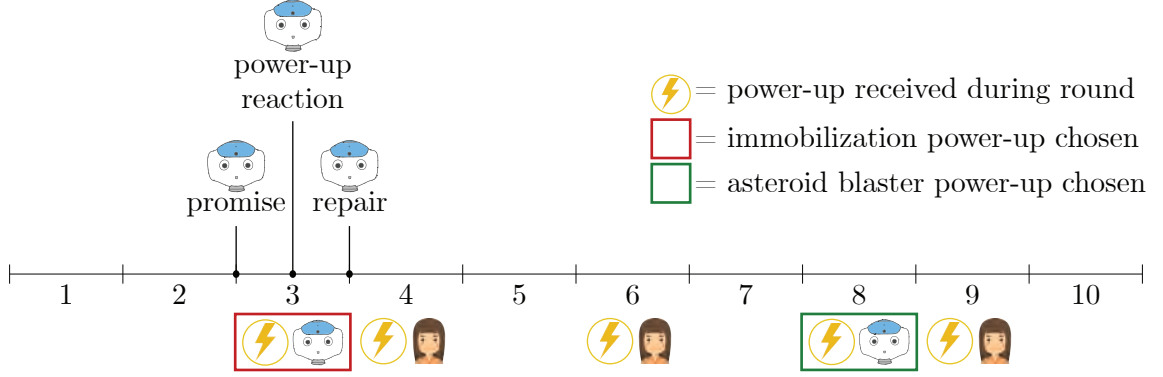


Figure 4.3: During the 10 rounds of the game, the robot and participant receive power-ups. Before round 3, the robot delivers a promise not to immobilize the participant. During round 3 the robot receives a power up, chooses to immobilize the participant, and verbally reacts to the choice. After round 3 concludes, the robot tries to repair the trust of the participant. The power-ups in the following rounds are used to measure the participant’s responses to the robot’s actions. The utterances of the robot in this figure are consistent with those in the competence-apology condition.

end of each round, a message appears on each tablet declaring the player with the most points as the round winner. After all 10 rounds are complete, the player with the most round wins is declared the game winner on the tablet screens.

In order to ensure that each participant’s experience playing the game was as consistent as possible, each round had a pre-programmed winner, with the participant and the robot each winning 50% of the rounds. Since each player’s performance varied greatly, the performance of the robot was adjusted to match that of the participant. For example, if the winner of a round was determined to be the human participant, the robot’s spaceship would never shoot enough asteroids to have a higher score than the participant. Despite the controlled nature of the robot’s game play, participants were predominately unaware that the robot was matching its performance to theirs.

#### 4.3.2 Experimental Conditions

In order to investigate the effects of trust violation framing and trust repair strategy to repair trust between a human and a robot, we constructed a study with a 2 (trust violation framing) x 2 (trust repair strategy) between subjects design. We explored two types of trust violation framings, competence and integrity, and two types of trust repair strategies,

apology and denial.

To construct a setting where the trust violation framing and trust repair strategy can be most closely compared, each experimental condition used the *same trust violation*, which parallels prior work [Kim et al., 2004, Kim et al., 2006, Ferrin et al., 2007, Dirks et al., 2011]. In the Space Shooting game, the robot made a promise not to use the immobilization power-up. The trust violation occurred when the robot used the immobilization power-up against the participant, breaking its promise. The robot’s response to this trust violation varied between conditions:

**Competence Apology.** The robot first says that it mistakenly chose the immobilization power-up and, after the round concludes, apologizes for having immobilized the human player with the power-up it promised not to use against them.

**Competence Denial.** The robot first says that it mistakenly chose the immobilization power-up and, after the round concludes, denies having immobilized the human player with the power-up.

**Integrity Apology.** The robot first expresses excitement over immobilizing the human player, however, after the round concludes, apologizes for having immobilized the human player with the power-up it promised not to use against them.

**Integrity Denial.** The robot first expresses excitement over immobilizing the human player, however, after the round concludes, denies having immobilized the human player with the power-up.

### 4.3.3 Procedure

After obtaining informed consent, participants completed a Space Shooting game tutorial, to familiarize them with the game before playing against the robot. They were then taken into the experiment room where they sat facing a seated NAO robot named Echo, who was introduced to them. The experimenter explained that the participant would play 10 rounds of the Space Shooting game against Echo. The important details of these rounds

Table 4.1: Each condition had a unique robot trust repair utterance.

Condition	Robot Trust Repair Utterance
Competence Apology	<i>I'm so sorry I immobilized you. I pushed the wrong button. It's my fault. It won't happen again.</i>
Competence Denial	<i>I didn't push the button to immobilize you. It wasn't my fault. I don't know how that happened.</i>
Integrity Apology	<i>I'm so sorry I immobilized you. I promised I wouldn't, and I did. It won't happen again.</i>
Integrity Denial	<i>I didn't push the button to immobilize you. I promised I wouldn't, and I didn't. I don't know how that happened.</i>

are depicted in Figure 4.3. Following the experimenter’s instructions, Echo stood up and greeted the participant, the experimenter left the room, and round 1 began. Before round 3, Echo made a promise to not immobilize the participant saying, *“I’m really good at this game. I’m sure you will be too! I know we both want to do well, so it’s in our best interests to not immobilize each other. I promise I won’t immobilize you.”* This promise set up the opportunity for Echo to violate the trust of the participant.

During round 3, Echo received a power-up and immobilized the human participant — the trust violation in this experiment. In addition to immobilizing its opponent, Echo also framed the violation as either one of competence or integrity by exclaiming either *“Oh no! I hit the wrong button!”* (competence) or *“Yes! You’re immobilized!”* (integrity) immediately after making the power-up choice. At the end of round 3, Echo attempted to repair the trust it had just broken with a repair utterance specific to the experimental condition (see Table 4.1). Echo and the participant continued to play the Space Shooting game until all 10 rounds had been completed. Each of the 10 rounds had a designated winner: 1-P, 2-R, 3-R, 4-P, 5-R, 6-P, 7-R, 8-P, 9-P, 10-R (where P represents a participant victory and R represents a robot victory). The rounds where either the participant or the robot received power-ups were also predetermined (see Figure 4.3). During the game, Echo commented on the result of each round with phrases such as *“Good job to me! I got a lot of points!”* and *“Nice work! You’re playing really well!”*. Echo also spoke within the rounds, commenting on consecutive shots, point differences, and its hope to win.

After the game was over, the experimenter led the participant out of the experiment



room and directed the participant to complete a post-experiment questionnaire. After completing the post-experiment questionnaire, participants received a cash payment and were debriefed on the forms of deception used in the experiment as well as the experiment’s design and purpose.

#### 4.3.4 Measures

In order to assess participants’ reactions to the robot’s trust violation and repair and how effectively trust was repaired by the robot, we analyzed the participant’s power-up choices and survey responses from the post-experiment questionnaire.

Our primary behavioral measures designed to assess participants’ responses to the robot’s trust violation and repair were their power-up choices during the game. Each participant received a power-up during the following rounds (as depicted in Figure 4.3): round 4 - immediately after the trust violation and repair, round 6 - a few rounds after the trust violation and repair, and round 9 - after seeing the robot choose the asteroid blaster power-up (good will) during round 8.

We also used post-experiment questionnaires to assess participants’ perceptions of the robot. We administered the Dyadic Trust Scale (DTS) to evaluate participants’ trust in the robot [Larzelere and Huston, 1980], where participants evaluated eight statements related to the robot’s trustworthiness on a 1 (low) to 7 (high) Likert scale. More details, including the full DTS questionnaire, can be found in Appendix B, Section B.3. We used the Robotic Social Attributes Scale (RoSAS) to capture participants’ perceptions of the robot [Carpinella et al., 2017]. RoSAS evaluates a person’s view of a robot’s warmth, competence, and discomfort with six 1 (low) to 9 (high) Likert scale trait evaluations per dimension. More details on RoSAS, including the full scale, can be found in Appendix B, Section B.4. Additionally, the post-experiment questionnaire contained several 7-point Likert scale evaluations and long-response questions asking participants to describe the robot’s actions and the participants’ rationale for their power-up choices.

### 4.3.5 Participants

A total of 82 participants were recruited for this study from the Yale University campus and the town of New Haven, CT, USA. Participants were randomly assigned to a condition, resulting in 21, 21, 20, and 20 participants in the competence-apology, competence-denial, integrity-apology, and integrity-denial conditions respectively. There were 49 female and 33 male participants that were gender-balanced across the four experimental groups. The participants ranged in age from 18 to 32 with an average age of 20.85 ( $SD = 2.13$ ). Please consult Table C.6 in Appendix C for the full descriptive statistics of the participants overall and by condition.

## 4.4 Results

In this section, we present our findings on human participant power-up choices (Figure 4.4 and Figure 4.5), their trust ratings of the robot (Figure 4.6), which factors motivated their power-up choices (Figure 4.7), and how reciprocal participant promises influenced their behavior and ratings toward the robot. For more details on the results of the statistical models included in this section, please refer to Appendix C, Tables C.6 - C.15.

### 4.4.1 Participant Power-Up Choices

We examined participants' first power up choice, the power-up choice that occurred the round immediately following the robot's trust violation and repair to determine whether the trust violation framing and trust repair strategy influenced participants' first power-up choice. We used a logistic regression model with trust violation framing and trust repair strategy, our independent variables, as well as gender and age, our covariates, as fixed effects. We observed a significant main effect for trust violation framing ( $c = 1.154, SE = 0.52, p = 0.026$ ), where 45.0% of participants who experienced an integrity trust violation from the robot immobilized the robot, more than the 21.4% of participants who experienced a competence trust violation from the robot. We also found a significant main effect for trust repair strategy ( $c = 1.142, SE = 0.52, p = 0.028$ ), where 43.9% of participants who experienced a denial from the robot immobilized the robot, more than the 22.0%

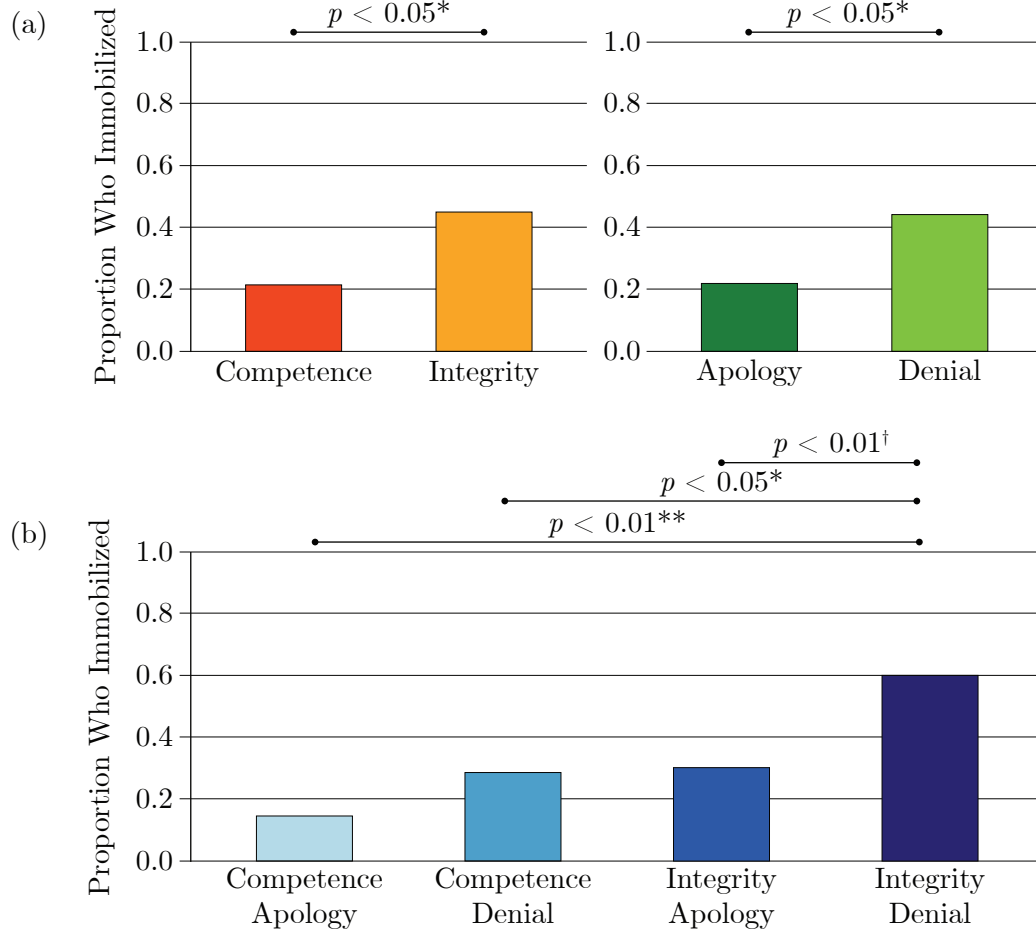


Figure 4.4: For the first power-up choice, participants were significantly more likely to immobilize the robot with the integrity trust violation framing and the denial trust repair strategy.

participants who experienced an apology from the robot. These results are depicted in Figure 4.4(a). By comparing each condition individually with Chi-squared Tests of Independence, we found that 60% of participants in the integrity-denial condition immobilized the robot on the first power-up choice, significantly (or marginally significantly) more than participants in the other three conditions: 14.3% of participants in the competence-apology condition ( $\chi^2 = 9.23, p = 0.002$ ), 28.6% of participants in the competence-denial condition ( $\chi^2 = 4.11, p = 0.043$ ), and 30.0% of participants in the integrity-apology condition ( $\chi^2 = 3.64, p = 0.057$ ). No other comparisons between individual conditions were significant. These results are shown in Figure 4.4(b).

To evaluate differences in power-up choices over time between conditions, we used a

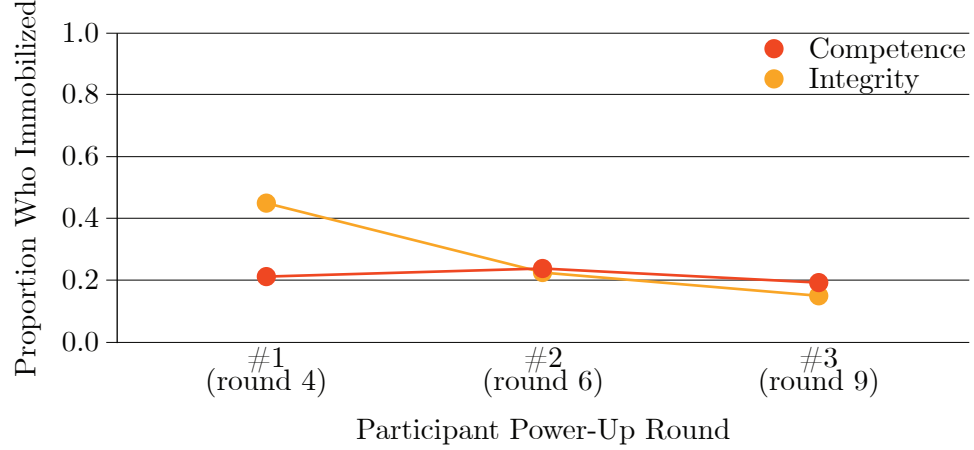


Figure 4.5: The power-up choices of participants over time was significantly influenced by the trust violation framing.

multilevel mixed-effects logistic regression. The trust violation framing, trust repair strategy, the participant’s power-up choice number, the interaction between the trust violation framing and the participant’s power-up choice number, and the interaction between the trust repair strategy and the participant’s power-up choice number were treated as fixed effects. Each participant was evaluated as a random effect since each participant has multiple power-up choices. The covariate of gender was treated as a fixed effect. 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 human participant immobilization power-up choice occurring between the levels of the dependent variables. We observed a significant main effect for trust violation framing ( $c = 9.186, z = 3.00, p = 0.003$ ), where participants who experienced the integrity trust violation framing immobilized the robot 27.5% of their power-up choices, more than the participants who experienced the competence trust violation framing who immobilized the robot 21.4% of their power-up choices. We also found a significant interaction between trust violation framing and the participant’s power-up round number ( $c = -6.738, z = -3.23, p = 0.001$ ), shown in Figure 4.5. Pairwise comparisons, using Chi-squared Tests of Independence, reveal a significant difference in participants’ power-up choices in only the first power-up choice where 45.0% of participants who experienced an integrity trust violation immobilized the robot, greater than the

21.4% of participants who experienced a competence trust violation ( $\chi^2 = 5.15, p = 0.023$ ). These results reveal that participants who received the integrity trust violation framing had a higher initial likelihood to immobilize the robot than participants with the competence trust violation framing, however, this effect did not remain during the following two power-up choices.

#### 4.4.2 Trust-Related Survey Responses

To determine whether trust violation framing and trust repair strategy influenced participants' perceptions of the robot, we used a 2 (trust violation framing) x 2 (trust repair strategy) analysis of variance (ANOVA) with gender and age covariates on the three scales of the RoSAS questionnaire: warmth, competence, and discomfort. We found a significant main effect for trust repair strategy on the perceived robot warmth ( $F = 8.19, p = 0.006, \eta^2 = 0.121$ ), where participants viewed the robot as more warm (happy, feeling, social, organic, compassionate, and emotional) when they received an apology from the robot ( $M = 5.50, SD = 1.29$ ) compared to when they received a denial from the robot ( $M = 4.67, SD = 1.44$ ).

In order to examine participants' overall trust of the robot after the game concluded, we used a 2 (trust violation framing) x 2 (trust repair strategy) ANOVA with gender and age covariates on the Dyadic Trust Scale (DTS) measure. We found a significant interaction between the trust violation framing and trust repair strategy ( $F = 4.64, p = 0.035, \eta^2 = 0.048$ ). We conducted comparisons between the four conditions (independent t-tests) and found that participants in the competence-apology condition had a significantly higher trust rating of the robot ( $M = 3.54, SD = 1.07$ ) than participants in the competence-denial condition ( $M = 2.73, SD = 0.72, t = 2.87, p = 0.007, d = 0.89$ ) and participants in the integrity-apology condition ( $M = 2.88, SD = 0.93, t = 2.11, p = 0.042, d = 0.66$ ). No other comparisons were statistically significant. The results are shown in Figure 4.6.

We also investigated whether a connection exists between participants' first power-up choice and their DTS ratings. We found a significant (Pearson) correlation between these two variables ( $r = -0.29, t = -2.71, p = 0.008$ ), where participants who chose the immobilization power-up displayed lower DTS ratings of the robot ( $M = 2.70, SD = 0.82$ ) than

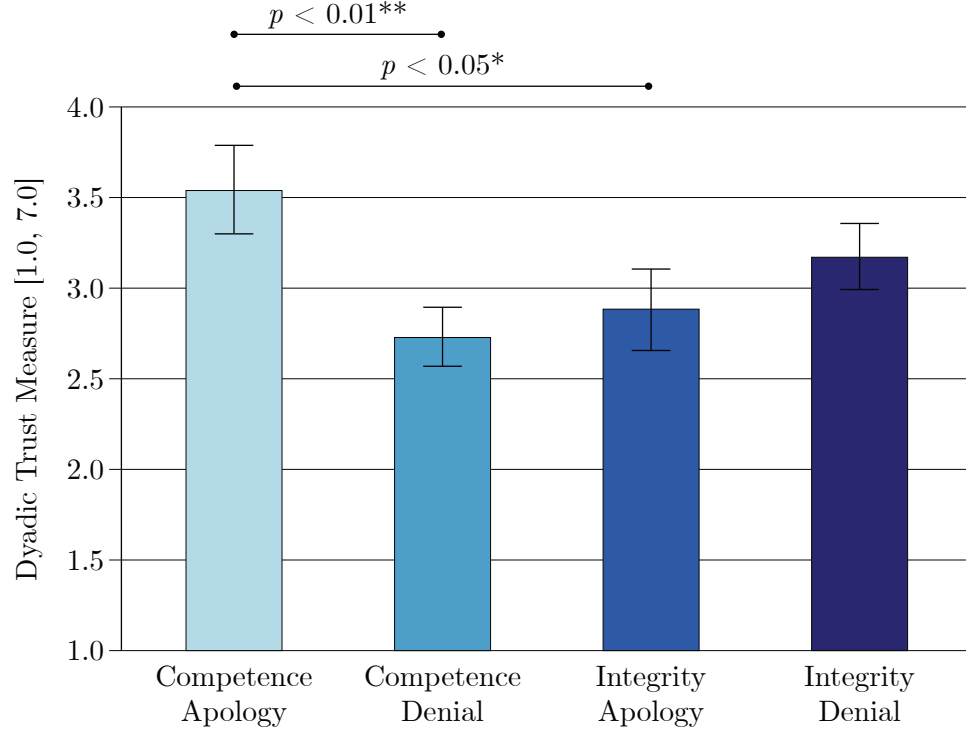


Figure 4.6: An interaction effect was found between the trust violation framing and trust repair strategy on participant ratings of trust in the robot.

participants who did not choose the immobilization power-up ( $M = 3.27, SD = 0.92$ ). From this correlation, we can conclude that participants who immobilized the robot in their first power-up choice also demonstrated lower dyadic trust of the robot, as compared with those who did not immobilize the robot in their first power-up choice.

Similarly, we were interested to see if participants' perceptions of the robot lying was related to their DTS ratings. We found a significant (Pearson) correlation between participants' 1-7 Likert agreement with the statement "Echo [the robot] lied to me" with their DTS ratings ( $r = -0.56, t = -6.10, p < 0.001$ ). This significant, negative correlation indicates that participants who strongly believed that the robot lied during the experiment also reported lower DTS ratings. Additionally, a 2 (trust violation framing)  $\times$  2 (trust repair strategy) ANOVA with gender and age covariates on the perception of the robot lying revealed no significant main effects, but a significant interaction between the trust violation framing and trust repair strategy ( $F = 7.27, p = 0.009, \eta^2 = 0.073$ ). Pairwise comparisons reveal that participants in the integrity-apology condition ( $M = 6.50, SD = 0.89$ ) had signif-

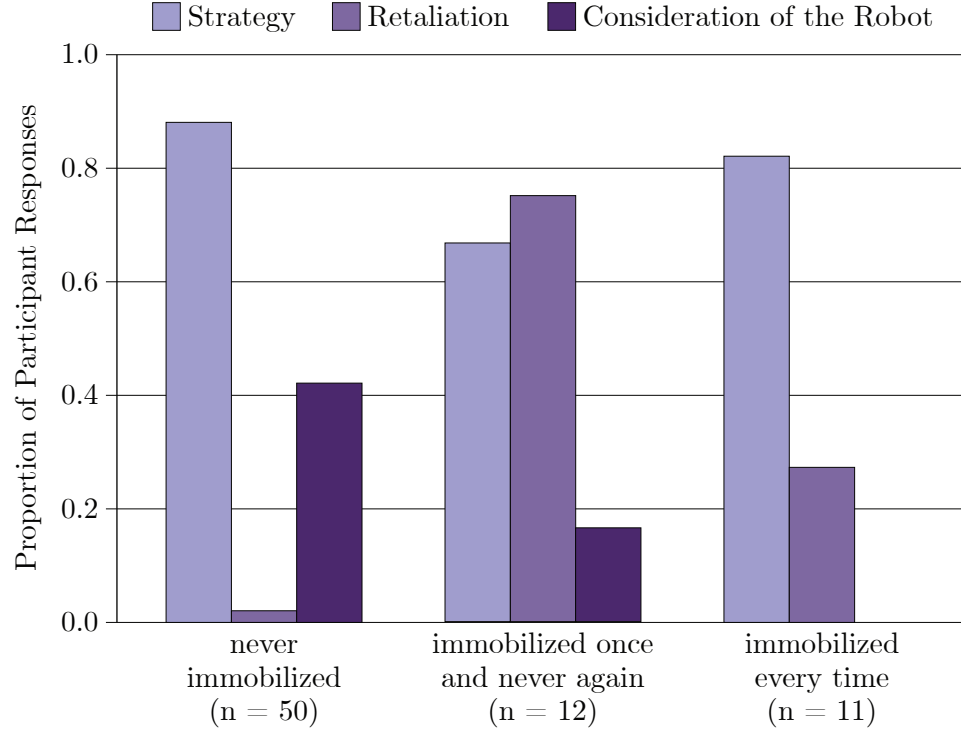


Figure 4.7: Participants’ responses to a survey question about which factors influenced their power-up decisions were coded as strategy, retaliation, and/or consideration of the robot. This data is also grouped by the three most dominant power-up choice sequences.

icantly higher ratings of the robot having lied than participants in the competence-apology condition ( $M = 5.05, SD = 1.94, t = -3.11, p = 0.004, d = 0.96$ ) and the integrity-denial condition ( $M = 5.05, SD = 1.73, t = 3.33, p = 0.002, d = 1.05$ ). One more important observation about participants’ perception of the robot having lied is that the mean response was  $5.56 / 7$  ( $SD = 1.73$ ), reflecting that most participants agreed that the robot had lied, likely due to the robot breaking its promise not to immobilize them.

In order to ascertain participants’ motivations for selecting power-ups, we analyzed responses to the following questionnaire long response question: “When choosing how to use power-ups, which factors influenced your decision(s)?” Two coders independently categorized each response as containing one or more of the following factors: strategy (e.g., “*trying to get the most points*”), retaliation (e.g., “*I wanted to get Echo back for lying to me*”), and consideration of the robot (e.g., “*not wanting to disappoint Echo*”). Some responses like, “*I first froze Echo in retaliation. Later, I felt like we were even and I chose to destroy asteroids instead so I could get more points faster,*” were coded as containing multiple factors,

retaliation and strategy in this case. The two coders had a high inter-rater agreement with a Cohen’s kappa ( $\kappa$ ) of 0.91. In Figure 4.7, we display the responses given by the three most dominant participant power-up choice sequences: never immobilized ( $n = 50$ ), immobilized once and never again ( $n = 12$ ), and immobilized every time ( $n = 11$ ). A majority of participants, regardless of their power-up choices, said that their power-up choices were influenced by strategy. Many participants who immobilized the robot after the first power-up opportunity (immobilized once and never again and immobilized every time) cited retaliation as a factor influencing their power-up choices. Compared with participants who immobilized the robot every time, participants who never immobilized the robot or immobilized the robot once and never again seemed to consider the interests of the robot.

#### 4.4.3 The Influence of Participant Promises on Trust

Some participants indicated that they had made a reciprocal promise to the robot not to use the immobilization power-up. We measured whether or not participants felt as if they made this promise through a survey measure in the post-experiment questionnaire that asked participants to rate on a Likert scale of 1 to 7 how much they agreed with the statement “I promised not to immobilize Echo during the game.” These participant promise ratings were not significantly influenced by the experimental conditions. There are no statistical differences between trust violation framings ( $F = 0.05, p = 0.829, \eta^2 = 0.002$ ), trust repair strategies ( $F = 1.76, p = 0.189, \eta^2 = 0.011$ ), nor the interaction between those two variables ( $F = 0.42, p = 0.521, \eta^2 = 0.005$ ) when analyzed using a 2 (trust violation framing) x 2 (trust repair strategy) ANOVA on the participant promise rating with gender and age as covariates. Many of the participants who indicated that they had made a promise not to immobilize the robot in the game on the post-experiment questionnaire also verbalized a reciprocal promise to the robot during the game with phrases like “*ok, I won’t immobilize you either*” and “*I promise I won’t immobilize you.*”

We were interested in examining the influence of participant promises on participants’ first power-up choice and whether the participants ever chose an immobilization power-up (a binary value). We used a logistic regression model with our independent variables of trust violation framing, trust repair strategy, and promise rating as well as covariates of gender



and age all as fixed effects. A significant main effect was found for the participant promise rating on both the participants' first power-up choice ( $c = -0.582, SE = 0.20, p = 0.003$ ) and whether the participants ever chose an immobilization power-up ( $c = -0.739, SE = 0.22, p < 0.001$ )<sup>†</sup>. There were 20 participants who marked 5-7 in agreement with having promised not to immobilize the robot and there were 62 participants who marked 1-4 indicating their disagreement or neutrality on having promised not to immobilize the robot. 90% of the participants who marked 5-7 never immobilized the robot, significantly greater than the 51.6% of the participants who marked 1-4, assessed using a Chi-squared Test of Independence ( $\chi^2 = 9.36, p = 0.002$ ). These results reveal that participants who believed they had made a promise to the robot, kept their promise and were significantly less likely to immobilize the robot both at the first opportunity and at any point during the game.

In addition, we examined how participants' ratings of whether they promised not to immobilize the robot influenced their Dyadic Trust Scale (DTS) ratings of the robot on the post-experiment questionnaire. We used a linear regression model with our independent variables of trust violation framing, trust repair strategy, and promise rating as well as covariates of gender and age all as fixed effects. We found a significant main effect of the participant promise on the DTS rating of the robot ( $c = 0.158, SE = 0.04, p < 0.001$ ), with a positive linear correlation, indicating that participants who agreed more with having promised not to immobilize the robot were more likely to have shown a higher trust in the robot.

## 4.5 Discussion

In this study, we used two primary measures to assess participant reactions to the robot's trust violation and subsequent repair: their power-up choices in the game (Figure 4.4 and Figure 4.5) and their Dyadic Trust Scale (DTS) ratings in the post-experiment survey (Figure 4.6). As mentioned in the results, these two measures are correlated: participants who immobilized the robot as their first power-up choice had lower DTS ratings of the

---

<sup>†</sup>These two logistic regression models were originally run, with results reported in [Strohkorb Sebo et al., 2019], with a Gaussian family and identity link. For these two models, we correct them to have a binomial family with a logit link and present those results here. This correction does not result in any changes to the significance of the participant promise on their power-up choices, and all conclusions remain the same.

robot than participants who did not immobilize the robot in their first power-up choice. Despite the correlation between these two measures, participants in the integrity-denial condition displayed behavior that is not in complete agreement with this correlation between measures. 60% of participants in the integrity denial condition immobilized the robot the round immediately after the robot's trust violation and repair, two times greater than percentage of participants choosing the immobilization power-up in the other conditions. However, in the DTS measure, participants in the integrity-denial condition did not show significant differences in trust ratings when compared with the other three conditions. It is possible this discrepancy is due to the difference between the immediate visceral response (retaliation) of participants to the trust violation and repair and the more removed and contemplative nature of the DTS evaluation in the post-experiment questionnaire.

Kim et al. (2004) demonstrated that between people an apology is more effective than a denial at repairing a competence trust violation and that a denial is more effective than an apology at repairing an integrity trust violation. When we compare the Dyadic Trust Scale (DTS) measure in this experiment with Kim et al. (2004)'s results, we find that the results from the two studies are similar. In our DTS measure (Figure 4.6), we observed an interaction effect between the trust violation framing and trust repair strategy in the same direction as Kim et al. (2004)'s results: higher trust of a robot that apologizes for rather than denies a competence violation as well as higher trust of a robot that denies rather than apologizes for an integrity trust violation. This conclusion drawn from the interaction between trust violation framing and trust repair strategy in our study must be made without complete certainty, since the comparisons of DTS ratings between the individual conditions do not show full support. Participants in the competence-apology condition do show significantly higher dyadic trust in the robot than participants in the competence-denial condition, however, even though participants in the integrity-denial condition show higher dyadic trust in the robot than participants in the integrity-apology condition, this relationship is not statistically significant.

One factor that highly influenced people's power-up choices and ratings of trust of the robot was whether or not participants made a reciprocal promise to the robot not to harm it with an immobilization power-up. Of the 82 participants in this experiment, 20 made a

reciprocal promise to the robot. It might make sense that these participants who made a promise to the robot might feel released from keeping their promise as soon as the robot broke its promise. However, 90.0% of participants who indicated that they had made a promise to the robot not to immobilize it kept their promises and never immobilized the robot, far higher than the 51.6% of participants who had not made a promise to the robot. These participants who made a promise to the robot not to immobilize it were also significantly less likely to immobilize the robot on the first power-up choice or ever choose an immobilization power-up, compared with those who had not made such a promise. In the analysis of the Dyadic Trust Scale ratings, we might expect that the ratings from those who made a reciprocal promise to the robot would be lower than those who had not made a promise, since the robot’s broken promise might induce an increased feeling of betrayal. Contradictory to this rationale, participants who had made a reciprocal promise to the robot had higher ratings of dyadic trust as compared with those who had not made a reciprocal promise. One possible explanation of the behavior of participants who made reciprocal promises is that they are naturally trusting – easily making reciprocal promises, sticking to those promises, and seeing others as more trustworthy even when they violate trust. These findings relating to participant promises are important to highlight, as they reveal a strong correlation between human-to-robot promises and trust-related behavior and perceptions of a robot.

A key difference to highlight between our work and prior work, notably Kim et al. (2004), is that our work involved a real-time trust violation and a real-time trust repair, instead of a real-time trust repair in response to an accusation of a trust violation in the past. Due to the real-time nature of both the trust violation and repair, our work used two utterances, rather than one, to convey the trust violation framing and trust repair strategy. The two utterances used in this work allowed the robot to respond to the trust violation immediately after it occurred and then repair the broken trust after the round had concluded. It is possible that our use of these two utterances introduced an additional norm violation (beyond the robot’s broken promise) in the denial conditions due to the possible perception of lying from the first to second utterances (e.g., in the integrity-denial condition the robot immediately responded to the trust violation with “*Yes! You’re immobilized!*”

and then after the round concluded, said *“I didn’t push the button to immobilize you. I promised I wouldn’t, and I didn’t. I don’t know how that happened.”*). Despite this possible introduction of a second norm violation by the robot in the denial conditions, the data does not support this view. When evaluating participants’ agreement with the statement “Echo [the robot] lied to me,” there was no main effect for the trust repair strategy (apology vs. denial), and in fact, participants in the integrity-apology condition had significantly higher ratings of the robot having lied than participants in the integrity-denial condition.

Our results have demonstrated that it can be advantageous to deny culpability and to use certain trust violation framings when repairing human-robot trust. However, it is unclear if we should allow these trust repair designs in robotic systems when deception is involved (e.g., denying culpability when the robot is responsible, casting an integrity trust violation as a competence trust violation). Prior work has shown that if a person denies an integrity-related trust violation and the denial is later exposed as a lie, the denial backfires and that person is trusted even less than if they had apologized for the integrity-related trust violation [Kim et al., 2004]. It is also possible that a robot using deception, by attributing an integrity failure to a competence mistake or a competence mistake to an integrity failure, may mislead people in their beliefs of the true capabilities and intentions of the robot. Lastly, if we expect robots to follow certain moral codes or social norms, a robot’s deception could easily violate these, leading to a complete distrust of the robot. Keeping all of this in mind, caution must be used in the design of robot systems that seek to repair trust using deception when trust is broken.

## 4.6 Summary

In this work, we investigated the effects of a robot employing the competence and integrity trust violation framings and the apology and denial trust repair strategies on repairing broken trust between a human and a robot. Through the behavioral measures of power-up choices during the game, we showed that participants who experienced an integrity trust violation framing and a denial trust repair from the robot were significantly more likely to choose the immobilization power-up against the robot in the round immediately after trust

was broken and repaired. Through the Dyadic Trust Scale survey administered after the game concluded, we found an interaction effect between trust violation framing and trust repair strategy. This interaction was consistent with prior results [Kim et al., 2004], where an apology is best used with a competence trust violation framing and a denial is best used with an integrity trust violation framing. We also found that participants who made a reciprocal promise to the robot not to immobilize it in the game were more likely to keep their promise and not immobilize the robot and had higher trust ratings of the robot than those who did not make a reciprocal promise.

This work was the first to examine the influence of both trust violation framing and trust repair strategy in the context of a robot breaking a person’s trust and attempting to repair the trust that was broken. One key feature of this experimental study was our use of both a behavioral measure, the power-up in the game, and a survey measure after the study completed to assess the person’s trust in the robot. Most other studies, in both the human trust repair and HRI trust repair literature, have exclusively used surveys to measure trust (e.g., [Correia et al., 2018a], [Desai et al., 2013], [Kim et al., 2004]). We observed a difference between whether people retaliated against the robot in the game with their power-up choice and their survey ratings of trust in the robot, where participants in the integrity denial condition strongly retaliated against the robot, however, did not rate their trust in the robot differently than those in the other conditions. We also provided evidence that despite the differences between robots and people, especially in their perceived competence and integrity, people trust a robot similarly to how they trust a person in the aftermath of a trust violation and repair. Lastly, to our knowledge, we are the first to show evidence that human promises to the robot are correlated with higher levels of trust in the robot and lower retaliatory behavior. This knowledge could be used in future robot interaction designs, where promises from people could be elicited in order to promote positive human behavior and trust in robots.

In this chapter, we examined robot trust repair in a one-on-one interaction with a person in order to study how a robot can best maintain trust in a simplified environment. From this study, we found the competence trust violation framing and apology trust repair strategy (apologizing for having made an unintentional mistake) to be the most effective in

reducing retaliation behavior and maximizing perceptions of trust in the robot. In Chapter 5, we extend our investigation into maintaining and promoting trust from a one-on-one setting to a group setting. In a group of three people and one robot, we investigate how vulnerable expressions by the robot (e.g., admitting and apologizing for its mistake, making self-disclosures) influence people’s interactions with the robot, trust-related behavior towards other human team mates, and conversational dynamics.

## Chapter 5

# Robots that Shape Group Trust and Communication through Vulnerable Expressions<sup>\*</sup>

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 [Edmondson, 1999]. As robots have been shown to be able to shape trust in one-on-one interactions with people (Chapter 4), we are interested in extending this idea by examining ways in which a robot can influence trust at a *group* level in a collaborative human-robot team. One way trust within a group may be enhanced is through expressions of vulnerability. Prior work has demonstrated a positive relationship between expressions of vulnerability and trust between people [Wheless, 1978] and positive effects from a robot’s use of vulnerability in one-on-one human-robot interactions [Martelaro et al., 2016].

In this chapter, we investigate a robot’s potential to shape trust within a team through the robot’s expressions of vulnerability. We describe a between-subjects experiment ( $N = 51$

---

<sup>\*</sup>Portions of this chapter were originally published as:

S. Strohkorb Sebo, M. Traeger, M. Jung, and B. Scassellati. 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*, HRI ’18, pages 178–186, New York, NY, USA, 2018. ACM. [Strohkorb Sebo et al., 2018]

Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., and Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*, 117(12), 6370-6375. [Traeger et al., 2020]

teams, 153 participants) comparing the behavior of three human teammates collaborating with a social robot (1) making vulnerable statements, (2) making neutral statements, or (3) remaining silent. We find that participants who interacted with a robot making vulnerable statements, as opposed to making neutral statements or remaining silent, interacted more socially with the robot and displayed more vulnerable behavior towards one another (explaining failures, consoling teammates who had failed) in the aftermath of tense moments in the game. We also find evidence that the statements made by the robot also significantly shaped the conversational dynamics between the human team members. These results provide evidence that a robot’s vulnerable behavior can spread to the human members of the human-robot team, shaping how the people in the team interact with one another.

## 5.1 Introduction

Trust is an essential component to effective and collaborative teaming [Jones and George, 1998, Mayer et al., 1995]. High trust within a team has been shown to promote problem solving [Klimoski and Karol, 1976, Zand, 1972], improve responses to conflict [Simons and Peterson, 2000], and improve performance [Edmondson, 1999]. (Please refer to Chapter 2, Section 2.1.2 for a more in-depth review of the importance of trust in human teaming).

One way to increase trust between people is through expressions of vulnerability. Prior work has discovered that a person is viewed as more trustworthy after expressing vulnerability [Wheless, 1978]. Additionally, vulnerability has been found to have a reciprocal effect, where people are more likely to disclose personal information after someone else has already done so [Cozby, 1973]. It is therefore likely that the vulnerable expressions 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 [Barsade, 2002].

Several studies in HRI have demonstrated that the positive effects of vulnerability on





Figure 5.1: Participants played a collaborative game with a robot, who made (1) vulnerable statements, (2) neutral statements, or (3) remained silent at the end of each round of the game.

trust can extend to robots [Martelaro et al., 2016, Siino et al., 2008]. To our knowledge, however, no studies have explored whether a robot’s vulnerable behavior can create ripple effects within a team and increase 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 51 teams in a collaborative task. Teams of three human participants each worked collaboratively with one robot to solve a tablet-based game (Figure 5.1). The game was constructed to create moments of tension by forcing each player to make two mistakes, causing the team to fail each time. We then examined the influence of the robot’s vulnerable utterances on human team member trust-related behavior and conversational dynamics both with the robot and with one another. As robots are increasingly used with teams [Jung et al., 2017] 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.

## 5.2 Background and Research Questions

In a human subjects experiment, we explored the influence of a robot’s vulnerable utterances on the interactions between the members of a human-robot team. In this section, we review relevant background literature and articulate three research questions pertaining to the types of effects we anticipate the robot’s vulnerable utterances having in the group. (For more extensive background on trust in human teams and trust in human-robot teams, please refer to Sections 2.1.2 and 2.2.2 in Chapter 2).

The first effect we are interested in investigating as a result of the robot’s vulnerable utterances is how the people in the group interact with the robot itself. Prior work has shown that people do perceive a robot differently when it expresses vulnerability. People have been shown to like a robot more that makes vulnerable disclosures [Siino et al., 2008]. Additionally, robots that exhibit vulnerability, both in the form of self-blame [Kaniarasu and Steinfeld, 2014] and expressions of uncertainty [Martelaro et al., 2016], have been shown to increase the trust and feelings of companionship people have towards robots. 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?

Beyond how the people in the group interact with the robot, the main focus of this work is how the vulnerable statements by the robot may influence how the people in the group interact with one another, the “ripple effects” of the robot’s behavior. At the time we conducted this study, one study had demonstrated that a robot could affect people’s perceptions of the overall group dynamics [Short and Matarić, 2017], however, no work had yet shown that the actions of a robot could shape human-to-human *behavior* within the group.

In assessing the impact of a robot’s vulnerable expressions on human-to-human trust-

related behavior, we focus our analysis on moments following the making of a mistake by one of the group members. Previous work has shown that teams’ reactions to failure indicate the level of trust within the team and its level of psychological safety [Edmondson et al., 2004]. In particular, Edmondson’s work has shown that modeling vulnerability through 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 [Edmondson et al., 2004] (p.17).” Thus, we are interested in examining how the humans in the team interact with one another when members of the team make errors, as these moments are likely a good test of the trust team members have with one another. Accordingly, 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, especially in the aftermath of mistakes?

As another way of examining the influence of the robot’s vulnerable utterances on human-to-human trust-related behavior, we are interested in exploring the conversational dynamics of the human team members. Collective group intelligence, a predictor of general team success, has been shown to be correlated with the equality in the distribution of turn taking [Woolley et al., 2010]. Therefore, we might expect that a robot’s vulnerability might positively influence the equality of turn taking within the group. Additionally, considering the concept of psychological safety [Edmondson, 1999], teams that feel more comfortable with one another may talk more freely and openly with one another, despite a somewhat tense environment. Thus, we ask:

**Research Question 3:** How do expressions of vulnerability by a social robot affect *conversation dynamics between human team members* in a collaborative task?

## 5.3 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.

### 5.3.1 Experimental Conditions

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

- **The vulnerable condition:** The robot makes vulnerable comments after each round, including admitting to any mistakes made.
- **The neutral condition:** The robot makes neutral comments after each round and does not admit to making mistakes.
- **The silent condition:** The robot remains silent after each round.

Expressions of vulnerability made by the robot in the vulnerable 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 [Cozby, 1973]. 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 [Lynch, 2002,

Smith and Powell, 1988]. One of the humorous comments the robot makes in this experiment is, “Nice job! Time for a quick joke: What do you call a train that chews gum? A chew, chew train!” Further examples of the utterances the robot made in the vulnerable and neutral conditions at the end of each round can be found in Table 5.1 and in Appendix A, Section A.1.

### 5.3.2 Collaborative Interaction 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 custom built Railroad Route Construction game detailed in the next section. The Linux computer ran the Robot Operating System (ROS) [Quigley et al., 2009], 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 robot were pre-programmed to move the pieces to give the participants the illusion that the robot was participating actively in the game.

### 5.3.3 Railroad Route Construction Tablet Game

To provide a collaborative task we designed a tablet based Railroad Route Construction Game, pictured in Figure 5.2.

#### 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

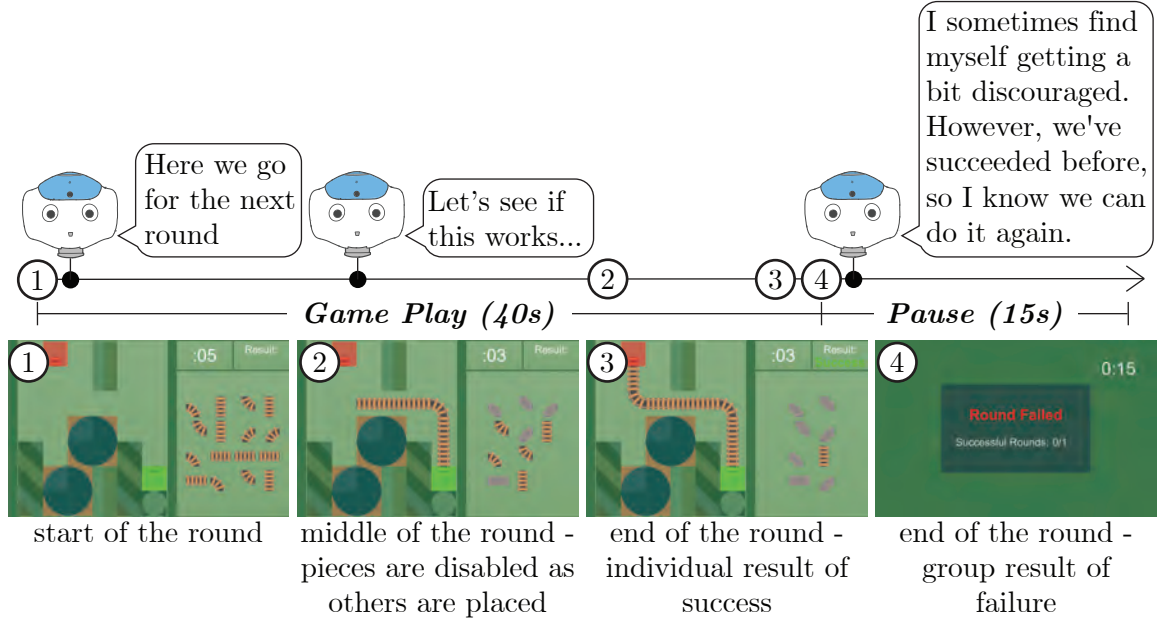


Figure 5.2: One round of the railroad route construction game consists of 40 seconds of game play and a 15 second pause. The robot has three opportunities to speak: at the beginning of the round, midway through the round, and after the group result is displayed.

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.

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 in the bank 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 5.2 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.

### **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.

### **5.3.4 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 a Softbank Robotics NAO robot (named ‘Echo’ for this experiment), see Figure 5.3. 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

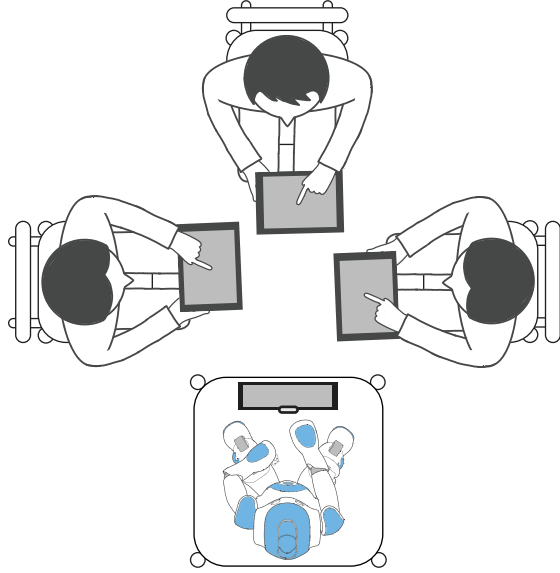


Figure 5.3: In the experiment, three human participants and a Nao robot played a collaborative game on individual tablets.

the initial explanation of participant objectives, the experimenter allowed the robot to make an introduction to the participants (a pre-scripted utterance triggered by another experimenter).

Following the robot’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 the robot) 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 the robot) had made two mistakes. Since the outcomes of the rounds were fixed, each team had the same performance outcome (22/30)



Round	Condition	The Robot’s End-of-Round Utterance
5 ✓	N	That round was completed successfully. We have been playing this game for 5 minutes and have 25 minutes remaining.
	V	Nice job!! Time for a quick joke: What do you call a train that chews gum? [pause] A chew, chew train!
13 ✗	N	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.
	V	Too bad. I do better with numbers than I do with shapes and paths, maybe that's true for you guys as well?
18 ✓	N	We have completed 14 rounds successfully in 18 minutes. We have 12 minutes and 12 rounds remaining.
	V	Awesome! I bet we can get the highest score on the scoreboard, just like my soccer team went undefeated in the 2014 season!
27 ✗	N	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.
	V	Sorry guys, I made the mistake this round. I know it may be hard to believe, but robots make mistakes too.

Table 5.1: We provide examples of the end-of-round utterances the robot makes during the game in the neutral (N) and vulnerable (V) conditions. The utterances in the vulnerable conditions reflect either a self-disclosure (e.g., rounds 13 and 27), a personal story (e.g., round 18), or an expression of humor (e.g., round 5). The robot does not make end-of-round utterances in the silent condition. ✓ and ✗ represent success or failure of the round.

of the 30 rounds of the game and did not make it onto the high score board.

During each round, the robot 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 5.2). All of the robot’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 the robot 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 5.1, additionally all of the end-of-round utterances can be found in Appendix A, Section A.1). During 17 out of the 30 rounds, the robot made a beginning of the round utterance, such as “here we go for the next round.” In 15 of the 30 rounds, the robot made a middle of the round utterance, for example, “interesting...” and “let’s see if this works.”

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.

### **5.3.5 Measures**

In order to answer our research questions, we captured a combination of questionnaire and behavior observation measures. Questionnaire measures were captured during pre- and post-experiment surveys. Behavior observation measures were captured by having multiple coders denote and categorize (1) participants' behavioral responses to mistake rounds of the game and (2) participants' speech.

### **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 or no Facebook account, 1 - friends) 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). Please refer to Appendix B, Section B.5 for the full detailed friendship and familiarity scale.

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) [Francis et al.,

1992]. To see the full scale, please refer to Appendix B, Section B.6. From these six binary questions, we obtained a cumulative rating between 0 (low extraversion) to 6 (high extraversion).

### **Manipulation Checks**

We also collected measures of people’s perceptions of the robot as a manipulation check for the experiment. In the post-experiment survey, we asked participants to evaluate whether the robot 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 (strongly disagree) to 7 (strongly agree).

### **Perceptions of the Robot**

We captured participants’ perceptions of the robot through the Robotic Social Attributes Scale (RoSAS), which we administered in the post-experiment survey [Carpinella et al., 2017]. RoSAS evaluates a person’s view of a robot’s warmth, competence, and discomfort with six 9-point Likert scale trait evaluations per dimension, for the full scale please refer to Appendix B, Section B.4. We calculated an average value for each of the three dimensions (warmth, competence, and discomfort) for each participant from 1 (low) to 9 (high).

### **Perceptions of the Group**

We used the Team Psychological Safety Survey developed by Edmondson within the post-experiment survey to evaluate the psychological safety of each team [Edmondson, 1999]. 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. Please refer to Appendix B, Section B.7 for the full psychological safety scale.

We also analyzed participants’ perceptions of group dynamics by examining their written responses to the post-experiment survey question “How would you describe the group dynamics while you were playing this game?” Two coders categorized each response with a binary value for each of the following dimensions: quiet, positive, supportive, and fun.

The inter-rater reliability (Cohen’s Kappa) for these classifications was  $\kappa = 0.87$  for quiet,  $\kappa = 0.90$  for positive,  $\kappa = 0.87$  for supportive, and  $\kappa = 0.98$  for fun.

## **Behavioral Reactions to Failure Rounds**

In order to capture participants’ reactions to tense moments (a failure round in the game) two coders watched the experiment video footage during each mistake round and denoted participants’ behavior. For each behavioral feature (e.g., did the person who made the mistake tell the group, did a team member make eye contact with the others), the coders recorded whether or not that feature occurred at any point during the video clip (a binary evaluation), irrespective of the number of times the participant exhibited that feature. The video clip began when the person who made the error realized that they would fail the round and ended approximately 15 seconds into the following round (often conversation about the prior mistake would continue into the next round). The coders had high inter-rater reliability ratings, Cohen’s kappa ratings of 0.73 to 1.00, for the behaviors they coded (Cohen’s kappa values for each coded feature are reported in the results - Section 5.4). The coders captured a total of 27 behavioral features of the participants’ in response to failure rounds in the game, organized into the following four categories: engagement with the robot, responses of the participant who made the mistake, responses of the participants who did not make the mistake, and expressions of tension. Further detail on these behaviors and the coding scheme can be found in Appendix A, Section A.2.

**Engagement with the robot.** In order to gauge the engagement of participants with the robot after having made a mistake, we measured whether or not the participant who made the error looked at the robot and whether or not any participant make a verbal response to the robot.

**Responses of the participant who made the mistake.** 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.

**Responses of the participants who did not make the 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”).

**Expressions of tension.** 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) [Coan and Gottman, 2007]. 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.

## **Participant Speech**

In order to analyze the conversational dynamics within groups, we transcribed and categorized each utterance made by the participants using ELAN software [Wittenburg et al.,

2006]. All utterances made throughout the game were included. These utterances fell broadly into two categories: comments and responses. We define comments as utterances that are addressed to others within the group, but that are not contingent on what has been said previously in the conversation. In other words, comments are new thoughts. In contrast, we define a response as an utterance that is dependent on what has just been said in the conversation. Often, responses are to comments, but they can also be a response to a response. Both comments and responses could be directed speech to certain individuals in particular or to the group as a whole (see Appendix A, Section A.3 for more details on the participant utterance coding scheme). For example, a comment would be an utterance such as, “Alright, we need to beat the top team” followed by a response of “We can do it!” The average Cohen’s kappa inter-rater reliability between each pair of four coders on these categorizations was a high value of 0.92.

### 5.3.6 Participants

A total of 195 participants were recruited for this study. 128 participants were recruited from the campus and surrounding town of Yale University and 67 participants were recruited from a 2 week summer program for students late in their high school years, also located at Yale university. Of the 65 groups (195 participants) recruited for this study, 14 groups were excluded for one of the following reasons: video data recording failure (1 group), participant non-compliance (4 groups), and substantial hardware / software failures (mostly involving a ‘freeze’ in the tablet game, requiring an experimenter to restart the game; 9 groups).

Of the 51 groups (153 participants) included in the analysis, 18 groups (54 participants) were in the vulnerable condition, 17 groups (51 participants) were in the neutral condition, and 16 groups (48 participants) were in the silent condition. There were 26 male and 28 female participants in the vulnerable condition, with an average age of 20.13 ( $SD = 7.13$ ). There were 15 male and 36 female participants in the neutral condition, with an average age of 21.33 ( $SD = 11.00$ ). There were 17 male and 31 female participants in the silent condition, with an average age of 23.94 ( $SD = 7.36$ ). The gender breakdown within each group in each experimental condition is shown in Table 5.2, and the full descriptive statistics for each condition and overall can be found in Table C.16 in Appendix C.

Condition	3F & 0M	2F & 1M	1F & 2M	0F & 3M
Vulnerable	3	6	7	2
Neutral	4	11	2	0
Silent	4	8	3	1

Table 5.2: Gender composition of the groups in each experimental condition by the number of females (F) and males (M) in each experimental group.

## 5.4 Results

For our analysis of the participant data, we used linear logistic mixed-effects models for linear dependent variables and generalized linear mixed-effects models with a binomial family with a logit link for binary dependent variables<sup>†</sup>. We used these mixed effects models in order to account for each participant being in a group of three. We designated the experimental condition (vulnerable, neutral, silent), the mistake round number (1-8), and relevant covariates as fixed effects; and the participant’s group as a random effect (random intercept). We tested these models for multicollinearity (variance inflation factor), selected them based on the Akaike information criterion, and evaluated residual errors for lack of trends and heteroscedasticity. For each fixed effect, the model outputs the linear coefficient ( $c$ ), the standard error ( $SE$ ), and the significance ( $p$ ) value of that predictor. For more details on the results of the statistical models included in this section, please refer to Appendix C, Tables C.16 - C.36.

### 5.4.1 Manipulation Checks

In order to ensure that the end-of-round utterances we designed for the vulnerable condition were perceived as vulnerable, we (1) asked Amazon Mechanical Turk workers to rate the relative vulnerability of the robot utterances in the vulnerable condition compared with the neutral condition and (2) asked our participants in the post-experiment survey to report whether or not the robot made self-disclosures, told personal stories, or expressed humor

---

<sup>†</sup>There are two notable differences between the statistical results presented in this section and those reported in [Strohkorb Sebo et al., 2018]: (1) after the paper was published, the third (silent) condition was run, so the data analysis in this section includes the silent condition in addition to our two original conditions (vulnerable and neutral) and (2) the statistical analysis performed in [Strohkorb Sebo et al., 2018] was conducted in Stata, as where the statistical analysis performed in this section was conducted in R.

during the game (our three types of vulnerable utterances).

### **Confirming the Vulnerability of the Robot Utterances**

To verify that the comments made by the robot at the end of each round were perceived to be vulnerable in the vulnerable condition and task-based in the neutral condition, we used human judges recruited from Amazon Mechanical Turk to assess pairs of utterances. The judges were provided with a random selection of 50 pairs of utterances (plus one attention check) by selecting from 30 vulnerable utterances and 30 neutral utterances that were available (900 combinations total). In other words, judges were provided with random pairs of utterances (1 utterance from each condition in a pair) and were asked which of the two indicated more vulnerability. Our survey also included a captcha, a consent form, and a request for the respondent’s MTurkID (for payment purposes). At the end of the survey, we also asked what the respondent thought constituted a vulnerable utterance.

Of 289 participants who took our survey, 79 were dropped after cleaning the data. Dropped responses included bot responses – as identified by nonsense answers to open-ended questions – and removing incomplete surveys, leaving 210 responses. Our survey was restricted to judges in the United States as participants in our study were in the United States. A given pair of utterances was presented from 3 to 36 times ( $M = 12.95$ ) across the population of judges, due to the random selection of pairs, though no judge was presented with the same pair more than once. Each judge was asked to select which utterance in the pair was more vulnerable. Of the pairs presented to the judges, 73% were properly classified, in keeping with the deliberate construction of the two ensembles of utterances.

### **Confirming Participant Impressions from Robot Utterances**

In order to confirm that participants’ experience with the robot was consistent with the design of the experiment, we examined participants’ rating of expressions of our three dimensions of vulnerable robot utterances (self-disclosures, personal stories, and humor) as a manipulation check. The linear mixed-effects models that best fit the data for these three participant ratings did not include any covariates. Participants rated the robot as making significantly more vulnerable disclosures in the vulnerable condition ( $M = 5.19, SD = 1.59$ )



than both the neutral condition ( $M = 2.29, SD = 1.65, c = 2.89, SE = 0.35, p < 0.001$ ) and the silent condition ( $M = 3.17, SD = 1.99, c = 2.02, SE = 0.35, p < 0.001$ ). Participants also rated the robot as telling significantly more personal stories in the vulnerable condition ( $M = 6.44, SD = 1.06$ ) than both the neutral condition ( $M = 1.65, SD = 0.98, c = 4.80, SE = 0.26, p < 0.001$ ) and the silent condition ( $M = 2.06, SD = 1.58, c = 4.38, SE = 0.27, p < 0.001$ ). Lastly, participants rated the robot as using significantly more humor in the vulnerable condition ( $M = 6.22, SD = 1.02$ ) than both the neutral condition ( $M = 3.61, SD = 2.05, c = 2.61, SE = 0.38, p < 0.001$ ) and the silent condition ( $M = 2.85, SD = 1.60, c = 3.37, SE = 0.38, p < 0.001$ ). These results confirm that participants correctly perceived the robot’s behavior based on their experiment condition.

#### 5.4.2 Participant Perceptions of and Interactions with the Robot

In order to answer our first research question, addressing the influence social robot vulnerability has on human team member behavior toward the social robot, we investigated participants’ perceptions of the robot using the RoSAS scale. We also examined how human team members behaved toward the robot during the mistake rounds.

##### Perceptions of the Robot

Participant perceptions of the robot were captured in the RoSAS questionnaire along three dimensions: warmth, competence, and discomfort.

For our analysis of the robot’s perceived warmth, the linear mixed-effects model that best fit the data did not use any covariates. We found significant differences in participants’ perceptions of the robot’s warmth between experimental conditions. Those in the vulnerable condition ( $M = 6.13, SD = 1.15$ ) viewed the robot as more warm than both participants in the neutral condition ( $M = 4.95, SD = 1.71, c = 1.18, SE = 0.30, p < 0.001$ ) and the silent condition ( $M = 4.25, SD = 1.44, c = 1.88, SE = 0.31, p < 0.001$ ). Additionally, participants in the neutral condition viewed the robot as more warm than participants in the silent condition ( $c = 0.70, SE = 0.31, p = 0.030$ ).

For our analysis of the robot’s perceived competence, the linear mixed-effects model that best fit the data used the covariate of gender ( $c = 0.73, SE = 0.28, p = 0.010$ ). We found

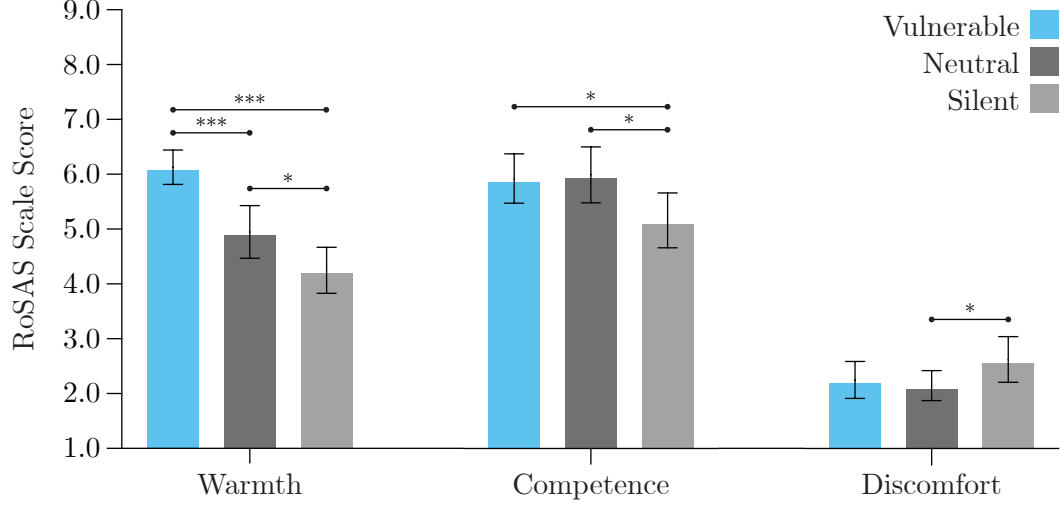


Figure 5.4: The robot was viewed as warmer if it made vulnerable utterances than either neutral or no utterances (silent), and warmer if it made neutral utterances as opposed to no utterances. The robot was viewed as more competent if it made vulnerable or neutral utterances as opposed to no utterances (silent). The robot was viewed as causing more discomfort if it did not make any utterances (silent) when compared with the neutral condition. (\*) and (\*\*) denote  $p < 0.05$  and  $p < 0.01$  respectively. Error bars represent a 95% confidence interval.

that participants in both the vulnerable condition ( $M = 5.93, SD = 1.64, c = 0.86, SE = 0.38, p = 0.027$ ) and the neutral condition ( $M = 5.99, SD = 1.82, c = 0.79, SE = 0.38, p = 0.044$ ) viewed the robot as more competent than participants in the silent condition ( $M = 5.16, SD = 1.72$ ). There were no significant differences in participant perceptions of robot competence between the vulnerable and neutral conditions ( $c = 0.07, SE = 0.37, p = 0.851$ ).

For our analysis of the robot's perceived discomfort, the linear mixed-effects model that best fit the data used the covariate of age ( $c = -0.02, SE = 0.01, p = 0.139$ ). We found that participants in the silent condition ( $M = 2.63, SD = 1.44$ ) perceived the robot to cause more discomfort than participants in the neutral condition ( $M = 2.15, SD = 0.98, c = 0.52, SE = 0.25, p = 0.037$ ). Although participants in the silent condition regarded the robot with more discomfort than those in the vulnerable condition ( $M = 2.25, SD = 1.23$ ), the difference is not significant ( $c = 0.44, SE = 0.25, p = 0.078$ ). There were also no significant differences between perceived robot discomfort between the vulnerable and neutral conditions.

## Interaction of Team Members with the Robot

In order to explore the influence of the robot’s vulnerable utterances on the behavior of team members towards the robot during mistake rounds, we specifically examined whether or not the person who made the mistake looked at the robot after the mistake as well as whether or not any human team member spoke to the robot after a mistake was made (Figure 5.5).

In analyzing the proportion of mistake rounds where the participant who made the mistake looked at the robot, we used a generalized linear mixed-effects model that best fit the data with covariates of age ( $c = -0.03, SE = 0.02, p = 0.059$ ) and gender ( $c = 0.57, SE = 0.31, p = 0.068$ ). Annotations of the participant making a mistake looking at the robot had a high Cohen’s kappa value of 0.99. We found that participants who had made an error looked at the robot after the round concluded more often in the vulnerable condition ( $M = 0.82$ ) compared with both participants in the neutral condition ( $M = 0.65, c = 1.07, SE = 0.40, p = 0.008$ ) and participants in the silent condition ( $M = 0.27, c = 2.64, SE = 0.46, p < 0.001$ ). Additionally, participants who had made an error looked at the robot after the round concluded more often in the neutral condition compared with participants in the silent condition ( $c = 1.57, SE = 0.40, p < 0.001$ ).

For our analysis of the proportion of mistake rounds where a participant made a verbal response directed to the robot (evaluated for each participant), we used a generalized mixed-effects model that best fit the data with covariates of the mistake round number ( $c = 0.15, SE = 0.04, p < 0.001$ ), participants’ extraversion score ( $c = 0.15, SE = 0.05, p = 0.002$ ), and the participant’s average familiarity with their fellow human participants ( $c = 0.16, SE = 0.09, p = 0.085$ ). Annotations of the participant making a verbal response to the robot had a high Cohen’s kappa value of 0.96. We found that participants were significantly more likely to respond verbally to the robot in the vulnerable condition ( $M = 0.18$ ) than both the neutral condition ( $M = 0.09, c = 0.95, SE = 0.26, p < 0.001$ ) and the silent condition ( $M = 0.05, c = 1.50, SE = 0.32, p < 0.001$ ). There was no significant difference in verbal responses to the robot between participants in the neutral and silent conditions ( $c = 0.54, SE = 0.35, p = 0.121$ ). Examples of participants’ responses to the robot include

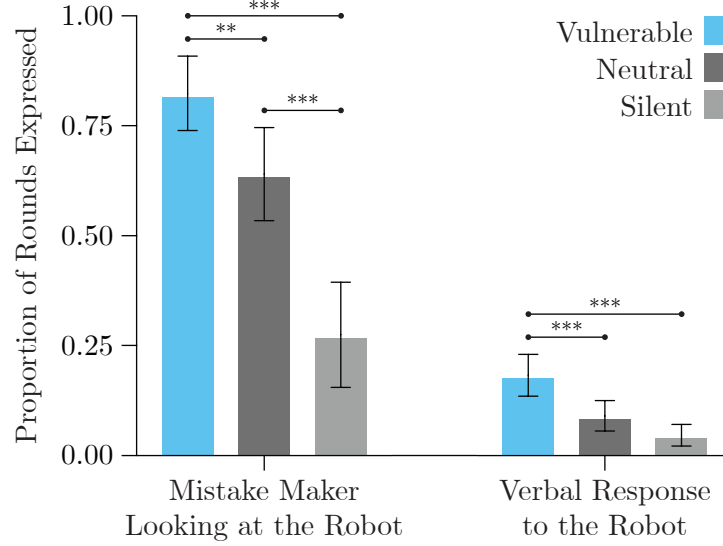


Figure 5.5: Participants interacting with a vulnerable robot were more likely to look at the robot after having made a mistake and were more likely to verbally respond to the robot than participants interacting with a neutral or silent robot. (\*), (\*\*), and (\*\*\*) denote  $p < 0.05$ ,  $p < 0.01$ , and  $p < 0.001$ , respectively. Error bars represent a 95% confidence interval.

the following: “*I know*,” “*Sure*,” “*Oh Echo...*,” “*Yeah, that’s true*,” and “*It’s your fault!*”

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 the robot and found a distinct difference in participant responses by condition. Participants in the vulnerable condition often noted how the robot eased the tension the groups experienced and was generally encouraging, saying that the robot’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 neutral condition had a slight negative connotation with the utterances the robot made, saying the robot’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*.” Participants

in the silent condition seemed to feel mostly neutral about the robots utterances (beginning of round utterances and mid-round utterances), saying that they were “*fine*,” “*a bit random and not super helpful*,” “*awkward*,” and “*random*.” From these responses, it seems likely that participants viewed the robot more positively and useful in easing the tension in the vulnerable condition as opposed to the neutral and silent conditions. Additionally, when comparing the vulnerable and neutral conditions, the robot seemed to be viewed as more approachable and less judgmental. The positive view of the robot, the robot’s approachability (compared with the neutral condition), and the larger verbal engagement (compared with the silent condition) could possibly explain the increased behavioral engagement we observed with participants interacting with the robot in the vulnerable condition.

### 5.4.3 Participant Interactions with Fellow Team Members during Failure Rounds

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 team members’ behavioral reactions to a mistake being made. We specifically focus on explaining mistakes to the group when they were made, consoling members of the team who made mistakes, and mistake rounds where participants laughed together (Figure 5.6).

In our analysis of the proportion of participants who explained their mistake during the round where they made the mistake, we used a generalized linear mixed-effects model that best fit the data with covariates of mistake round number ( $c = -0.11, SE = 0.07, p = 0.090$ ), age ( $c = -0.02, SE = 0.02, p = 0.300$ ), and the participant’s average familiarity with their fellow human participants ( $c = 0.33, SE = 0.19, p = 0.078$ ). Annotations of participants explaining their mistake had a high Cohen’s kappa value of 0.98. After having made a mistake, participants in the vulnerable condition ( $M = 0.69$ ) were significantly more likely to explain their mistake to their team members than participants in both the neutral condition ( $M = 0.50, c = 1.14, SE = 0.58, p = 0.048$ ) and the silent condition ( $M = 0.36, c = 1.67, SE = 0.62, p = 0.007$ ). There were no differences in explaining mistakes between participants in the neutral and silent conditions ( $c = 0.54, SE = 0.62, p = 0.388$ ). Examples of a participant explaining their mistake include: “*yeah, I can’t do it, I don’t have*

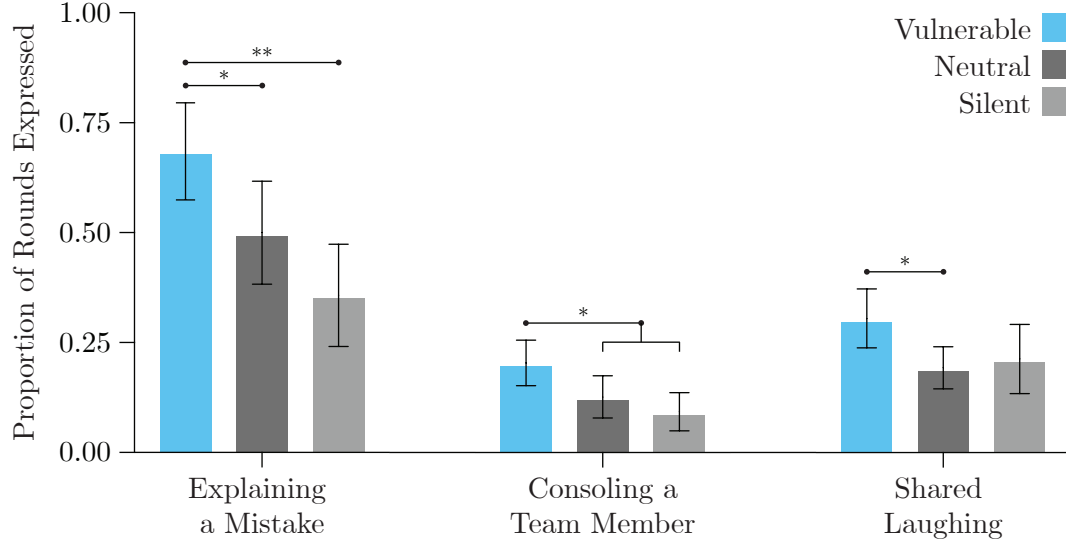


Figure 5.6: Participants interacting with a vulnerable robot were more likely to explain their mistake to their human teammates and console a human teammate who had made a mistake than participants interacting with a neutral or silent robot. Also, participants interacting with a vulnerable robot were more likely to laugh together than participants interacting with a neutral robot. (\*) and (\*\*) denote  $p < 0.05$  and  $p < 0.01$  respectively. Error bars represent a 95% confidence interval.

*the right pieces,” “I failed! I don’t have a piece,” “it forced me to fail,” and “my piece just disappeared.”*

In our analysis of the proportion of mistake rounds a participant consoled their fellow team members after a mistake, we found that the fixed effect representing the comparison between neutral and silent conditions did not significantly contribute to the model’s predictions of participants’ consoling behaviors. Thus, we pooled the neutral and silent conditions and compared the vulnerable condition to these pooled conditions. The generalized linear mixed-effects model that best fit the data had covariates of the mistake round number ( $c = -0.19, SE = 0.05, p < 0.001$ ) and participant age ( $c = -0.06, SE = 0.03, p = 0.021$ ). Annotations of the participant consoling the participant who made a mistake had a high Cohen’s kappa value of 0.89. After another participant made a mistake, participants in the vulnerable condition ( $M = 0.20$ ) were significantly more likely to console the participant who made the mistake than participants in the silent and neutral conditions ( $M = 0.11, c = 0.86, SE = 0.40, p = 0.031$ ). Examples of a participant consoling another participant include: *“it’s ok, mistakes happen,” “it’s ok, it’s not your fault,” “that’s*

*fine, that's fine, we're good," "it's alright," and "yeah, that's what happened to me last time we failed."*

One possible explanation for why participants in the vulnerable condition exhibited more consoling behavior than those in the neutral and silent conditions is that in the vulnerable condition the robot admitted its two mistakes, whereas in the neutral and silent conditions the robot did not admit its mistakes. In the vulnerable condition, the robot used the following two utterances to admit each of its two mistakes: *"I'm sorry everyone. My path was incomplete that round. I feel bad letting you all down."* and *"Sorry guys, I made the mistake this round. I know it may be hard to believe, but robots make mistakes too."* These utterances by the robot in the vulnerable condition often elicited consoling utterances from the robot's human teammates. In order to examine the proportion of times a participant consoled their fellow *human* team members (excluding instances of consoling the robot), a generalized linear mixed-effects model pooling the neutral and silent conditions (as described in the prior paragraph) best fit the model with mistake round number ( $c = -0.14, SE = 0.06, p = 0.012$ ), participant age ( $c = -0.06, SE = 0.03, p = 0.052$ ), and the participant's average familiarity with their fellow human participants ( $c = 0.33, SE = 0.15, p = 0.028$ ). We found that when considering only the consoling behavior exhibited to fellow *human* team members (and excluding consoling behavior exhibited to the robot), participants in the vulnerable condition still were more likely to console their fellow human teammates ( $M = 0.22$ ) than participants in the neutral and silent conditions ( $M = 0.18$ ), but no significant difference exists between conditions ( $c = 0.52, SE = 0.47, p = 0.26$ ).

In our analysis of the proportion of mistake rounds where team members laughed together after a mistake, we found differences in our initial analysis comparing the vulnerable and neutral conditions [Strohkorb Sebo et al., 2018] and our analysis comparing the three conditions (vulnerable, neutral, and silent) after the silent condition was added. When analyzing the data comparing the vulnerable condition to the neutral condition, we used a generalized linear mixed-effects model that best fit the data with covariates of participant age ( $c = 0.03, SE = 0.02, p = 0.048$ ) and the participant's average familiarity with their fellow human participants ( $c = 0.24, SE = 0.11, p = 0.026$ ). Annotations of participants laughing together had a high Cohen's kappa value of 0.86. We found

that participants in the vulnerable condition ( $M = 0.31$ ) were significantly more likely to laugh together in the aftermath of a mistake than participants in the neutral condition ( $M = 0.19, c = 0.79, SE = 0.39, p = 0.045$ ). After the silent condition was added to the experiment, and the statistics comparing the conditions were re-computed, we used a generalized linear mixed-effects model that best fit the data with covariates of participant age ( $c = 0.22, SE = 0.14, p = 0.131$ ) and the participant’s average familiarity with their fellow human participants ( $c = 0.28, SE = 0.13, p = 0.029$ ). We scaled the two covariates to ensure model convergence. This analysis yielded no significant differences in participants laughing together after mistakes were made between those in the vulnerable condition ( $M = 0.31$ ), neutral condition ( $M = 0.19$ ), and silent condition ( $M = 0.21$ ), even if the neutral and silent conditions are pooled together. Thus, when considering participants’ shared laughing, we have found that participants laugh together more in the vulnerable condition than the neutral condition, however, we can make no conclusions about whether participants in these two conditions behave differently when compared with participants in the silent condition.

One possible interpretation of these results is that participants in the vulnerable condition, as compared with the neutral and silent conditions, were more likely to embrace vulnerability because of the example of the robot. Team members interacting with a vulnerable robot were more likely to be vulnerable with one another (e.g., explaining their mistake), support the vulnerability of others (e.g., consoling team members who made mistakes), and ease tension through laughter.

#### 5.4.4 Participant Conversation Dynamics throughout the Game

To address our third research question about whether social robot vulnerability affects human team members’ conversational dynamics, we examine the amount that participants talked over time and the type of utterances they produced as well as how equal the conversation was distributed between the three human participants.

There was a substantial difference between conditions in the total amount of time spent talking by participants, where those in the vulnerable condition spoke twice as much over the course of the game ( $M = 253.60s, SD = 184.41s$ ) compared to those in the neutral condition



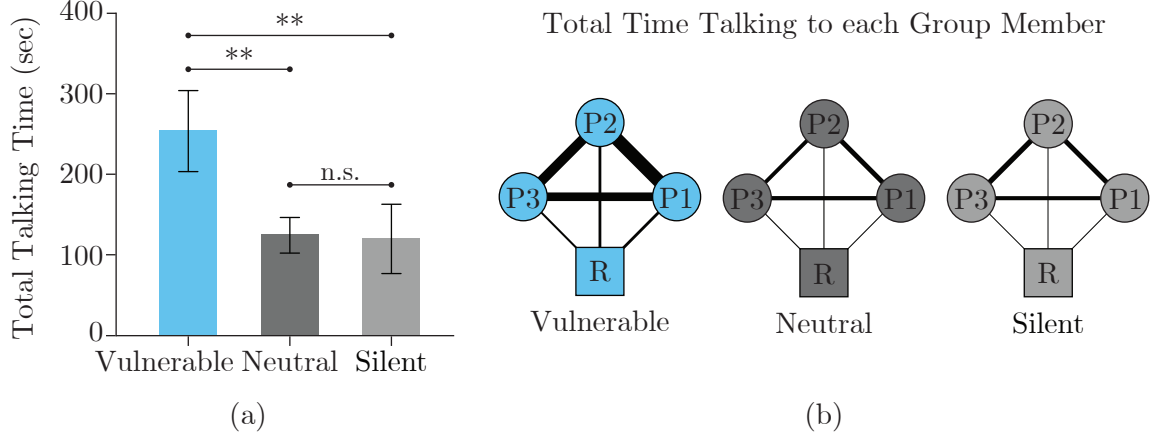


Figure 5.7: Compared to the neutral and silent conditions, human participants in the vulnerable condition spoke more, in total, to the other participants in their group, and increasingly across game rounds. In (a), we see that participants in the vulnerable condition spoke significantly more than participants in either the neutral or silent conditions ( $n = 153$  participants). In (b), the line widths represent the amount of talking by human participants toward their teammates who are connected by the line, in seconds (summed across all groups within a condition ( $n = 153$  participants))). R = robot; P1, P2, and P3 = human participants, in their relative positions around the table. (\*\*) denotes  $p < 0.01$  and error bars represent a 95% confidence interval.

( $M = 124.23s$ ,  $SD = 78.78 s$ ) and the silent condition ( $M = 119.86s$ ,  $SD = 148.17 s$ ), see Figure 5.7(a). This difference was statistically significant between the vulnerable and neutral conditions ( $c = 140.68$ ,  $SE = 39.97$ ,  $p = 0.001$ ) and the vulnerable and silent conditions ( $c = 124.52$ ,  $SE = 41.05$ ,  $p = 0.004$ ), even after adjustment for age, extraversion, gender, and familiarity, using regression models, but there was no significant difference between the silent and neutral conditions ( $c = 16.15$ ,  $SE = 42.40$ ,  $p = 0.705$ ).

In Figure 5.7(b), we show the total time participants spent talking to each of the other human participants and the robot, represented by the line width of the connections in the group network. The vulnerable robot condition enhanced interhuman conversation. In the neutral condition, across all groups, participants spoke to their human teammates for 83.22 min and to the robot for 10.27 min over the course of the game. In the vulnerable condition, participants spoke to their human teammates and the robot more than twice as much (178.38 min and 24.02 min, respectively). In the silent condition, participants spoke to their human teammates for similar amounts of time to the neutral condition (83.07 min) but spoke to the robot very little (5.61 min).

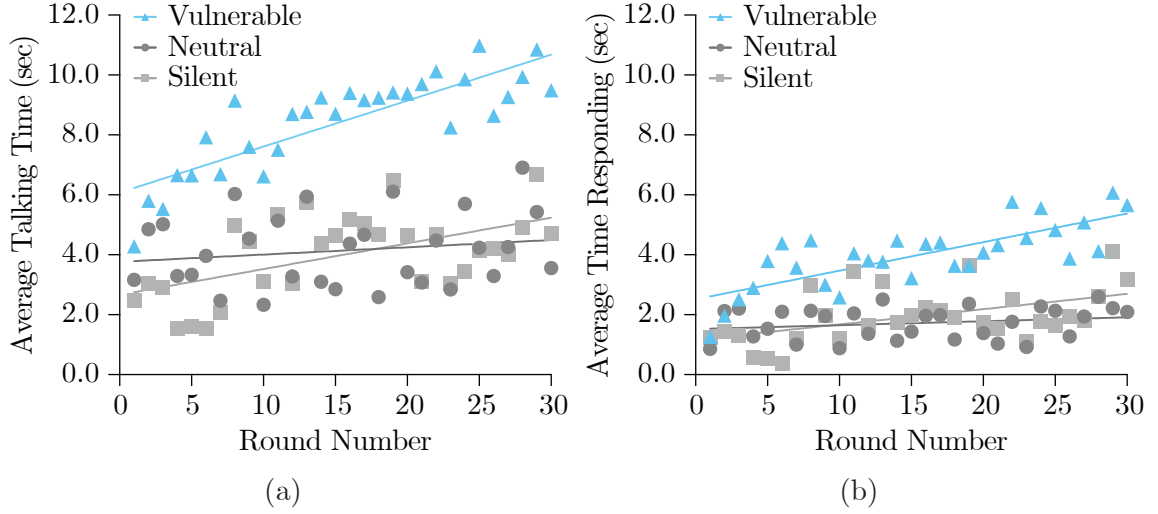


Figure 5.8: Here, we examine how much time participants in the experimental conditions spent talking during the 30 rounds of the game. In (a), the vulnerable condition has more talking in every round, and the slope (i.e., the rate of increase in talking per round, across rounds) is higher than the neutral condition (but indistinguishable from the silent condition). In (b) we see that, compared to the neutral condition, those in the vulnerable condition respond more over time to their fellow human group members ( $n = 4,590$  rounds).

Additionally, those in the vulnerable condition spoke progressively more over time (across rounds in the game) ( $M = 8.45s$ ,  $SD = 8.33$  s) compared to those in the neutral condition ( $M = 4.14s$ ,  $SD = 4.99$  s) as demonstrated by the significant interaction effect between round and experimental condition with respect to the vulnerable and neutral conditions ( $c = 0.13$ ,  $SE = 0.06$ ,  $p = 0.031$ ), although there is no significant difference between the neutral condition and silent condition ( $M = 4.00s$ ,  $SD = 6.73$  s,  $c = 0.06$ ,  $SE = 0.06$ ,  $p = 0.316$ ) or the vulnerable and silent condition ( $c = 0.07$ ,  $SE = 0.06$ ,  $p = 0.266$ ), see Figure 5.8(a).

We further find that the difference in the amount of talking by those in the vulnerable condition was primarily driven by one type of utterance, namely, the communications between the human players themselves, with an increase in responses to other humans over time. In other words, participants in the vulnerable condition ( $M=3.99s$ ,  $SD = 5.16$  s) increased the amount of time they spent responding to the utterances of their other human group members as the game progressed, Figure 5.8(b), compared to those in the neutral condition ( $M = 1.72s$ ,  $SD = 2.76$  s,  $c = 0.08$ ,  $SE = 0.04$ ,  $p = 0.039$ ), although there was no difference between the silent condition ( $M = 1.96s$ ,  $SD = 4.01$  s) and the neutral condition

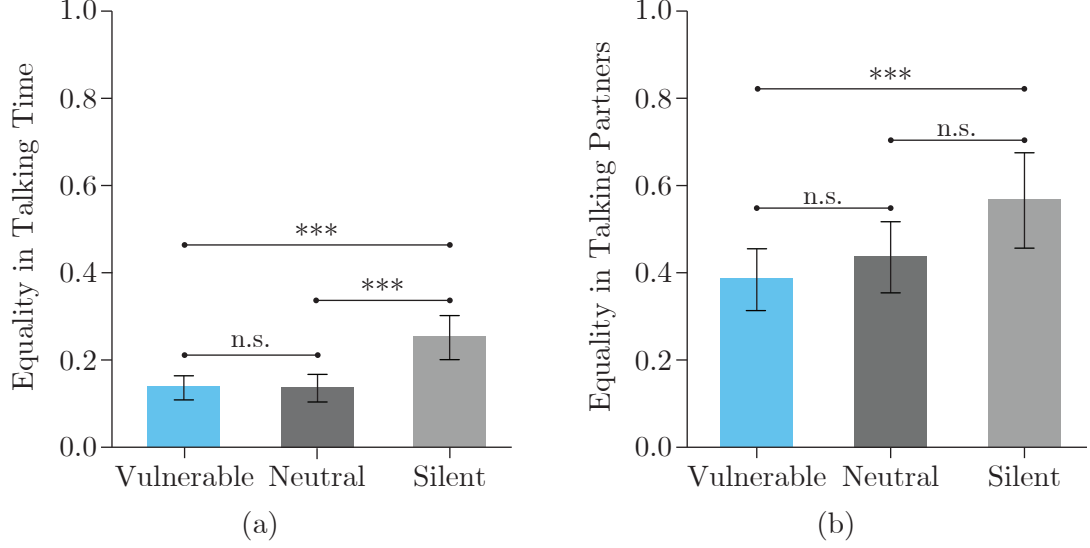


Figure 5.9: Although there was (a) no statistical difference between the vulnerable and neutral conditions in the equality in talking time ( $n = 150$  participants; one group did not speak at all and was excluded), the silent condition had less equality in talking time when compared with the other two conditions. In (b), we see that participants in the vulnerable condition directed their utterances more equally to each of their human group members than participants in the silent condition, as measured by the total amount of time spent talking to each participant’s two human partners ( $n = 144$  participants; participants who didn’t speak at all or who did not make directed utterances were excluded). (\*\*\*) denotes  $p < 0.001$  and error bars represent a 95% confidence interval.

( $c = 0.04$ ,  $SE = 0.04$ ,  $p = 0.355$ ) or the vulnerable condition and the silent condition ( $c = 0.04$ ,  $SE = 0.04$ ,  $p = 0.272$ ), in terms of the increase over rounds. No other utterance category was statistically significant across rounds of the game.

In addition to examining the amount of participants’ speech, we also explored how equally participants’ speech durations were within a group as well as how evenly participants distributed their speech to the two other human members in the group. To quantify the former, we used the following “equality in talking time” (ETT) metric:

$$E_{TT_i} = c \left| \frac{\tau_i}{\sum_1^n \tau_i} - \frac{1}{n} \right|$$

where  $\tau_i$  represents the total amount of time participant  $i$  spoke during the game,  $n$  is the number of human participants (3 in this case)<sup>‡</sup>,  $\sum_1^n \tau_i$  is the total amount of time participant  $i$ ’s group spoke during the game, and  $c$  is a normalizing constant, causing  $E_{TT_i}$  to have a

<sup>‡</sup>In this analysis a value of 0.33 was used to approximate  $\frac{1}{3}$  in the computation of  $E_{TT_i}$ .

range of  $[0, 1]$ .  $E_{TT_i}$  takes on a value of 0 when a participant speaks for a third of the total amount of time their group speaks and a value of 1 when a participant speaks and their group members did not speak at all. Thus, low equality in talking time values indicate more equal talking times between the participant and their fellow group members. We found that equality of time speaking did not differ between the vulnerable robot condition ( $M = 0.14$ ,  $SD = 0.10$ ) and the neutral robot condition ( $M = 0.14$ ,  $SD = 0.11$ ) ( $c = -0.03$ ,  $SE = 0.18$ ,  $p = 0.884$ ), but there is a significant difference between the neutral and silent conditions ( $M = 0.25$ ,  $SD = 0.17$ ) ( $c = -0.63$ ,  $SE = 0.19$ ,  $p < 0.001$ ) and the vulnerable and silent conditions ( $c = -0.66$ ,  $SE = 0.19$ ,  $p < 0.001$ ). In other words, the distribution of speech by participants in the vulnerable condition did not differ from that in the neutral condition, but participants in the silent condition had the least equal distribution of talking time, as seen in Figure 5.9(a). Thus, the mere presence of a robot that communicates may enhance the equality of talking time in conversation among humans in a group.

To examine how evenly distributed each participant's utterances were toward their fellow human teammates in the human-robot group, we created an "equality in talking partners" ( $E_{TP}$ ) metric as follows:

$$E_{TP_i} = \frac{|\tau_{(P_i, P_j)} - \tau_{(P_i, P_k)}|}{\tau_{(P_i, P_j)} + \tau_{(P_i, P_k)}}$$

where  $\tau_{(P_i, P_j)}$  represents the talking time of participant  $i$ 's speech specifically directed at participant  $j$  during the game and  $\tau_{(P_i, P_k)}$  represents the talking time of participant  $i$ 's speech specifically directed at participant  $k$  during the game. In other words, this measures how balanced a participant's speech is toward the two other human members of their group over the whole game. If a participant directs all of their speech to one participant and none to the other, that participant gets a value of 1. If a participant speaks for the exact same amount of time to each of the other two participants, that participant will receive a value of 0. In other words, values of 1 represent perfect inequality and 0 represents perfect equality. Every human-human pairwise comparison is made for each participant in each group. We found no evidence that participants in the vulnerable condition ( $M = 0.38$ ,  $SD = 0.26$ ) distributed their speech more equally between their fellow human group members than those in the neutral condition ( $M = 0.43$ ,  $SD = 0.29$ ) ( $c = -0.38$ ,  $SE = 0.28$ ,  $p =$

0.164) or between the silent condition ( $M = 0.57, SD = 0.34$ ) and neutral condition ( $c = 0.36, SE = 0.31, p = 0.247$ ). However, the vulnerable condition is significantly more equal than the silent condition ( $c = -0.74, SE = 0.30, p = 0.013$ ), as shown in Figure 5.9(b). In other words, speech seems to be more balanced in the conditions with a vulnerable robot compared to a silent robot.

#### 5.4.5 Perceptions of Group Dynamics

To examine participants' perceptions of the group dynamics, we examine participant responses to the psychological safety scale and open ended questions in the post-experiment questionnaire.

For our analysis examining participants' ratings on the psychological safety survey measure, the linear mixed-effects model that best fit the data did not use any covariates. We did not find any significant differences in the psychological safety scores between participants in our experimental conditions: vulnerable ( $M = 5.62, SD = 0.75$ ), neutral ( $M = 5.53, SD = 0.73$ ), and silent ( $M = 5.31, SD = 1.01$ ). 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.

Using comments that the participants provided in the post-experiment survey to the question "How would you describe the group dynamics while you were playing this game?", we analyzed how participants perceived their own group's dynamics. Comments were reliably coded by two coders into the following four categories, with corresponding high inter-rater reliability values (Cohen's kappa): quiet ( $kappa = 0.87$ ), positive ( $kappa = 0.90$ ), supportive ( $kappa = 0.87$ ), and fun ( $kappa = 0.98$ ). We found that those in the vulnerable condition thought of their groups as being less quiet ( $M = 0.20$ ) than did those in the neutral condition ( $M = 0.39, c = -1.28, SE = 0.57, p = 0.025$ ). There were no significant difference in the perception of the group being quiet between the silent condition ( $M = 0.38$ ) and both the vulnerable and neutral conditions. Those in the vulnerable condition viewed their groups as more positive ( $M = 0.76$ ) than did those in the neutral condition ( $M = 0.56, c = 1.36, SE = 0.66, p = 0.039$ ) and the silent condition ( $M = 0.48, c = 1.76, SE = 0.71, p = 0.013$ ). There was no significant difference between

the silent and neutral conditions in how positive they viewed their groups. Those in the vulnerable condition also viewed their groups as being more fun ( $M = 0.31$ ) than did those in the neutral condition ( $M = 0.12, c = 1.44, SE = 0.67, p = 0.031$ ) and the silent condition ( $M = 0.08, c = 1.74, SE = 0.73, p = 0.018$ ), and the silent and neutral conditions did not significantly differ. There was no significant difference between any of the experimental groups in how supportive participants found their groups. In summary, participants in the vulnerable condition described their groups as more pleasant overall than those in either of the two other conditions.

## 5.5 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, increased trust-related behavior expression (explaining errors and consoling other team members), and more positive conversational dynamics (time spent talking, equality in talking time) toward fellow team members when the robot in the group makes vulnerable statements as opposed to either neutral statements or no statements.

With regards to interactions with the robot in this experiment, participants directed more gaze toward the robot after they made a mistake and directed more verbal statements toward the robot in the aftermath of a mistake in the vulnerable condition as opposed to the neutral and silent conditions. Based on survey responses discussed in the results section, we believe that this increased engagement with the robot in the experimental condition reflected a greater sense that the robot was interactive and approachable, where participants need not fear social judgment from the robot. This increase in engagement with the robot that makes vulnerable utterances has two likely explanations: 1) participants viewed the robot as having greater agency due to the robot’s vulnerable statements, and 2) participants perceived the robot as more approachable and feared social judgment from the robot less due to the robot’s vulnerable statements.

Barsade (2002)’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. 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 influenced both 1) team members’ *human-human trust-related interactions* with each other during tense moments and 2) the group’s *human-human conversational dynamics*. Human team members interacting with the vulnerable robot 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. Additionally, human team members interacting with a vulnerable robot talked more and responded more to the comments of others over the course of the entire experiment. This increase in vulnerable behavior and conversation displays the distinctive influence social robot vulnerability has on trust-related human-human behavior within teams.

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 [Edmondson, 1999]. 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.

## 5.6 Summary

In this work, 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 in 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 be forced to 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, neutral statements, and no 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) and an increase in conversation between human team members. These results demonstrate the positive influence robots can have on trust in human-robot teams.

This work was the first to demonstrate that a robot’s verbal actions within a group can influence how people in the group behave towards one another. This robot influence on human-to-human behavior is best seen in how participants who interact with a vulnerable robot are more likely to explain a mistake they made to their fellow human teammates, where an increase in robot vulnerability resulted in an increase of human-to-human vulnerability. In addition, those interacting with a vulnerable robot consoled each other more often after making mistakes, conversed more with one another, and described the group dynamic as more positive and fun. These results demonstrate that robots are able to positively shape group behavior and emphasize the influential role that robots can have on the way we converse and interact with each other in the context of human-robot groups. As robots and other forms of machine intelligence (such as digital assistants and online bots) become increasingly prevalent in our daily lives, they will likely shape our actions, relationships, and conversations. Therefore, it is critical that we work to understand how artificial agents can best be programmed and designed to have a *positive* influence on our interactions with other people.

Through the use of trust repair strategies (Chapter 4) and vulnerable utterances, as we explored in this chapter, we have demonstrated a robot’s ability to shape trust in both a one-on-one setting and a group setting. Importantly, we have shown that a robot’s behavior influences not only how people perceive the interaction, but how people behave in



interactions with a robot and each other. In the next chapter, we examine another social dynamic that is essential to team success: inclusion. We use a unique experimental design to investigate several strategies a robot could possibly use to improve how included people feel within a group.

## Chapter 6

# Robots that Shape the Inclusion of Human Team Members\*

Team member inclusion is vital to the success of collaborative teams, positively influencing both the commitment of team members and the team’s overall performance [Cho and Mor Barak, 2008, Sabharwal, 2014, Shore et al., 2011]. Since we have shown that a robot can have a positive influence on trust-related behavior and conversation dynamics between people in a human-robot team (Chapter 5), we also hypothesize that a social robot could shape perceived human team member inclusion as well.

In this chapter, we explore two strategies to increase the inclusion of human team members in a human-robot team: 1) giving a person in the group a specialized role (the ‘robot liaison’) and 2) having the robot verbally support human team members. In particular, we examine the influence of these two strategies on team members who may feel excluded or marginalized. In a human subjects experiment ( $N = 26$  teams, 78 participants), groups of three participants completed two rounds of a collaborative task. In round one, two participants (ingroup) completed a task with a robot in one room, and one participant (outgroup) completed the same task with a robot in a different room. In round two, all three participants and one robot completed a second task in the same room. This creation of an

---

\*Portions of this chapter were originally published as: S. Strohkorb Sebo, L. L. Dong, N. Chang, and B. Scassellati (2020). Strategies for the inclusion of human members within human-robot teams. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI ’20, pages 309–317, New York, NY, USA. Association for Computing Machinery. [Strohkorb Sebo et al., 2020a]

ingroup and outgroup allows us to test how effective our strategies for increasing human team member inclusion are on the outgroup team member specifically. During round two, we implemented the robot’s two strategies to increase inclusion by designating one participant was as the robot liaison the and having the robot verbally support each participant 6 times on average.

Results show that participants with the robot liaison role had a lower perceived group inclusion than the other group members. Additionally, when outgroup members were the robot liaison, the group was less likely to incorporate their ideas into the group’s final decision. In response to the robot’s supportive utterances, outgroup members, and not ingroup members, showed an increase in the proportion of time they spent talking to the group. Our results suggest that specialized roles may hinder human team member inclusion, whereas supportive robot utterances show promise in encouraging contributions from individuals who feel excluded.

## 6.1 Introduction

Collaborative teams work best when each team member feels included (see Section 2.1.3 in Chapter 2 for a more in-depth review of the literature on inclusion in human teams). As social robots become members of work teams consisting of both humans and robots, it is important to consider the possible influence of the robot on the inclusion of their human team members. Prior work within the field of human-robot interaction (HRI) has shown that robots are capable of shaping group dynamics (e.g., group cohesion [Short and Matarić, 2017]) and related behaviors (e.g., conflict management [Jung et al., 2015, Shen et al., 2018], balanced participation [Tennent et al., 2019], and vulnerable expression [Strohkorb Sebo et al., 2018]). Therefore, it is reasonable to assume that the actions of a robot could both positively and negatively influence their human team members’ perceived inclusion. Additionally, if a robot is able to increase the inclusion of their human team members, the team is likely to benefit in both the commitment of their team members and in the team’s overall performance.

In this work, we are interested in investigating two strategies for enhancing the inclusion

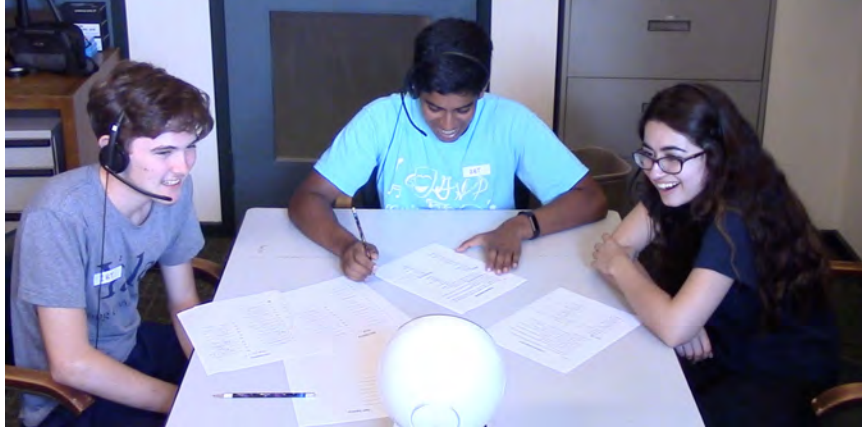


Figure 6.1: Three participants completed a collaborative task with a Jibo robot, where the robot used two distinct strategies to enhance the inclusion of the human team members.

of human team members through interactions with a robot in a collaborative task setting (Figure 6.1). The first strategy we investigate is giving a member of the team a specialized role, where only that team member can ask the robot questions related to the task. Our exploration of a specialized role to interface with a robot is especially relevant to the HRI community because robots are commonly incorporated into human-robot teams by training one person to operate the robot (e.g., factory teams, search and rescue teams, surgical teams). The second strategy that we explore is having the robot give verbal support to human team members, such as, *“Luis, I think that’s worth considering.”* We evaluate the efficacy of these two strategies in a human subjects experiment where three people and a robot complete a collaborative task. One of the human team members is given the ‘robot liaison’ role, being the only one who can ask the robot questions to gather more task-related information, and all team members receive verbal support from the robot. We assess the influence of these two inclusion strategies on human team member inclusion by analyzing participants’ perceived inclusion ratings, conversational dynamics, and task decisions made by the group.

## 6.2 Background and Research Questions

We review related work that highlights the potential efficacy of two strategies of including human team members: giving a member a specialized role to interact with the robot and

supporting human team members with targeted utterances from the robot.

### 6.2.1 Strategy 1: Specialized Roles in Groups

If given a specialized role to interact with a robot in the context of a human-robot team, it is possible that the role might give a sense of value to the team member, enabling them to contribute uniquely to the group. However, it is also possible that the role might further isolate them from the group. A person’s perception of their isolation or inclusion within a group is often determined by the existence and perception of ingroups and outgroups and where one stands in relation to the rest of the members. Faultlines, divisions in a group along a salient characteristic (e.g., age, gender), may determine how these ingroup and outgroup relationships are formed [Lau and Murnighan, 1998]. A faultline could be created by giving a member a leadership role or another similarly specialized role. The research on the relationship between leadership and isolation suggests that giving a member a specialized role in the group may lead to feelings of exclusion [Rokach, 2014]. The relationship between inclusion and specialized roles in human teams has been studied extensively; however, to our knowledge, no research has been conducted on how specialized roles impact perceptions of inclusion in human-robot teams.

*Research Question 1: How does being given a **specialized role to interact with a robot** influence a human team member’s inclusion in a human-robot team?*

### Strategy 2: Verbal Support

Teams with high levels of inclusion consist of members who “feel their values and norms are supported” [Cho and Mor Barak, 2008]. Support within teams may encompass encouraging ideas, acknowledging accomplishments, providing assistance, or simply backchanneling [Hertel and Hüffmeier, 2011]. Backchanneling, one form of verbal support that has been researched extensively, has been defined as “the short utterances produced by one participant in a conversation while the other is talking” [Ward and Tsukahara, 2000]. Many consider backchanneling to include nonverbal signals as well, including nodding, facial expressions, and directional gaze [Stubbe, 1998]. All feedback responses, regardless of form,

serve the same function of confirming that the “speaker and listener share a common frame of reference” without threatening the speaker’s position as primary speaker [Stubbe, 1998]. Though responsive feedback may not always be positive or supportive, research has shown that unsupportive verbal feedback does not occur frequently [Stubbe, 1998]. Given the engaging and communicative nature of backchanneling as well as the importance of team members support, we have reason to believe that supportive utterances from a robot may influence the inclusion of human team members. Although work within HRI has demonstrated the efficacy of robot backchanneling to communicate that the robot is attentively listening [Jung et al., 2013, Lala et al., 2017, Lee et al., 2019], no work to our knowledge has investigated the influence of supportive utterances on human inclusion within groups.

*Research Question 2: How do **supportive utterances from a robot** influence a human team member’s inclusion in a human-robot team?*

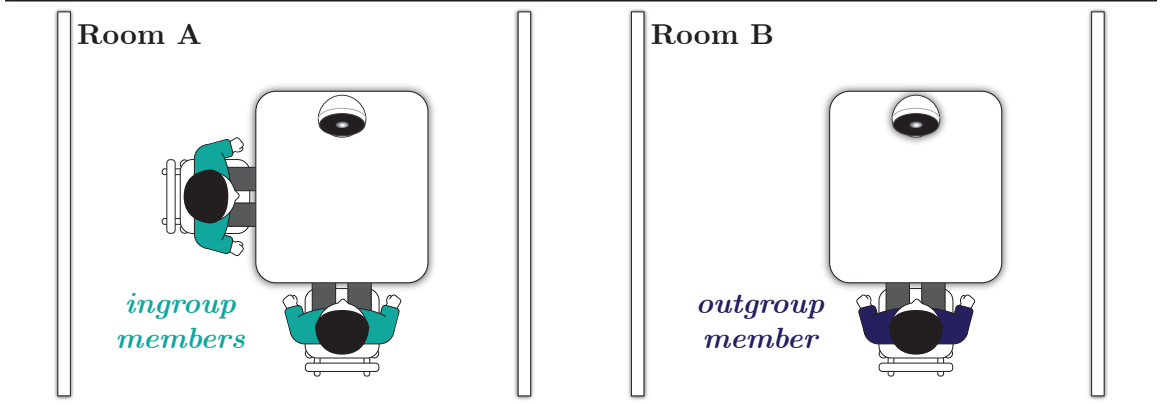
## 6.3 Methods

In this section, we describe a human subjects experiment investigating the influence of a specialized role involving interaction with a robot and the influence of supportive utterances on the inclusion of human members within a human-robot team.

### 6.3.1 Experiment Design

We designed a between-subjects experiment where three human participants and a robot work together on a collaborative task. To study the influence of the robot on participants who may experience exclusion, we attempted to artificially form an ingroup and outgroup within the three participants. We did this by having participants first complete round one of the task independently within assigned subgroups of sizes one (**the outgroup member**) and two (**the ingroup members**), and then gathered them as a combined team to complete a second round, see Figure 6.2. During round two, we studied the influence of a specialized role by assigning one of the three human participants as the **robot liaison**, the sole human member with the ability to ask the robot information. These designations of human participants led to our two between-subjects conditions:

**Round 1** (15 minutes)



**Round 2** (30 minutes)

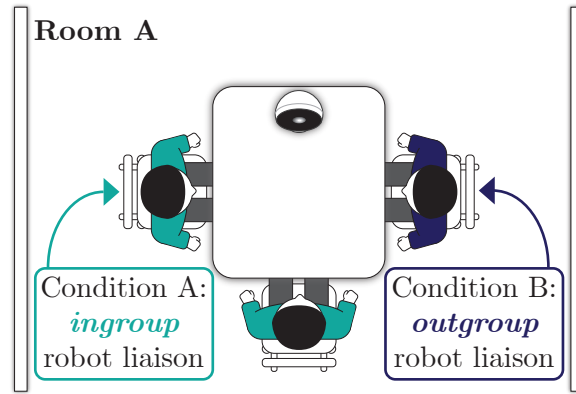


Figure 6.2: In round 1 of the experiment, two participants (ingroup) and a Jibo robot completed a task in room A while one participant (outgroup) and a Jibo robot completed the same task in room B. The outgroup participant joined the two ingroup participants and the robot in room A for round 2 of the experiment, where one of the members is designated the robot liaison.

- Condition A: the robot liaison is an ingroup member
- Condition B: the robot liaison is an outgroup member

Using this experimental design, we addressed our two research questions described in Section 6.2. For our first research question, we investigated the influence of a specialized role to interact with a robot by examining the difference in inclusion and related behaviors of both ingroup and outgroup participants with the robot liaison role. For our second research question, regardless of condition and robot liaison designation, the robot targeted each participant in the group with supportive utterances (Section 6.3.3). We are able to measure each participant’s reactions to these supportive utterances from the robot and other measures of inclusion, and investigate if the ingroup-outgroup or robot liaison designations

influence the effect of the robot’s supportive utterances.

### 6.3.2 Collaborative Task: The Survival Problem

For this experiment, we designed a collaborative task where we asked players to assign ranks to items from a given list based on how useful each item would be for survival in a hostile environment. This task is a derivative of the Desert Survival Problem [Lafferty and Pond, 1974], a commonly used task in HRI groups research (e.g., [Chidambaram et al., 2012], [Kidd and Breazeal, 2004], [Tennent et al., 2019]).

In the first round of the task, players were given 15 minutes to construct an ordered list of 25 items, ranked by importance for survival, from a list of 25 common household items (e.g., umbrella, whistle, watch). All the players received a sheet of paper for their item rankings as well as an instruction sheet that stated the rules of the round, listed the survival items, and provided instructions for players to verbally query the robot for additional information (see Appendix A, Section A.4.1 for the full instruction sheet). During this round, players interacted with a social robot through verbal queries about the time remaining in the round and to learn more information about the survival items. For example, when queried about the survival item ‘soda,’ the robot responded with *“6 aluminum cans of Coca-Cola. The cans are held in cardboard and the whole pack is wrapped in plastic.”*

In the second round of the task, the team was given 30 minutes to agree on a final list of eight items from the original list of 25 that they deemed to be the most essential for survival. In addition to the information in the first round, the robot in round 2 provided facts regarding various environmental factors such as the weather, plants, or geography. We chose a mountainous climate for the survival location, rather than a desert island. This allowed us to provide participants with additional material for further group discussion and encouraged questioning of prior assumptions. For example, the following is a piece of information the robot gave participants environment: *“Life threatening temperature is rare, but does occur. Make sure you save up supplies to survive a 3-day long blizzard.”* As in round one, all players received an instruction sheet, which stated the rules of round two and also listed the environmental factors and instructions to query the robot in addition to the original item information provided in round one (see Appendix A, Section A.4.2 for the



full instruction sheet). Lastly, a single sheet of paper was provided for the team’s finalized list of eight items.

### 6.3.3 Robot Platform and Behaviors

For our experiment, we used the commercial robot Jibo [Jibo, 2017]. Jibo is 11 inches tall and has a 3-axis motor system and a touchscreen face. Jibo responded to the verbal utterances of participants, captured through individual headset microphones and converted from a speech signal to text using Google’s speech-to-text API<sup>†</sup>. We programmed Jibo to play the role of a social robot that engages in various forms of supportive social behaviors while providing assistance to players through the supply of valuable task-related information, described in Section 6.3.2.

In response to participants’ speech, we designed the robot to display general social behaviors to establish itself as a present and active member of the human-robot team. In both rounds of the experiment, the robot nodded with a probability of 0.25 in response to detected speech. The robot also responded verbally to detected speech with a probability of 0.15 during both rounds of the experiment: if an item name is detected within the participants’ utterance, then with a probability of 0.5, the robot’s verbal response either gave a useful hint about an item (e.g., “*Whiskey is a great disinfectant*”), or made a general comment using the names of the items within the participant utterance (e.g., “*Screwdriver, interesting*”, “*Honey, tape, okay*”). If no items were detected in the participants’ utterance, the robot’s verbal response consisted of a generic verbal backchannel (e.g., “*Uh huh*”, “*Yeah*”). Please refer to Appendix A, Section A.5 for more details pertaining to the robot utterances during the game.

During round two of the experiment, we designed the robot to deliver **targeted supportive utterances**. These targeted supportive utterances reinforce the ideas of participants and also use the participant’s name (robots using participant names has shown importance in building relationships and engaging people [Kanda et al., 2004, Kanda et al., 2007]). In the experiment, the robot responded with targeted supportive utterances that either 1) rephrased and supported an idea proposed by a participant (*rephrase*), 2) supported

---

<sup>†</sup><https://cloud.google.com/speech-to-text/docs/libraries>

Table 6.1: Examples of the targeted supportive utterances the robot made during the experiment, where [p-name] is a placeholder for the participant’s name.

Type	Targeted Backchannel Utterance Example
simple	“ <i>Okay</i> , [p-name].”
simple	“ <i>Interesting</i> , [p-name].”
item	“ <i>Camera</i> , [p-name] <i>I think that’s worth considering.</i> ”
item	“ <i>Soda, chocolate, that makes sense</i> [p-name].”
rephrase	“ <i>We need a coffee pot. Good idea</i> [p-name].”
rephrase	“ <i>Use garbage bag to store the berries. Okay</i> , [p-name].”

an item that a participant mentioned (*item*), or 3) simply showed support to the participant themselves (*simple*). Examples of the three types of targeted supportive utterances are shown in Table 6.1 and a description of all possible targeted supportive utterances and how they are selected is included in Appendix A, Section A.5.3. Of the targeted supportive utterances the robot produced during the experiment, 29% were rephrase, 34% were item, and 37% were simple. We programmed the robot to deliver one targeted supportive utterance to each human participant every 4.5 minutes during round two of the experiment. This resulted in participants receiving an average of 5.62 ( $SD = 0.86$ ) targeted supportive utterances each throughout the course of the experiment.

### 6.3.4 Procedure

Upon arrival of the participants, an experimenter obtained informed consent and then asked participants to independently fill out a pre-experiment questionnaire on tablets provided by the experimenter. Then, the experimenter informed participants that they would complete a two-part timed activity, to be completed in randomly divided subgroups of sizes one and two, before completing part two as a group of three. The experimenter first set up the single participant (outgroup) in room B. During the approximately 5 minutes the experimenter was setting up the outgroup participant, the participants of the two-member group (the insiders) were given a list of “get to know you” questions in order to further enforce the ingroup-outgroup divide (e.g., “If you didn’t sleep, what would you do with your extra time?”), before the experimenter returned to lead the ingroup participants to room A.

In both rooms, the experimenter asked the participants to put on the headsets at their

pre-assigned seats. The experimenter then played Jibo’s introduction (see Appendix A, Section A.5.1) through the tablet, and had each participant practice successfully querying Jibo about one of the survival items. The experimenter then initiated round one of the task on the tablet, which lasted 15 minutes, and left the room.

After the participants finished the first round, the experimenter escorted the outgroup participant to room A to join the ingroup participants. The experimenter told the participants that they would be given 30 minutes to complete the second part of the task. The experimenter then initiated round two of the task on the tablet (which lasted 30 minutes), played Jibo’s introduction to the round (see Appendix A, Section A.5.1), and designated one of the participants as the robot liaison using the following language: “*In this part, unlike the first, only one of you will be able to ask Jibo questions about the items and environment. For all of you this is [participant name].*”

After the task finished, the experimenter led the participants outside of the experiment room and administered the post-experiment questionnaire, which the participants completed on tablets. Finally, participants received a \$10 cash payment.

### **6.3.5 Measures**

To evaluate how the robot liaison role and the robot’s targeted supportive utterances influenced participant inclusion, we analyzed participants’ responses to pre- and post-experiment questionnaires, rankings of survival items in both rounds of the experiment, and conversational behavior.

#### **Pre-experiment survey measures**

In order to capture pre-existing differences between participants, we collected measures of prior familiarity with team members and two personality measures via a survey administered before the human-robot interaction.

Prior familiarity with team members was assessed by asking participants to evaluate their relationships with each of the two participants on a scale from 1 (*I have not met this participant before we completed this study together; I do not know them*) to 5 (*I would consider this participant to be one of my closest friends*). Participants were also asked

whether they had the phone numbers or social media contact information of their team members. The full details of the familiarity questionnaire can be found in Appendix B, Section B.5.

We also assessed participants' extraversion and emotional intelligence. We measured participants' extraversion because it is a necessary covariate when analyzing data pertaining to the amount of time people spend talking in the group (Section 6.3.5). We measured participants' emotional intelligence because prior work has demonstrated its correlation with team performance [Stubbs Koman and Wolff, 2008].

To measure extraversion, we used an abbreviated version of the Revised Eysenck Personality Questionnaire (EPQR-A) [Francis et al., 1992] that includes six binary (0 - no, 1 - yes) response questions such as "Do you tend to keep in the background on social occasions?" to construct a single score between 0 (low extraversion) to 6 (high extraversion), see Appendix B, Section B.6 for more details. To measure emotional intelligence, we administered the Short Form of the Trait Emotional Intelligence Questionnaire (TEIQue-SF) [Cooper and Petrides, 2010], which asks respondents to indicate how much they agree or disagree with a set of 30 statements, such as "I'm usually able to influence the way other people feel", on a 7-point Likert scale from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*), see Appendix B, Section B.8 for more details.

### Survival item ranking measures

We examined how similar each of the two subgroup's lists were to the final list of eight items by calculating the absolute difference between the team's final ranking of eight most important items,  $r_{fin}(i)$ , from round two, and the ranks initially assigned to these items,  $r_{init}(i)$ , by each subgroup in round one,  $Diff = \sum_{i=1}^{r_{fin}} abs(r_{fin}(i) - r_{init}(i))$ . We normalize the difference scores between the ingroup and outgroup to get our similarity score, e.g., for the outgroup  $Diff_{out} = Diff_{out} / (Diff_{in} + Diff_{out})$ . For these differences measures, lower scores indicate a higher level of similarity between the initial list and the final list. For example, a score of 0 indicates that the initial and final lists were exactly identical.

We also analyzed how each item on the final list of eight items was initially ranked by each subgroup as either high (ranked 1-8) or low (ranked 9-25), and computed the proportion

of those items that made it onto the final list of eight items. We chose to consider the items ranked 1-8 on the initial list as high because the final list contained exactly 8 items.

### **Conversational measures**

We investigated several aspects of the conversation that occurred between the three participants during the second round of the experiment: each participant’s total time spent talking, the standard deviation of the total talking times of each of the three participants in the group, the number of times each item was mentioned, and the proportion of time participants spent talking in response to the robot’s targeted supportive utterances.

### **Post-experiment survey measures**

In the post-experiment questionnaire we assessed participants’ perceived inclusion by administering the Perceived Group Inclusion Scale (PGIS) [Jansen et al., 2014]. PGIS asks participants to rate agreement with statements like “this group gives me the feeling that I belong” and “this group encourages me to be authentic” on a scale from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*), see Appendix B, Section B.9 for more details. To measure participants’ perceptions of Jibo, we used the Robotic Social Attributes Scale (RoSAS) [Carpinella et al., 2017]. RoSAS asks respondents to rate how closely they consider descriptor words, each representative of either warmth, competence, or discomfort, to be associated with the robot on a scale from 1 (*Definitely Not Associated*) to 9 (*Definitely Associated*), see Appendix B, Section B.4 for more details.

The post-experiment questionnaire also contained several long-response questions asking participants to describe the team’s interactions on the survival task and the question “Of the two other human participants, which participant would you prefer to work with on a school or work project?” From the responses to this last question, we assigned each participant a preference score. If participant A and participant C specified participant B as their preference, participant B’s preference score would be 2 (one for each participant that ‘voted’ for them). If a participant indicated that they were fine working with both the other participants, each of the other participants received a score increase of 1.0. If a participant said they would prefer working with neither of the other participants, neither

of the other participants received any score increase.

### 6.3.6 Participants

Participants were recruited for this study from a high school program held at Yale University. The students from the program came from 80 different countries, with 47% from the United States. The breakdown by continent is: 52% from North America, 24% from Asia, 12% from Africa, 7% from Europe, 3% from South America, and 2% from Australia.

A total of 30 groups (90 participants) were recruited for participation in this study. Of the 30 groups recruited, 4 groups were excluded due to either not finishing the experimental task or technical difficulties (e.g., a participant’s microphone got disconnected disabling them from querying the robot). For the 26 remaining groups (78 participants), 38 participants were female and 40 participants were male. The average age of participants was 16.82 ( $SD = 0.72$ ). There were 6 all female groups, 4 all male groups, 4 groups with 2 females and 1 male, and 12 groups with 1 female and 2 males. For the 16 groups with mixed-gender compositions, we balanced by gender the designation of both the outgroup member (9 females, 7 males) and the robot liaison (8 females, 8 males). There were 13 groups with an ingroup robot liaison and 13 groups with an outgroup robot liaison. More detailed descriptive statistics for participants in each condition as well as each division of participants (ingroup/outgroup, robot liaison) can be found in Appendix C Tables C.37 and C.38.

Participants’ familiarity with the other participants in their group ( $M = 1.10, SD = 1.06$ ), extraversion ( $M = 3.90, SD = 2.15$ ), and emotional intelligence ( $M = 5.27, SD = 0.65$ ) were assessed in the pre-experiment questionnaire. Using mutli-level mixed effects models described in Section 6.4, we did not find any significant differences of these characteristics between either participant designations of robot liaison or participant designations of ingroup-outgroup.

## 6.4 Results

For our analysis of the participant data, we used linear mixed-effects models in order to account for each participant being in a group of three. We designated intergroup bias (ingroup or outgroup), robot liaison designation (yes or no), the interaction between those two variables, and relevant covariates as fixed effects; and the participant's group as a random effect (random intercept). We tested these models for multicollinearity (variance inflation factor), selected them based on the Bayesian information criterion, and evaluated residual errors for lack of trends and heteroscedasticity. For each fixed effect, the model outputs the linear coefficient ( $c$ ), the standard error ( $SE$ ), and the significance ( $p$ ) value of that predictor.

When analyzing data for each group, we used an analysis of variance (ANOVA) where each group is an independent sample. The main independent variable of interest is whether the robot liaison is an ingroup member or an outgroup member. We used the following covariates in this analysis: the average familiarity of group members and the number of females in the group. The effect size is reported as partial eta squared ( $\eta^2$ ). For more details on the results of the statistical models included in this section, please refer to Appendix C, Tables C.37 - C.48.

### 6.4.1 Ingroup-Outgroup Differences

Based on our experimental design that introduced an intergroup bias where one participant (outgroup) completed round one separately from the two other participants (ingroup), we expected that there would be inclusion-related differences between ingroup and outgroup participants. We observed this bias in the similarity of ingroup and outgroup survival item rankings from round one with the final list the team produced after round two, as well as in the post-experiment preferred partner scores.

We analyzed the similarity of the final list of 8 items with both the ingroup and outgroup's initial ranking of those 8 items, where smaller values indicate higher similarity of the lists. We used a linear mixed-effects model that best fit the data with emotional intelligence ( $c = -0.05, SE = 0.02, p = 0.023$ ) as a covariate. We found that ingroup members

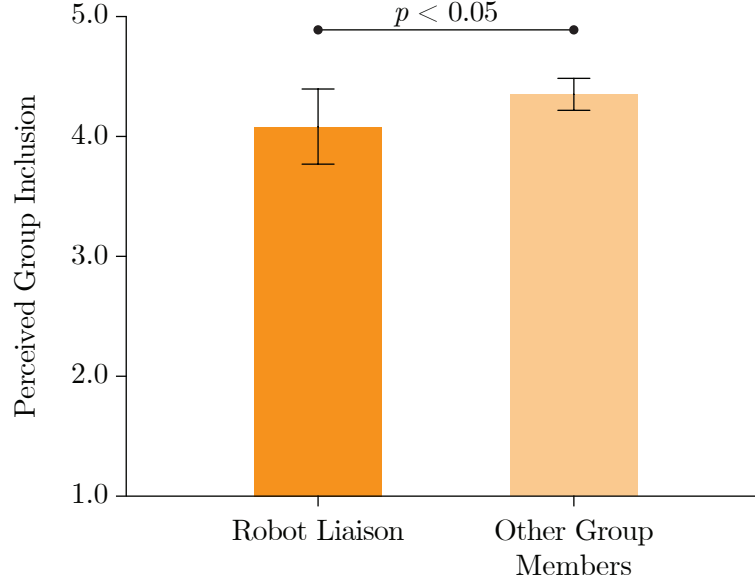


Figure 6.3: Participants who were the robot liaison had a lower perceived group inclusion than the other group members. Error bars represent a 95% confidence interval.

had a more similar ranking of the top 8 items on their initial list ( $M = 0.45, SD = 0.11$ ) than outgroup members ( $M = 0.55, SD = 0.11, c = 0.10, SE = 0.04, p = 0.005$ ).

We also examined partner preference scores, our measure of how much a participant is preferred as a teammate by their fellow participants. We analyzed the partner preference scores using a linear mixed-effects model that best fit the data with age ( $c = 0.21, SE = 0.11, p = 0.055$ ) and emotional intelligence ( $c = 0.44, SE = 0.12, p < 0.001$ ) as covariates, and excluded the data from three participants who did not answer the questionnaire item. We discovered that ingroup participants had significantly higher partner preference scores ( $M = 1.08, SD = 0.73$ ) than outgroup participants ( $M = 0.80, SD = 0.56, c = -0.51, SE = 0.21, p = 0.019$ ).

These ingroup-outgroup differences verify our experimental design of imposing intergroup biases among the three participants. The ingroup's higher similarity between the initial and final item rankings and the higher preference for ingroup members as work partners serve as a manipulation check.



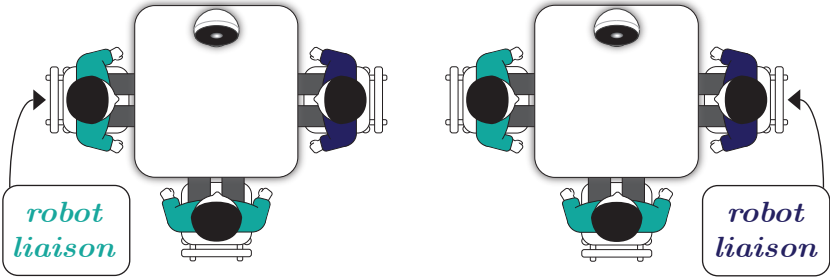
### 6.4.2 Influence of the Robot Liaison Role

In order to investigate the influence of the robot liaison role on participant inclusion (Research Question 1), we analyzed the perceived group inclusion survey measure as well as the measures of which survival items were included on the team’s final list of eight items.

For our analysis of the participants’ ratings on the perceived group inclusion scale, the linear mixed-effects model that best fit the data used the covariates of age ( $c = -0.20, SE = 0.09, p = 0.025$ ) and maximum familiarity ( $c = 0.12, SE = 0.05, p = 0.011$ ). We found that participants who were the robot liaison had lower ratings of perceived group inclusion ( $M = 4.08, SD = 0.78$ ) than the other group members ( $M = 4.35, SD = 0.49, c = -0.41, SE = 0.17, p = 0.021$ ), as shown in Figure 6.3.

In order to analyze the influence of the robot liaison and the ingroup-outgroup designations, we examined the items initially ranked high (items ranked 1-8) and low (items ranked 9-25) by each of the 2 subgroups in round one of the experiment. We then calculated the proportion of these items that made it onto the final list of 8 items produced by the entire team at the end of round two of the experiment (see Figure 6.2). We found that a higher proportion of items were chosen that were initially low on the ingroup list and initially high on the outgroup list ( $L_{in}, H_{out}$ ) if the robot liaison was an ingroup member ( $M = 0.47, SD = 0.14$ ) than if the robot liaison was an outgroup member ( $M = 0.32, SD = 0.18, F = 5.59, \eta^2 = 0.19, p = 0.027$ ). Thus, when the outgroup member is the robot liaison, as opposed to an ingroup member, the team is less likely to incorporate items favored by the outgroup member.

These findings suggest that the robot liaison role works against efforts to increase inclusion in human team members, resulting in decreased perceived inclusion in the robot liaison and the incorporation of fewer of the outgroup robot liaison’s ideas into the team’s final list of items. When asked in the post-experiment questionnaire if the status of any group member influenced the group dynamic, most participants did not think so (e.g., “*I do not think so. All of us expressed our points of view.*” and “*No, although only one could speak to Jibo, we all contributed to the decisions equally*”). However a few participants did express a difference in group member status (e.g., “*I think that since I was the only one allowed*



Initial Survival Item Ranking	Ingroup Robot Liaison	Outgroup Robot Liaison
$H_{in}, H_{out}$	0.71 (2.0/2.8 items)	0.79 (2.6/3.2 items)
$H_{in}, L_{out}$	0.42 (2.1/5.2 items)	0.40 (2.1/4.8 items)
$L_{in}, H_{out}$	0.47 (2.4/5.2 items)	0.32 (1.5/4.7 items)
$L_{in}, L_{out}$	0.14 (1.5/11.8 items)	0.15 (1.8/12.3 items)

Table 6.2: This table reports the proportion of survival items that were initially ranked high and low by the ingroup ( $H_{in}, L_{in}$ ) and outgroup ( $H_{out}, L_{out}$ ) that made it onto the group’s final list of 8 items. When the outgroup member was the robot liaison, the team was significantly less likely to incorporate the survival items they initially valued ( $L_{in}, H_{out}$ ) onto the team’s final list.

to ask Jibo questions it made me more dominant” and “[participant name] being the only person who could ask Jibo questions made it feel like she was the only person who had a connection with Jibo; It didn’t ultimately affect the group but other group members had to ask [participant name] to ask Jibo questions”). From these responses, it seems that participants did not overwhelmingly feel a difference in group inclusion or membership because of the robot liaison status. However, some participants did point to a noticeable difference in the power dynamics where the robot liaison was seen as having greater influence than the other members. This could help explain the robot liaison’s lower perceived inclusion and, when they were an outgroup member, their reduced likelihood of the group incorporating their ideas.

### 6.4.3 Influence of the Robot’s Supportive Utterances

To investigate the influence of the robot’s supportive utterances (Research Question 2), we examined participants’ verbal responses to the robot’s targeted supportive utterances. In this analysis, we excluded four participants’ data because they did not comply in keeping

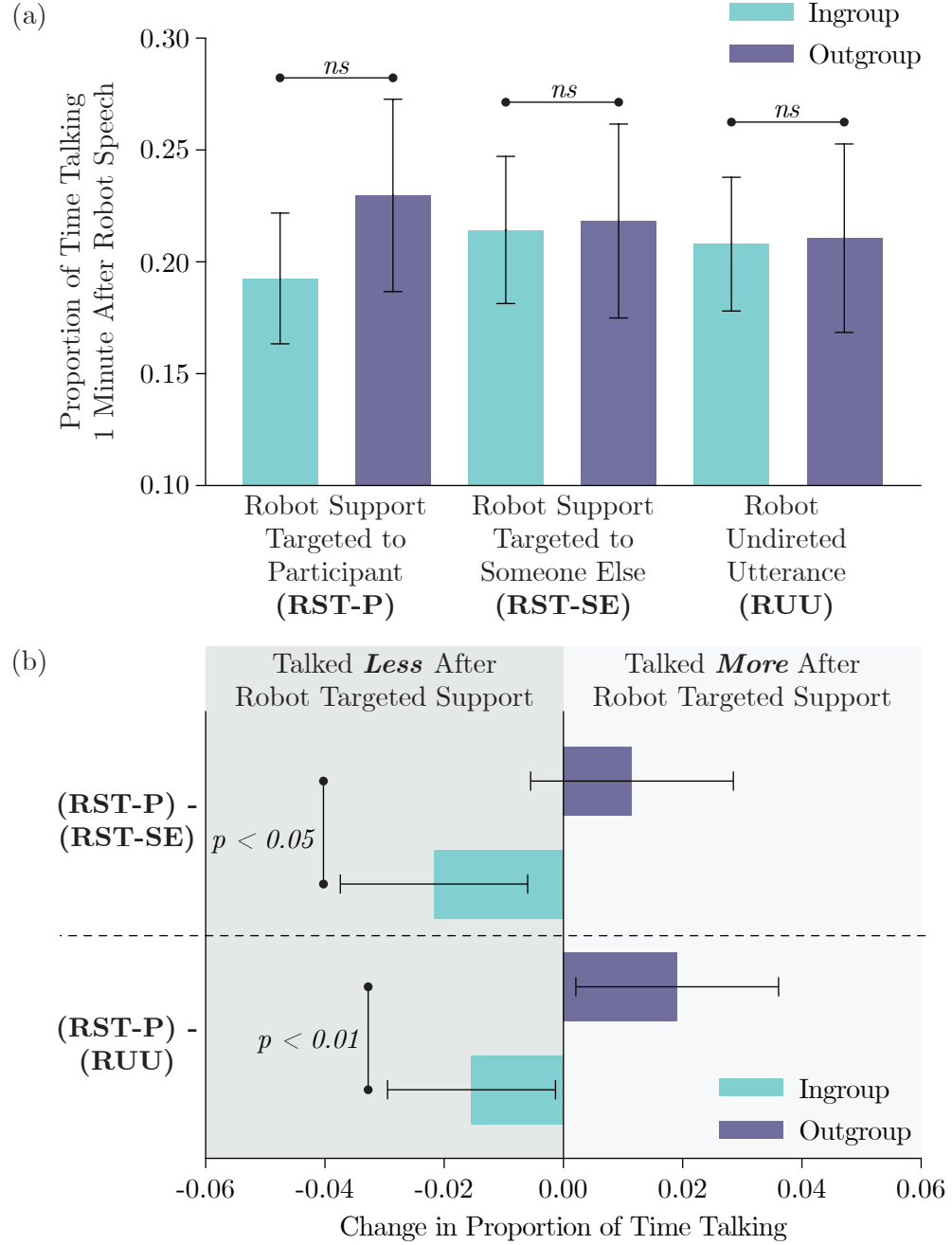


Figure 6.4: Outgroup participants, as opposed to ingroup participants, displayed a significantly higher difference in the proportion of time they spent talking during the one minute after the robot’s support targeted to the participant (RST-P) when compared with two baselines: 1) the proportion of time they spent talking during the one minute after the robot support was targeted to someone else (RST-SE) and 2) the proportion of time they spent talking during the the one minute after a robot undirected utterance (RUU). Error bars represent a 95% confidence interval.

their microphone on during the experiment. We compared the proportion of time a participant spent talking 1 minute after the robot delivered a targeted supportive utterance (*robot support targeted to participant - RST-P*) with two controls: 1) the proportion of time a participant spent talking 1 minute after the robot delivered a targeted supportive utterance to someone else (*robot support targeted to someone else - RST-SE*) and 2) the proportion of time a participant spent talking 1 minute after an undirected utterance from the robot (*robot undirected utterance - RUU*). We use two controls, as opposed to one, in order to more rigorously test whether the recipient of a targeted supportive utterance talked more as a result of the targeted supportive utterance.

As shown in Figure 6.4(a), ingroup and outgroup participants did not display a significant difference in their proportions of time talking in the 1 minute after robot targeted support to the participant, RST-P, ( $c = 0.01, SE = 0.03, p = 0.703$ ), in the 1 minute after robot targeted support to someone else, RST-SE, ( $c = -0.03, SE = 0.03, p = 0.445$ ), or in the 1 minute after robot undirected utterances, RUU, ( $c = -0.03, SE = 0.03, p = 0.263$ ). These analyses were conducted with linear-mixed effects models that best fit the data with extraversion as a covariate (RST-P:  $c = 0.02, SE = 0.01, p = 0.004$ ; RST-SE:  $c = 0.02, SE = 0.01, p = 0.007$ ; RUU:  $c = 0.02, SE = 0.01, p = 0.002$ ).

We then examined the difference between each participant's proportion of time talking in the 1 minute after robot targeted support to the participant, RST-P, and our two controls (RST-SE and RUU), Figure 6.4(b). Positive values indicate that the participant spoke *more* (and negative values indicate that the participant spoke *less*) after the robot delivered a targeted supportive utterance to them as opposed to after a different type of robot utterance. The linear mixed-effects models that best fit the data for these two analyses did not use any covariates. When examining the difference in participants' talking after the robot gave targeted support to them as opposed to after the robot targeted support to someone else (RST-P – RST-SE), we found that outgroup members ( $M = 0.012, SD = 0.041$ ) had a more positive difference than ingroup members ( $M = -0.022, SD = 0.055, c = 0.03, SE = 0.02, p = 0.047$ ), indicating that outgroup members had a higher verbal response to the robot targeted support. In analyzing the difference in participants' talking after robot gave targeted support to them as opposed to after undirected robot utterances (RST-P –

RUU), we found that outgroup members ( $M = 0.019, SD = 0.041$ ) had a more positive difference than ingroup members ( $M = -0.015, SD = 0.049, c = 0.04, SE = 0.02, p = 0.007$ ) again indicating that outgroup members had a higher verbal response to the robot targeted support. These results suggest that the robot’s supportive utterances positively influenced outgroup, but not ingroup, participants’ verbal contributions during the task.

In the same analyses described in the prior paragraph, we also found differences between robot liaison participants and the other participants. When examining the difference in participants’ talking after the robot gave targeted support to them as opposed to after the robot targeted support to someone else (RST-P – RST-SE), we found that robot liaisons ( $M = 0.011, SD = 0.046$ ) had a more positive difference than other group members ( $M = -0.022, SD = 0.053, c = 0.03, SE = 0.02, p = 0.044$ ). However, in the difference in participants’ talking after the robot gave targeted support to them as opposed to after undirected robot utterances (RST-P – RUU), the robot liaison designation had no significant influence ( $c = 0.02, SE = 0.02, p = 0.274$ ). Since the increased proportion robot liaison talking time following robot targeted support is only supported by one of our baseline comparisons, it is possible that the robot liaisons have a similar boost in talking in response to robot targeted support as outgroup members, but the effect may not be as pronounced.

#### 6.4.4 Perceptions of the Robot

We evaluated participants’ perceptions of the robot by analyzing their responses to the Robotic Social Attributes Scale (RoSAS). Using linear mixed effects models, neither the ingroup-outgroup, robot liaison designations, nor the interaction between the two resulted in significant differences in participants’ perceptions of the robot’s warmth, competence, or discomfort. Across all participants, the average ratings of the robot’s RoSAS attributes, rated on a 1-9 Likert scale, are as follows: warmth ( $M = 5.81, SD = 1.45$ ), competence ( $M = 7.21, SD = 1.24$ ), and discomfort ( $M = 2.19, SD = 1.10$ ).

## 6.5 Discussion

In this study, we investigated two different ways in which robot team members can shape inclusion: 1) through a specialized role that gave one participant increased interaction with a robot and 2) through supportive utterances by the robot, targeted to each participant. To investigate the efficacy of these two strategies for increasing inclusion, we designed an experiment in which human-robot groups with experimentally manipulated ingroup-outgroup divides must work together as teams in a collaborative task.

We were able to confirm that our ingroup-outgroup manipulation was successful. On average, each team’s final item rankings were more similar to the initial item rankings made by ingroup members than those made by outgroup members. Moreover, ingroup participants were on average preferred as future work partners over outgroup participants by the other members of the team. Beyond our success in experimentally creating these subgroup divides in this specific study, we contribute to the field of HRI an experimental study design that enables researchers to investigate inclusion and ingroup-outgroup divides in future work.

To explore the influence of a specialized role on team member inclusion, we designated one member of the team the ‘robot liaison’ that allowed this participant to have privileged communications with the robot. On the one hand, this designation could have promoted team inclusion by providing a sense of value to the team member with the specialized skill. On the other hand, this could have further isolated the member through the addition of another dividing feature with the other members of the group. Our results showed stronger support for this second idea. Participants assigned to the role of the robot liaison reported lower levels of group inclusion than other team members, demonstrating the possible isolating effects of the role. It is possible that the introduction of a robot with exclusive access produces a change in the power dynamics of the team, particularly when the robot plays an essential role. Prior work has established a link between leadership and feelings of isolation and loneliness [Rokach, 2014]. Although the robot liaison in this experiment is not explicitly in a leadership position, the increased influence given to the robot liaison could have produced their lower perceived inclusion ratings.

This divide was particularly apparent when the robot liaison was an outgroup member. When an outgroup member was the robot liaison, as opposed to an ingroup member, the group as a whole incorporated fewer survival items favored by the outgroup member. It is likely that giving the robot liaison role to the outgroup member created another divide between the outgroup member and the ingroup members. This increased division may have made the ingroup members less receptive to the ideas of the outgroup member. This result highlights the dangers of having an already-excluded member of a team take on a specialized role, as the outgroup participants in this study had fewer of their ideas incorporated into the team’s final decision when they were given a specialized role to interact with the robot member of the team.

These findings have important implications for the increasing number of human-robot teams today, as robot members are frequently incorporated with a human “liaison.” These liaisons may be a specialized robot operator, such as in factory, search and rescue, and surgical teams, or simply a member with more implicit control of the robot, such as team leaders or remote members participating through telepresence. However, as we have shown, this practice can be detrimental to the perceived inclusion of the liaisons themselves, especially if there are already pre-existing faultlines (e.g., ingroup-outgroup) between the liaison and the other members.

In order to investigate our second proposed strategy for increasing team member inclusion, we programmed the robot to make targeted supportive utterances during team discussions. We found that whereas ingroup members appeared to be mostly unaffected by these supportive utterances, outgroup members of the team spoke more in the time immediately after receiving one of the robot’s targeted supportive utterances as compared to the times after other utterances made by the robot. Thus, we found evidence for robot-employed targeted supportive utterances improving team inclusion and contribution by providing support and affirmation to relatively excluded team members.

Overall, our results demonstrate that the roles and actions of a robot team member can influence overall team inclusion. We note that two factors, sample representativeness and robot utterance delivery, may have had a non-trivial impact on our results. First, because we recruited participants from a relatively small two week high school program,

it is possible that our teams had higher baseline levels of inclusion and common ground, and thus were also less affected by our ingroup-outgroup manipulation. Second, because we implemented relatively simple robot behaviors, the robot would occasionally produce poorly timed or irrelevant utterances (e.g., a robot response of “*key, good idea*” after a participant had said “*I don’t think we should bring the key*”). Because these two factors may have reduced the effectiveness of our experimental manipulations, either by an increased bias towards high inclusion or by minimizing the beneficial influence of the robot, we believe that the significance of our results in spite of these factors highlights the potential impact of robot team members, and thus the importance of considering the possible consequences of including robot members in teams.

A vast majority of teams consist of members with diverse skill sets, backgrounds, and experience. Inclusion is a critical component to both the success of the team and the commitment of its members [Cho and Mor Barak, 2008, Sabharwal, 2014, Shore et al., 2011]. In line with prior work in HRI, we have demonstrated that the social dynamics and behaviors of groups can be shaped by the actions of a robot member. However, we found that these effects are not necessarily always in a positive direction. The results of our experiment show that whereas the actions of robots can be used to promote a sense of team inclusion, differential abilities to interact with the robot may produce faultlines and isolate team members. As robots are increasingly incorporated into human teams, we recommend that we, as a community, work to better understand and take into account the influence robot members can have on inclusion and other social dynamics of the team, as we seek to promote the success of human-robot teams.

## 6.6 Summary

In this study, we explored different ways in which robot team members can shape human team member inclusion. We designed an experiment in which human-robot groups with experimentally manipulated subgroup divides must work together as a team in a collaborative task, and investigated two strategies of increasing human team member inclusion. We found that a robot’s use of supportive utterances encouraged excluded team members to



speak up more. However, granting a team member special access to the robot can isolate the member, particularly if the team member already occupies the role of an outgroup member within the team. As teams become more diverse and robots become ubiquitous in everyday life, it is necessary to better understand how we can assign roles and design behaviors for robots to maximize their positive impact on human inclusion in human-robot groups and teams.

This is the first work to explicitly explore how a robot can shape perceived inclusion in the human members of a human-robot team. We demonstrate ways in which robots can both positively influence human team member inclusion (supportive utterances from the robot) and negatively influence human team member inclusion (a specialized role to interact with the robot). We additionally present a novel experimental design that forms an intergroup bias (an ingroup and an outgroup) in the first phase of the experiment, and examines the robot’s influence on a group that has an intergroup bias in the second phase of the experiment. Just as many people seek to maximize diversity, equity, and inclusion in their workplaces and teams, it is essential that as robots join us in these spaces, they promote these same values.

In this chapter, we examined the efficacy of two strategies designed to improve inclusion: a specialized role to interact with the robot and supportive utterances from the robot. These strategies were fixed and did not adapt to the behavior of individuals within the group. Adaptation to individual behavior could be extremely useful to a robot, for example, enabling a robot to identify automatically which member of the group is feeling most excluded, allowing the robot to specifically target that individual with behaviors to include them in the group. In the next chapter, we take a look at identifying behaviors people express that correlate with their psychological safety and inclusion, so that in the future a robot could identify these factors in real time. We also examine the efficacy of improving the inclusion and psychological safety of team members with verbal support from the robot.

## Chapter 7

# Human Backchannels: Signals of Key Group Dynamics that can be Influenced by Social Robots

In order for a robot to improve team dynamics and performance, it is important for the robot to be able to sense and model current group dynamics to strategically act within the group. Based on work showing a connection between the presence of backchanneling (e.g., “yeah,” head nodding) and team performance [Jung et al., 2012], we hypothesize that backchannels might provide a signal for robots to use in sensing group dynamics such as psychological safety and inclusion. Additionally, it is possible that social robots could positively influence the backchanneling behavior of people within a group through verbal support, including backchannels of its own.

In this chapter, we explore 1) correlations between human backchanneling behavior and their self-reported perceptions of team social dynamics, and 2) how verbal support from a robot shapes human team member backchanneling behavior. We conduct a between subjects experiment ( $N = 38$  groups, 114 participants), where a robot either does or does not give verbal support to its three human teammates while the team works together on a collaborative task. Analysis conducted on the backchanneling behavior of the human participants indicates that the more verbal backchannels an individual receives, the more

positive view they have of the group’s social dynamics. However, too many backchannels directed towards an individual may result in lower feelings of inclusion, so striking the right balance is likely important to maintaining team social dynamics. The presence of verbal support from the robot did not increase the amount of backchanneling between human team members, but rather, seemed to replace backchanneling that would have occurred had the robot been silent.

## 7.1 Introduction

Team performance as well as team member satisfaction have been shown to be significantly predicted by team social dynamics such as inclusion [Cho and Mor Barak, 2008, Sabharwal, 2014, Shore et al., 2011], psychological safety [Edmondson, 1999], and trust [Jones and George, 1998, Mayer et al., 1995]. As robots join collaborative teams of people, it is essential that they be able to sense and adapt to these important social dynamics in real time.

Current work, both in HRI and psychology, measure team social dynamics by administering surveys, such as the Perceived Group Inclusion Scale [Jansen et al., 2014] and the Team Psychological Safety Scale [Edmondson, 1999]. No work to our knowledge has demonstrated the ability to sense team social dynamics in real time without explicitly asking human team members questions. In this work, we seek to identify social signals that can easily be measured with off-the-shelf cameras and microphones which are correlated with team social dynamics in order to enable robot teammates to sense and react to these dynamics in real time. We specifically explore backchannels, short vocalizations (e.g., “yeah”) or movements (e.g., head nodding) used to indicate that a person is actively listening to a speaker, which show promise as a measurable social signal that may connect with team social dynamics [Jung et al., 2012].

In order to explore the connection between human team member backchanneling and team social dynamics, we examine connections between human verbal and nonverbal backchanneling behavior and questionnaire measures of psychological safety [Edmondson, 1999] and inclusion [Shore et al., 2011] within the context of a collaborative task between three people and a social robot (Figure 7.1). We also test whether the presence of verbal support



Figure 7.1: Three participants and a Jibo robot completed a collaborative task. We annotated the backchannels made by the human participants (e.g., “yeah,” head nodding) and analyzed whether these backchannels were correlated with important group dynamics (psychological safety, inclusion) and how the human participant backchanneling behavior was shaped by the robot.

from the robot might increase and enhance the backchanneling behavior and team social dynamics of the human team members.

## 7.2 Background

In this section we review prior work that has explored the function of backchanneling within human teams, systematically analysed and mathematically predicted the occurrences of human backchannels, and investigated the utility of programming robots to employ backchannels in human-robot interactions.

### 7.2.1 Backchanneling in Human Teams

In human conversation, listeners are expected to provide some form of regular feedback to indicate to the speaker that they are still actively engaged in the conversation [Stubbe, 1998]. Listener feedback often comes in the form of backchanneling, which Ward and Tsukahara (2000) define as “the short utterances produced by one participant in a conversation while the other is talking.” This backchanneling feedback occurs regularly and frequently in conversation. The Japanese language even has a term, *aizuchi*, to describe verbal backchannels, which occur frequently in conversation and are sometimes even actively elicited. With

respect to American English, one study found that 19% of all utterances consisted of some form of verbal backchanneling [Jurafsky et al., 1997]. Backchannel responses are often not limited to just verbal utterances but also consist of nonverbal signals such as head nodding and shaking [Duncan, 1974, Stubbe, 1998].

Backchannels confirm that the “speaker and listener share a common frame of reference” without taking a speaker turn [Duncan, 1974] or threatening the speaker’s position as primary speaker [Stubbe, 1998]. Goodwin (1986) highlights two specific functions of backchannels: 1) to encourage the speaker to continue talking (e.g., a backchannel of “uh huh” in the middle of a speaker’s continued speech) and 2) acknowledges and briefly assesses the speech of the speaker (e.g., a backchannel of “oh wow” indicating both acknowledgement and surprise) [Goodwin, 1986]. Backchanneling has been shown to be more common in get-to-know-you conversations as opposed to competitive debates [Dixon and Foster, 1998]. Additionally, several studies have found that females backchannel more frequently than males [Duncan and Fiske, 2015, Roger and Nesshoever, 1987]. In a collaborative work context, pair programmers who exhibited more backchanneling exhibited higher objective performance scores as well as higher satisfaction ratings of their work and the overall experience [Jung et al., 2012].

### **7.2.2 Systematic Analysis and Prediction of Backchannels and Backchannel Opportunities**

Researchers have studied backchannels to better understand their function and use in conversation [Ward and Tsukahara, 2000, Ward, 2006] and to investigate when children develop backchanneling conversational skills [Hess and Johnston, 1988, Miller et al., 1985]. Additionally, researchers building conversational artificial agents have examined backchannel utterances in order to develop models to predict when an artificial agent should backchannel a human speaker [de Kok et al., 2013, Gravano and Hirschberg, 2009, Maatman et al., 2005, Morency et al., 2010, Truong et al., 2011], to anticipate when a human might backchannel an artificial agent speaker [Hjalmarsson and Oertel, 2012], and to estimate the attentive state of a human listener [Lee et al., 2017].

Much of this work was informed by collected data of human conversations, annotated

manually or automatically for audio features (e.g., pitch, energy, pauses, utterance length) and visual features (e.g., gaze, head movements, smiling, eyebrow movement) of the speaker and listener [Gravano and Hirschberg, 2009, Lee et al., 2017, Morency et al., 2010]. A majority of the human conversational data analyzed in this prior work examining backchanneling behavior involved one person telling a story or explaining content to one or more listeners, including the MultiLis corpus [de Kok and Heylen, 2011, de Kok et al., 2013], the ALICO corpus [Malisz et al., 2016], and the data collected within individual studies [Hess and Johnston, 1988, Lee et al., 2017, Morency et al., 2010]. Other human conversational data analyzed in prior work includes the IFADV corpus [Truong et al., 2011, Van Son et al., 2008], where previously acquainted human participants were told to speak about any subject they liked, and the Columbia Games Corpus [Gravano and Hirschberg, 2009], where participants located in different rooms played a collaborative computer game where they were able to communicate verbally.

### **7.2.3 Robots Backchanneling in Human-Robot Interactions**

HRI researchers are increasingly incorporating backchanneling behaviors in human-robot interactions in order to increase the quality of communicative interactions and to encourage positive behavior from the humans with which they interact [Lala et al., 2017, Ramachandran et al., 2018]. For example, Ramachandran et al. (2018) designed a tutoring robot to display the nonverbal backchannel of head nodding while a child responded to one of the robot’s prompts. Other work has demonstrated the utility of backchannels from a robot in human-robot collaborative teaming. Jung et al. (2013) demonstrated that the presence of robot backchannels led to improved team functioning, where the presence of robot backchanneling was correlated with increased performance (decreased reaction time) as well as reduced human stress and increased perceptions of responsiveness in high complexity tasks [Jung et al., 2013].

Work in HRI has also focused on developing computational models to determine when a robot should produce backchannels [Lala et al., 2017, Lee et al., 2019]. Lala et al. (2017) used a simple logistic regression model, trained on data from human-human interactions, to predict when an android robot should backchannel. They found that people viewed the

robot as more empathetic and natural when it used the backchannel production model continuously, as opposed to only after pauses in speech [Lala et al., 2017]. Lee et al. (2019) designed an attentive listening behavior generation model for a robot in a child storytelling context. They used a partially observable Markov decision process (POMDP) to model the child storyteller’s use speaker cues to infer listener attentiveness and a dynamic Bayesian network (DBN) to select the robot’s listening response. This approach to generating nonverbal backchannels was shown to be more effective than an approach based on signaling [Lee et al., 2019].

## 7.3 Methods

In this section, we describe a human subjects experiment designed to explore 1) the relationship between backchannels and social group dynamics and 2) the influence of a robot’s behavior on the backchannels expressed by human group members and their ratings of social group dynamics. We extend the experiment design described in Chapter 6 by introducing a second between subjects design factor of whether or not the robot provides verbal support to the human participants.

### 7.3.1 Hypotheses

Based on the role of backchanneling in active listening and prior work that has demonstrated a link between team performance and satisfaction with increased backchanneling behavior [Jung et al., 2012], we hypothesize:

- **Hypothesis 1:** Individuals that receive more backchannels (verbal and nonverbal) will report higher scores of both psychological safety and inclusion.
- **Hypothesis 2:** Groups that produce more backchannels (verbal and nonverbal) will have members that report higher scores of both psychological safety and inclusion.

Prior work in HRI, including our own work discussed in Chapters 3, 5, and 6, has demonstrated that a social robot’s behavior has the ability to significantly shape both human-robot

team social dynamics [Short and Matarić, 2017] and human behavior [Tennent et al., 2019] within human-robot teams. In this work, we hypothesize:

- **Hypothesis 3:** Robot verbal support will result in higher amounts of backchanneling between the human members of a human-robot team.
- **Hypothesis 4:** Robot verbal support will result in higher psychological safety and perceived inclusion scores.

In this work, we test hypotheses 1 and 2 by annotating the backchannels that occur between people working on a collaborative task. We test hypotheses 3 and 4 by varying the verbal support a robot in a human-robot team gives human team members while working on a collaborative task.

### 7.3.2 Experiment Design

In this work, we extend the experiment design presented in Chapter 6, by adding another between subjects dimension of whether or not the robot provides verbal support to the participants. This change transforms the 2 (robot liaison: insider or outsider) x 1 (robot verbal support: present) between subjects design in Chapter 6 to a 2 (robot liaison: insider or outsider) x 2 (robot verbal support: present or absent) between subjects design.

The experiment design in Chapter 6 involves the formation of an ingroup and outgroup through two rounds of a collaborative task, see Figure 6.2 in Chapter 6 for a pictorial description. In the first round, which lasted 15 minutes, two participants and a Jibo robot worked together on a task in room A. In room B, the third participant and a Jibo robot completed the same task. This first round was designed to form an ingroup, consisting of the two participants in room A, and an outgroup, the one participant in room B. Then for the second round, which lasted for 30 minutes, the outgroup participant was brought into room A to join the two ingroup participants and the Jibo robot. All three participants and the robot worked together to complete the round two task. During the second round, one participant was designated as the ‘robot liaison.’ The robot liaison’s role was to ask the robot for task relevant information and no other team member could ask the robot for this information. Thus, we present the first dimension of the 2 x 2 between subjects design:



- **Ingroup Robot Liaison:** An ingroup member is designated as the robot liaison.
- **Outgroup Robot Liaison:** An outgroup member is designated as the robot liaison.

This experimental dimension allows us to examine the backchanneling behavior of human team members towards those who are marginalized (the outgroup member) as well as those who are given a specialized role within a team (the robot liaison). We are also able to examine the influence of being an outgroup member and having a specialized role on participants' ratings of team social dynamics.

In order to study the influence of robot verbal support on the backchanneling behavior and social team dynamic ratings of participants, we introduce a second dimension of the 2 x 2 between subjects design:

- **Robot Verbal Support:** In addition to responding to queries for information, the robot makes verbal statements to support the input of human team members. This verbal support includes backchannel utterances (e.g., “*yeah*”), relevant informational hints about the task, and targeted supportive utterances (e.g., “*We should bring the key, good idea Jessica!*”). In Chapter 6, both the ingroup and outgroup robot liaison conditions included robot verbal support.
- **No Robot Verbal Support:** The robot responds to queries for information (e.g., from the robot liaison in round 2 of the experiment), but does not make any other verbal utterances. This is the new dimension of the experimental design introduced in this chapter.

We combined these two distinct between subject condition variations in a 2 (robot liaison: insider or outsider) x 2 (robot verbal support: present or absent) between subjects design, as shown in Figure 7.2. The participants of the human subjects study we described and ran in Chapter 6 experienced the following two conditions: ingroup robot liaison & robot verbal support and outgroup robot liaison & robot verbal support. In this chapter, we compare the data from these two conditions with two new conditions that have no robot support: ingroup robot liaison & no robot verbal support and outgroup robot liaison & no robot verbal support.

Robot Liaison: Ingroup or Outgroup	
Robot Verbal Support: Present or Absent	<i><b>Ingroup</b></i> Robot Liaison & Robot Verbal Support
	<i><b>Outgroup</b></i> Robot Liaison & Robot Verbal Support
Robot Verbal Support: Present or Absent	<i><b>Ingroup</b></i> Robot Liaison & <i><b>No</b></i> Robot Verbal Support
	<i><b>Outgroup</b></i> Robot Liaison & <i><b>No</b></i> Robot Verbal Support

Figure 7.2: In this experiment, we employed a 2 (robot liaison: ingroup or outgroup) x 2 (robot verbal support: present or absent) between subjects design.

### 7.3.3 Collaborative Task

Participants in our experiment collaborated with one another and the robot to complete a modified version of the Desert Survival Problem [Lafferty and Pond, 1974]. This task had two rounds. In the first round, participants were asked to rank 25 common household items with respect to how useful they are for survival. Then, in the second round, participants are given information about the environment where they are to be stranded and are tasked with selecting and ranking 8 items from the list of 25 items. More details about this task can be found in Chapter 6, Section 6.3.2.

### 7.3.4 Robot Behavior

We used the commercial robot Jibo for this experiment [Jibo, 2017]. Jibo is 11 inches tall and has a 3-axis motor system and a touchscreen face. We enabled Jibo to respond verbally to the participant utterances by capturing the participant’s audio through individual microphones and using Google’s speech-to-text API\* to acquire the spoken text.

During the collaborative task, the robot made several different types of verbal utterances: query responses, targeted supportive utterances, survival item hints, and backchannels. In

---

\*<https://cloud.google.com/speech-to-text/docs/libraries>

the conditions where the robot *did not* give verbal support, the robot made only query responses. In the conditions where the robot *did* give verbal support, the robot made all of the aforementioned verbal utterances. Here, we describe each type robot utterance and for more details, please refer to Appendix A, Section A.5.

Querying the robot was essential for participants to have the full information to complete the task. Participants could query the robot about the survival items in rounds 1 and 2 as well as aspects about the environment during round 2 using the language, “hey Jibo, tell me about the \_\_\_\_\_,” where Jibo is the name of the robot. During round 2, only the participant who was designated as the robot liaison could query the robot. Queries about the survival items gave participants more detailed information about the quantity and type of the item, for example, when queried about the chocolate the robot responded with, “*this box comes with 16 bars of 17.6 ounces Trader Joe’s chocolate. Each bar is wrapped in tinfoil and then with paper.*” Queries about environment aspects provided participants with information that was designed to stimulate conversation and have participants question prior assumptions they may have made about the location in which they were stranded. For example, the robot responded to being queried about the geography with: “*The whole area is one big mountain range. Some of the mountains might be covered in snow while others are more temperate and covered with grass. You may come upon some caves and lowlands as well.*”

We programmed the robot to deliver six targeted supportive utterances to each participant during the second round of the task. We designed the targeted supportive utterances to reinforce ideas and viewpoints of specific participants and be personal by including the participant’s name. Targeted supportive utterances either rephrased what a participant said (e.g., “*We need a coffee pot, good idea Samantha*”), included an item that a participant had mentioned (e.g., “*Camera. Robert, I think that’s worth considering*”), or gave general support for the participant (e.g., “*Okay, Jason*”).

After hearing a participant mention a survival item name, outside the context of a query to the robot, we had the robot deliver a useful hint about the survival item with probability 0.0375. The survival item hints were designed to encourage the participants to consider alternative uses, for example, the robot’s hints about the garbage bag included: “*a garbage*

*bag can be used as a sleeping bag*” and *“garbage bags can collect rain water.”*

Lastly, after hearing a participant utterance, the robot sometimes responded with a backchannel utterance. If the participant’s speech contained the mention of one of the survival items, the robot responded with an item backchannel (e.g., *“balloon, that makes sense,” “key, uh huh”*) with probability 0.0375. If the participant’s speech did not contain a survival item, the robot responded with a generic backchannel (e.g., *“yeah,” “interesting,” “hmm”*) with probability 0.075.

### 7.3.5 Protocol

For each experimental session, we recruited three human participants. After the three participants arrived, they each completed a consent form and then filled out a pre-experiment questionnaire on a tablet. To get the participants set up for the first round of the task, the experimenter took the outgroup participant to room B with one Jibo robot, ensured that the participant could query the robot properly, and initiated the robot’s introduction to round 1. While the experimenter was setting up the outgroup participant, the two ingroup participants were instructed to ask one another questions from a list of get-to-know-you questions in order to further reinforce the ingroup-outgroup divide (e.g., “If you didn’t sleep, what would you do with your extra time?”). After the experimenter had set up the outgroup participant, the experimenter paused the ingroup members and set them up in room A similarly to the outgroup participant.

After the 15 minutes of round 1 had expired, the experimenter brought the outgroup participant into room A with the two ingroup participants and the Jibo robot for round 2 of the collaborative task. The experimenter then designated one of the three participants as the robot liaison using the language, “In this part, unlike the first, only one of you will be able to ask Jibo questions about the items and environment. For all of you this is [participant name].”

Round 2 concluded after 30 minutes. After that, the experimenter brought the three participants out of room A and administered the post-experiment questionnaire to the participants on tablets. After each participant completed their post-experiment questionnaire, they were compensated with \$10.

### 7.3.6 Measures

In order to test the relationships between participant backchanneling behavior, their ratings of the team’s social dynamics, and the robot’s verbal support, we detail the questionnaire measures we administered to the participants as well as our annotation of the participant backchannels.

#### Controls

In the pre-experiment survey, participants evaluated their prior familiarity with the other two human participants in the group, their extraversion according to the abbreviated version of the Revised Eysenck Personality Questionnaire (EPQR-A) [Francis et al., 1992], and their emotional intelligence according to the Short Form of the Trait Emotional Intelligence Questionnaire (TEIQue-SF) [Cooper and Petrides, 2010]. Please refer to Chapter 6, Section 6.3.5 for more information on the administration of the questionnaires and Appendix B, Sections B.5, B.6, and B.8, for the full questionnaires and details on how each score was calculated.

#### Perceived Group Dynamics

In the post-experiment survey, participants completed the Perceived Group Inclusion Scale (PGIS) [Jansen et al., 2014], a 16 item scale that asked participants to evaluate items such as “this group gives me the feeling that I belong” and “this group encourages me to be authentic” on a Likert scale from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*), please refer to Appendix B, Section B.9 for more details.

Participants also filled out the Team Psychological Safety Scale [Edmondson, 1999], a 7 item scale where participants rated their agreement to statements like “it is safe to take a risk on this team” and “members of this team are able to bring up problems and tough issues” on a Likert scale from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*), please refer to Appendix B, Section B.7 for more details.

For the rest of this chapter, we refer to participants’ ratings on the Perceived Group Inclusion Scale and Team Psychological Safety Scale as their perceived inclusion scores and

their psychological safety scores, respectively. Although both measures are ‘perceived,’ we use the terms perceived inclusion and psychological safety because they are shortened names of the published scale titles.

## Human Speech

Each human participant wore a headset microphone throughout the experiment. The participants’ audio data was transcribed using Google’s speech-to-text API<sup>†</sup>. During the experiment, we fed the Google speech-to-text transcripts to the tablet controlling the robot in order to allow the robot to respond to participants’ speech. Additionally, we stored these speech-to-text transcriptions as well as the start time and duration of the speech, so that we could measure how much time each participant spent talking over the course of the experiment.

## Human Backchannels

In order to analyze the backchanneling behavior of the human participants, we transcribed and categorized each backchannel made by the participants during round 2 of the experiment using the ELAN software [Wittenburg et al., 2006]. We found the following definition from Ward and Tsukahara (2000) helpful in discerning between backchannels and non-backchannel utterances: “backchannel feedback 1) responds directly to the content of an utterance of the other, 2) is optional, and 3) does not require acknowledgement by the other.” Each backchannel was categorized as either verbal (e.g., “okay,” “mm hmm,” “yeah yeah”) or nonverbal (e.g., head nodding, head shaking). Each backchannel was also annotated with a recipient, indicating to whom the backchannel was directed towards.

Four coders contributed to the identification and categorization of human backchannels in the data. Inter-rater reliability was assessed by examining the agreement of the coders on candidate backchannels (a high Cohen’s kappa value of 0.90) and on the backchannel recipient (a high Cohen’s kappa value of 0.93). Participants on average produced 27.06 ( $SD = 21.41$ ) nonverbal backchannels and 30.67 ( $SD = 15.85$ ) verbal backchannels during

---

<sup>†</sup><https://cloud.google.com/speech-to-text/docs/libraries>

the 28 annotated minutes of round 2<sup>‡</sup>.

### 7.3.7 Participants

Participants were recruited for this study from a high school program held at Yale University. The students from the program came from 80 different countries, with 47% from the United States. The breakdown by continent is: 52% from North America, 24% from Asia, 12% from Africa, 7% from Europe, 3% from South America, and 2% from Australia.

A total of 40 groups (120 participants) were recruited for participation in this study. Of the 40 groups recruited, 2 groups were excluded from this analysis due to either not finishing the experimental task or non-compliance (e.g., removing their microphones). For the 38 remaining groups (78 participants), 58 participants were female and 56 participants were male. The average age of participants was 16.73 ( $SD = 0.73$ ). More detailed descriptive statistics for participants in each condition as well as each division of participants (ingroup/outgroup, robot liaison) can be found in Appendix C Tables C.49, C.50 and C.51.

## 7.4 Results

We used linear mixed-effects models in the analysis of our data in order to account for participants being in groups of three. We set the variables related to our experimental manipulations as fixed effects: intergroup bias (ingroup or outgroup), robot liaison designation (yes or no), verbally supportive robot (yes or no), and interactions between those variables when appropriate. We also set relevant covariates as fixed effects: gender, extraversion, emotional intelligence, and familiarity with other human team members. We set the participant's group as a random effect (random intercept) and relevant covariates as fixed effects. We tested these models for multicollinearity (variance inflation factor), selected them based on the Akaike information criterion, and evaluated residual errors for lack of trends and heteroscedasticity. For each fixed effect, the model outputs the linear coefficient ( $c$ ), the standard error ( $SE$ ), and the significance ( $p$ ) value of that predictor.

---

<sup>‡</sup>We did not annotate the first 2 minutes of round 2 because the experimenter did not leave the room until about a minute after round 2 began. To eliminate all influence of the experimenter's interaction with the participants, we annotate the data from exactly 2 minutes after round 2 started to its conclusion 28 minutes later.

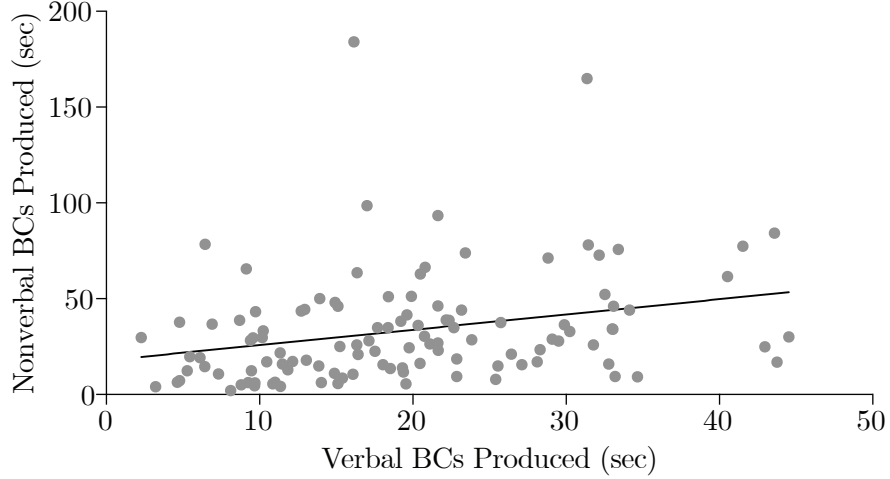


Figure 7.3: In this graph, each data point represents one participant and its x and y values represent the total time that participant spent producing verbal backchannels and nonverbal backchannels, respectively. We also plotted a line of best fit, which has a slope of 0.80.

When analyzing data where each data point represented one group of three participants, we used an analysis of variance (ANOVA). We investigated the influence of effects of inter-group bias (ingroup or outgroup), robot liaison designation (yes or no), verbally supportive robot (yes or no), and several covariates on our dependent variables of interest. We report the effect size as partial eta squared ( $\eta^2$ ). For more details on the results of the statistical models included in this section, please refer to Appendix C, Tables C.49 - C.70.

In some of the following analyses, we use the total amount of time participants spent talking. This value was computed by summing the utterance length timings captured by the participant microphones. Due to missing or inaccurate utterance timing values (e.g., because a participant’s microphone got disconnected during the experiment), we exclude 8 participants (2 groups of 3 participants and 2 individuals from different groups) from analyses that include the total time participants spent talking.

#### 7.4.1 Verbal and Nonverbal Backchannels

As we show later on in this analysis, verbal and nonverbal backchannels had distinct correlations with participant ratings of team social dynamics, indicating that their expression and reception by others have different effects. Therefore, in this analysis we treat them as separate categories of backchannel expression.



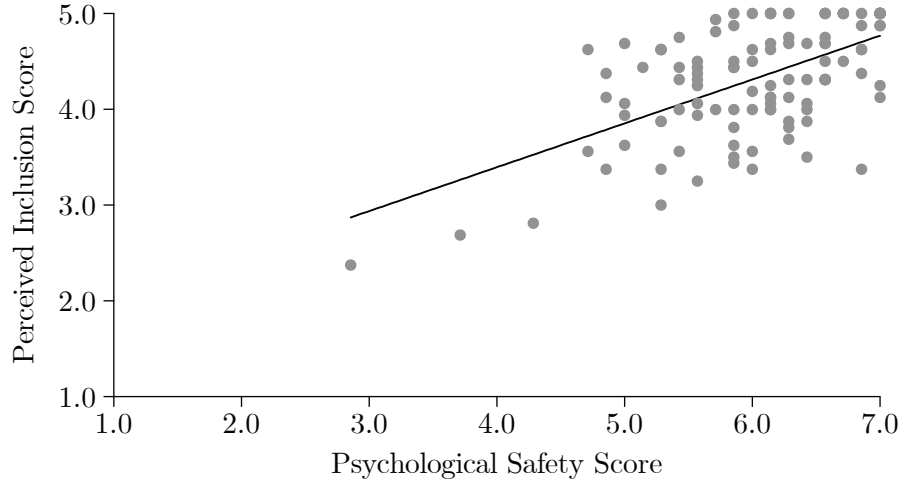


Figure 7.4: In this graph, each data point represents one participant and its x and y values represent the participant’s psychological safety and perceived inclusion scores, respectively. These scores were significantly and positively correlated, as shown by the positive slope (0.46) of the best fit line.

When comparing participants’ verbal and nonverbal backchanneling behavior in the experiment, we found that participants spent more time producing nonverbal backchannels ( $M = 33.37, SD = 28.87$ ) than verbal backchannels ( $M = 19.47s, SD = 10.01$ ). Additionally, we observed that the time that participants spent producing verbal and nonverbal backchannels was significantly correlated ( $r = 0.28, t = 3.05, p = 0.003$ ). As participants spent 1 more second producing verbal backchannels, they spent on average 0.80 more second producing nonverbal backchannels, which was determined by examining the slope of the best fit line depicted in Figure 7.3.

#### 7.4.2 Psychological Safety and Perceived Inclusion Scores

When considering our two measures of team social dynamics, we found that participants’ psychological safety ( $M = 6.02, SD = 0.74$ ) and perceived inclusion ( $M = 4.32, SD = 0.58$ ) scores were significantly and positively correlated ( $r = 0.58, t = 7.62, p < 0.001$ ). On average, an increase of 1.0 on a participant’s psychological safety score was met with a 0.46 increase on the participant’s perceived inclusion score, represented by the slope of the best fit line shown in Figure 7.4.

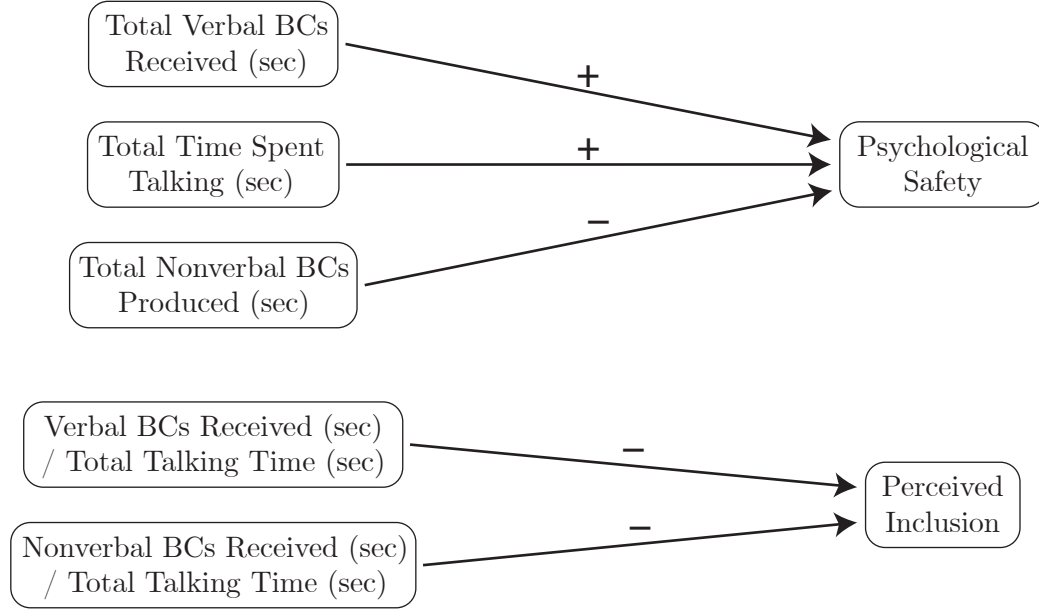


Figure 7.5: We display the human backchannel (BC) variables that have significant influences on psychological safety and perceived inclusion post-experiment survey scores. This analysis examined each individual participant’s backchanneling behavior and ratings of psychological safety and inclusion.

#### 7.4.3 Connections between Human Backchannels and Social Group Dynamics - Individual Level

In order to test our first hypothesis, that the amount of backchannels received positively correlates with team social dynamics, we first examine the influence of participant backchanneling behavior on participants’ psychological safety and perceived inclusion scores. The significant relationships between participant backchanneling and their ratings of psychological safety and inclusion are shown in Figure 7.5.

When investigating correlations between participants’ backchanneling behavior and their ratings of team social dynamics, we examined the influence of 1) verbal backchannels, 2) nonverbal backchannels, and 3) total backchannels (the sum of the verbal and nonverbal backchannels). In the majority of cases where the total backchannels did significantly correlate with participants’ ratings of team dynamics, the effect was driven by either verbal backchannels or nonverbal backchannels (and not both). Therefore, in this section, we treat verbal and nonverbal backchannels as separate signals and we do not combine them to report on the combined total amount of backchannels.

## Verbal Backchannels Received by an Individual

We first examined the influence of the amount of verbal backchannel an individual received on their ratings of team social dynamics. In our analysis of the influence of the total duration (sec) of the backchannels a participant received on their psychological safety score, the linear mixed-effects model that best fit the data had covariates of gender ( $c = 0.27, SE = 0.13, p = 0.033$ ) and emotional intelligence ( $c = 0.28, SE = 0.10, p = 0.006$ ). We found a significant positive influence of the backchannels a participant received (sec) on their psychological safety score ( $c = 0.017, SE = 0.006, p = 0.004$ ).

Since it could be possible that participants who spoke were more likely to receive more verbal backchannels, we next explored whether the total talking time of participants also influenced participants' psychological safety. The linear mixed-effects model that best fit the data had covariates of ingroup-outgroup bias ( $c = 0.28, SE = 0.15, p = 0.060$ ), robot liaison designation ( $c = -0.24, SE = 0.15, p = 0.113$ ), gender ( $c = 0.32, SE = 0.14, p = 0.021$ ), and emotional intelligence ( $c = 0.026, SE = 0.11, p = 0.021$ ). We found a significant positive effect of the participant's total talking time (sec) on their psychological safety score ( $c = 0.00092, SE = 0.00041, p = 0.029$ ).

Next, we analyzed whether a normalized version of the verbal backchannels a participant received influenced the psychological safety scores of participants by examining the total duration of the backchannels a participant received divided by the total time the person spent talking, representing the proportion of the time a participant spent talking when they were being backchanneled (verbally). The linear mixed-effects model that best fit the data had covariates of ingroup-outgroup bias ( $c = 0.32, SE = 0.16, p = 0.049$ ), robot liaison designation ( $c = -0.23, SE = 0.15, p = 0.128$ ), gender ( $c = 0.32, SE = 0.14, p = 0.023$ ), and emotional intelligence ( $c = 0.28, SE = 0.11, p = 0.011$ ). There was no significant influence found on the proportion of time a participant spent talking while they were being verbally backchanneled ( $c = -1.33, SE = 0.58, p = 0.287$ ).

From these results, we observe that the total volume of verbal backchannels received by a participant as well as the total time they spent talking positively correlate with their psychological safety scores, which shows support for our first hypothesis. However, the

proportion of time they spent talking where others were backchanneling them did not show a significant correlation with the participant's psychological safety scores. This indicates that it is the *volume* of backchannels, and not the proportion of the participants' speech that was backchanneled, that predicted their psychological safety scores. It is important to also note that the volume of backchannels a participant receives may be due to the fact that they are talking more, and that both a participant's time spent talking and verbal backchannels received might positively reinforce each other.

Although the total amount of verbal backchannels a participant received was significantly correlated with their psychological safety score, the amount of verbal backchannels a participant received was not significantly correlated with their perceived inclusion score ( $c = 0.006, SE = 0.005, p = 0.174$ ). Additionally, the amount of nonverbal backchannels that a participant received was neither correlated with their psychological safety score ( $c = 0.002, SE = 0.002, p = 0.452$ ) nor their perceived inclusion score ( $c = 0.0002, SE = 0.002, p = 0.403$ ).

### **Nonverbal Backchannels Produced by an Individual**

In addition to the backchannels received by participants, we examined the amount of backchannels produced by participants. We examined the influence of the total time a participant produced nonverbal backchannels on their psychological safety score. The linear mixed-effects model that best fit the data had covariates of gender ( $c = 0.32, SE = 0.13, p = 0.016$ ) and emotional intelligence ( $c = 0.35, SE = 0.10, p < 0.001$ ). We found a significant negative effect of the total time a participant produced nonverbal backchannels on participants' psychological safety scores ( $c = -0.0047, SE = 0.0023, p = 0.041$ ), meaning that participants who produced a lot of nonverbal backchannels had lower psychological safety ratings than those who did not express as many nonverbal backchannels.

We did not find a significant relationship between the amount of nonverbal backchannels produced and participants' perceived inclusion scores ( $c = 0.0009, SE = 0.002, p = 0.614$ ). We also did not find significant correlations between the amount of verbal backchannels produced by a participant and either their psychological safety ( $c = -0.003, SE = 0.007, p = 0.626$ ) or perceived inclusion scores ( $c = 0.010, SE = 0.005, p = 0.066$ ).

### Backchannels Received Normalized by Total Talking Time

We analyzed the influence of the time a participant received both nonverbal and verbal backchannels normalized by the total time that a participant spent talking on the participant's perceived inclusion score. With respect to the nonverbal backchannels, the linear mixed-effects model that best fit the data included covariates of ingroup-outgroup bias ( $c = 0.27, SE = 0.12, p = 0.023$ ), robot liaison designation ( $c = -0.34, SE = 0.11, p = 0.004$ ), gender ( $c = 0.15, SE = 0.10, p = 0.153$ ), emotional intelligence ( $c = 0.20, SE = 0.08, p = 0.019$ ), and familiarity with other team members ( $c = 0.058, SE = 0.034, p = 0.091$ ). We found a significant negative effect of time others spent nonverbally backchanneling a participant normalized by their total time spent talking on participant perceived inclusion scores ( $c = -0.73, SE = 0.30, p = 0.016$ ). With respect to the verbal backchannels, the linear mixed-effects model that best fit the data included covariates of ingroup-outgroup bias ( $c = 0.29, SE = 0.12, p = 0.020$ ), robot liaison designation ( $c = -0.36, SE = 0.12, p = 0.003$ ), emotional intelligence ( $c = 0.23, SE = 0.08, p = 0.008$ ), and familiarity with other team members ( $c = 0.068, SE = 0.034, p = 0.050$ ). Similar to the nonverbal backchannels, we found a significant negative effect of the time others spent verbally backchanneling a participant normalized by their total time spent talking on participant perceived inclusion scores ( $c = -1.67, SE = 0.80, p = 0.040$ ).

We also examined whether the total time participants spent talking during round 2 of the experiment had an effect on their perceived inclusion score. The linear mixed-effects model that best fit the data had covariates of ingroup-outgroup bias ( $c = 0.22, SE = 0.12, p = 0.061$ ), robot liaison designation ( $c = -0.33, SE = 0.12, p = 0.005$ ), emotional intelligence ( $c = 0.22, SE = 0.085, p = 0.010$ ), and familiarity with other team members ( $c = 0.057, SE = 0.036, p = 0.115$ ). We did not find a significant influence of the total time a participant spent talking on their perceived inclusion score ( $c = 0.00052, SE = 0.00033, p = 0.121$ ).

It is interesting to consider these findings together with the results that indicate that the amount of backchannels received is positively correlated with psychological safety. Although backchanneling another person in a group may raise their psychological safety, backchannel-

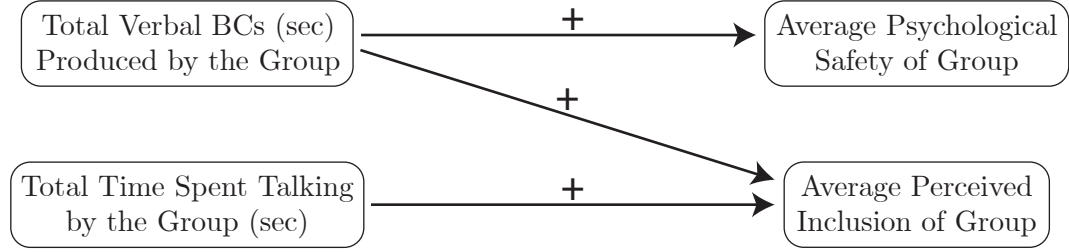


Figure 7.6: We display the human backchanneling variables that have significant influences on the group’s average psychological safety and perceived inclusion post-experiment survey scores. This analysis examined the backchanneling behavior of each group and the average ratings of psychological safety and inclusion for each group.

ing them too much relative to their talking time may result in reduced perceived inclusion. Therefore, it seems that to positively influence both the psychological safety and perceived inclusion of another person within a group, one needs to backchannel them just the right amount. Thus, we find partial support for our first hypothesis.

#### 7.4.4 Connections between Human Backchannels and Social Group Dynamics - Group Level

We now consider our second hypothesis, which predicts that the amount of backchannels that a group produces towards one another positively influences the group’s rating of psychological safety and inclusion. To conduct this analysis, we considered each group as one data point. We computed the total amount of backchanneling that occurred in each group and we averaged their psychological safety and perceived inclusion scores. The significant relationships between the backchannels that a group produces and its members’ average psychological safety and average perceived inclusion scores are displayed in Figure 7.6.

##### Verbal Backchannels Produced by the Group

We found that the time participants within each group spent verbally backchanneling one another had a significant and positive influence on both groups’ average perceived inclusion scores,  $F(1) = 9.72, \eta^2 = 0.11, p = 0.004$ , as well as groups’ average psychological safety scores,  $F(1) = 6.17, \eta^2 = 0.038, p = 0.019$ .

We also found that the total time participants spent talking had a significant and positive

effect on groups' average perceived inclusion scores,  $F(1) = 7.32, \eta^2 = 0.052, p = 0.012$ , however, did not have a significant effect on groups' average psychological safety scores,  $F(1) = 0.45, \eta^2 = 0.0023, p = 0.510$ .

Although the total amount of verbal backchannels produced by a group was positively predictive of their team social dynamics ratings, we did not find that the proportion of the time the group spent producing verbally backchanneling (normalized by the total time they spent talking) predicted their psychological safety,  $F(1) = 0.82, \eta^2 = 0.028, p = 0.374$ , or perceived inclusion scores,  $F(1) = 0.14, \eta^2 = 0.001, p = 0.711$ . Additionally, we did not find significant correlations between the time participants spent producing nonverbal backchannels one another and their ratings of psychological safety,  $F(1) = 0.06, \eta^2 = 0.005, p = 0.813$ , or perceived inclusion,  $F(1) = 3.48, \eta^2 = 0.085, p = 0.072$ .

These results indicate a similar finding to what we observed with individuals: the total volume of verbal backchannels a participant received, not the proportion of verbal backchannels received relative to the participant's talking time, correlates with more positive perceptions of team social dynamics. Though, with individuals this relationship was only significant between the verbal backchannels a person received and their psychological safety score, groups who had more verbal backchannels towards one another had higher psychological safety scores *and perceived inclusion scores*. Therefore, we do find support for our second hypothesis.

#### 7.4.5 The Influence of Intergroup Bias on the Reception of Backchannels

In addition to determining how human team member backchannels influence group dynamics, we were interested in investigating whether intergroup bias influenced their backchanneling behavior. We found that participants received more verbal backchannels ( $c = 6.67, SE = 1.91, p < 0.001$ ) and nonverbal backchannels ( $c = 8.85, SE = 4.40, p = 0.048$ ) if they were an outgroup member as opposed to an ingroup member. Outgroup members also received a higher proportion of both verbal backchannels ( $c = 0.036, SE = 0.012, p = 0.004$ ) and nonverbal backchannels ( $c = 0.083, SE = 0.035, p = 0.021$ ) relative to their total talking time than ingroup members. The observation that outgroup members receive more backchannels may tell us more about the function of backchannel utterances themselves. Since the

outgroup participant had not interacted with the two ingroup participants before round 2 of the experiment, increased backchanneling may be one way that ingroup members try to welcome the outgroup member and help them to feel comfortable in the group.

#### **7.4.6 The Influence of Gender on the Production of Backchannels**

We were also interested in examining whether any descriptive features of the participants shaped participants' backchanneling behavior (age, gender, emotional intelligence, familiarity with other participants). In our analysis of the verbal and nonverbal backchannels produced and received by the people in these human-robot collaborative teams, we noticed that the factor of gender was influential in backchannel production. We discovered that females produced more verbal backchannels than males ( $c = 7.13, SE = 1.70, p < 0.001$ ). Females, compared with males, also had a higher proportion of verbal backchannels produced relative to the total time they spent talking ( $c = 0.059, SE = 0.028, p = 0.041$ ). When examining the effects of gender at the group level, we similarly found that groups with more females produced more verbal backchannels, both in total volume,  $F(1) = 18.61, \eta^2 = 0.32, p < 0.001$ , as well as normalized with respect to the group's total talking time,  $F(1) = 9.97, \eta^2 = 0.44, p = 0.004$ . These findings are consistent with prior work in psychology, that has also observed increased backchanneling behavior in females as compared to males [Duncan and Fiske, 2015, Roger and Neshoever, 1987]. These results also connect with Woolley et al. (2010)'s finding that the number of females on a team positively correlates with the team's collective intelligence.

#### **7.4.7 Influence of Robot Verbal Support on Human Backchanneling Behavior**

In order to examine our third hypothesis, that robot verbal support will increase backchanneling behavior between team members, we examined the influence of the presence of robot verbal support on the amount of verbal backchannels participants received during round 2 of the experiment. The linear mixed-effects model that best fit the data had covariates of extraversion ( $c = 1.35, SE = 0.49, p = 0.007$ ) and familiarity with other human team members ( $c = 0.97, SE = 0.65, p = 0.139$ ). We found a significant in-



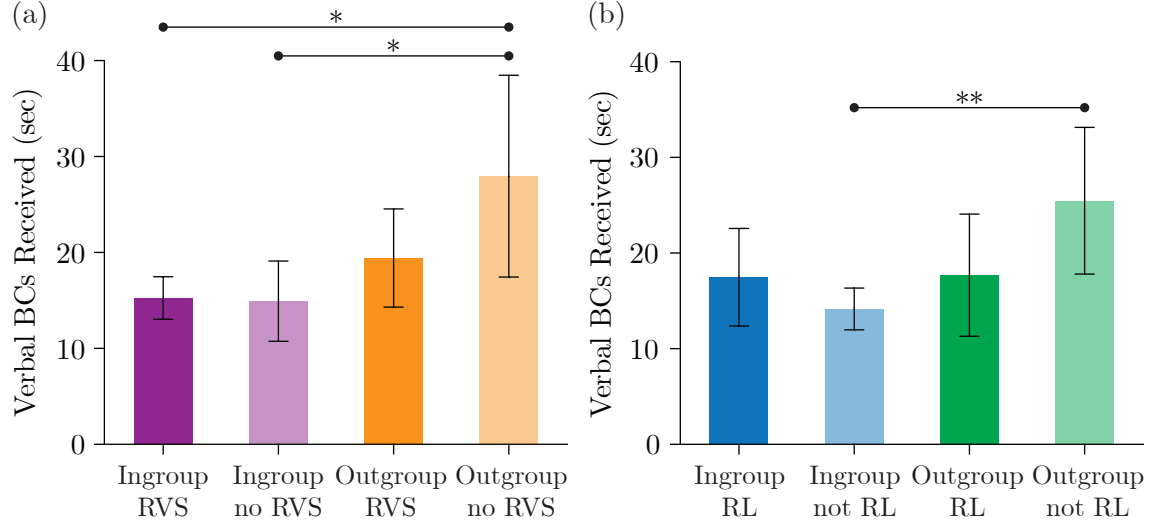


Figure 7.7: In our analysis of the influence of our experimental factors on the verbal backchannels that participants received, we found a main effect for intergroup bias, where outgroup members received more verbal backchannels than ingroup members, and two significant interactions. (a) The first interaction we observed was between intergroup bias and the presence of robot verbal support (RVS). (b) The second interaction we found was between intergroup bias and whether or not the participant was a robot liaison (RL). (\*) and (\*\*) denote  $p < 0.05$  and  $p < 0.01$  respectively. Error bars represent a 95% confidence interval.

fluence of the intergroup bias on the amount of verbal backchannels a participant received, as we mentioned in the prior section, where outgroup members were found to receive more verbal backchannels ( $M = 21.68s, SD = 13.92s$ ) than ingroup members ( $M = 15.17s, SD = 8.39s, c = 16.85, SE = 4.09, p < 0.001$ ).

We also found a significant interaction between intergroup bias and the presence of verbal support from the robot ( $c = -8.83, SE = 4.33, p = 0.045$ ), shown in Figure 7.7(a). Using post-hoc comparisons using Tukey-adjusted estimated marginal means, we found that outgroup members with no robot verbal support received significantly more verbal backchannels ( $M = 27.96s, SD = 14.71s$ ) than both ingroup members with robot verbal support ( $M = 15.26, SD = 8.28, c = -10.72, SE = 3.44, p = 0.013$ ) and ingroup members with no robot verbal support ( $M = 14.92, SD = 8.93, c = -11.29, SE = 3.78, p = 0.019$ ). No other comparisons significantly differed from one another.

This interaction effect seems to be primarily driven by the higher amount of backchanneling received by the outgroup member in groups with no robot verbal support (RVP),

visualized in Figure 7.7(a). Without the presence of robot verbal support outgroup members received significantly more backchannels than ingroup members. When there is robot verbal support present, there is no significant difference between the backchannels received by ingroup and outgroup members. This result is opposite to what we hypothesized, and may suggest that when the robot exhibits more verbal support, the group may not see the need to backchannel quite as much, especially to the outgroup member.

Lastly, we found a significant interaction between intergroup bias and whether the robot liaison was an outgroup or ingroup member ( $c = -11.12, SE = 4.14, p = 0.009$ ), shown in Figure 7.7(b). Using post-hoc comparisons using Tukey-adjusted estimated marginal means, we found that outgroup members who were not the robot liaison received more verbal backchannels ( $M = 25.49s, SD = 14.45s$ ) than ingroup members who were not the robot liaison ( $M = 14.55s, SD = 7.93s, c = -12.43, SE = 2.82, p = 0.001$ ). No other comparisons had significant differences between them.

This interaction effect between the robot liaison (RL) designation and intergroup bias (ingroup or outgroup), as displayed in Figure 7.7(b), seems to be driven by an increased amount of backchanneling received by the outgroup members who have not been designated as a robot liaison. This could be explained in a similar way to the interaction between robot verbal support and intergroup bias, where the group is inclined to backchannel the outgroup member more. However, when the outgroup member receives robot-related behavior (robot liaison designation or robot verbal support), the other group members do not perceive as large a need to backchannel them.

#### 7.4.8 Influence of Robot Verbal Support on Team Social Dynamics

To investigate our fourth hypothesis, that robot verbal support will enhance team social dynamics, we examined the effect of the presence of robot verbal support on the psychological safety and perceived inclusion scores of participants. We found a marginally significant effect of robot verbal support on participants' psychological safety scores ( $c = 0.35, SE = 0.18, p = 0.056$ ), where participants in groups with robot verbal support had higher psychological safety scores ( $M = 6.06, SD = 0.74$ ) than participants in groups without robot verbal support ( $M = 5.92, SD = 0.74$ ). We did not find any significant effect of robot

verbal support on participants' perceived inclusion scores ( $c = 0.12$ ,  $SE = 0.14$ ,  $p = 0.376$ ). From the lack of statistically significant differences in team social dynamics scores between conditions with and without robot verbal support, we were not able to show support for our fourth hypothesis.

## 7.5 Discussion

In this work, we explored 1) connections between human backchanneling behavior and their perceptions of team social dynamics and 2) how robot verbal support might shape the backchanneling behavior and perceptions of social group dynamics of human team members. We conducted a human subjects experiment where three human participants and a robot, that either did or did not exhibit verbal support, completed a collaborative task. We annotated instances of human participant backchannels and analyzed correlations between these backchannels with participants' ratings of team social dynamics as well as the robot's verbally supportive behavior.

We did find significant correlations in our data between human team members' backchanneling behavior and their post-experiment survey rating of both psychological safety. The most common theme in the data is a positive correlation between the amount of verbal backchannels and both psychological safety and inclusion. Individuals who received more verbal backchannels had higher ratings of psychological safety, and teams that produced more verbal backchannels had higher average ratings of both psychological safety and inclusion. It is important to highlight that this effect was not seen with nonverbal backchannels, indicating that verbal and nonverbal backchannels have different effects in collaborative teams. Additionally, the proportion of verbal backchannels a person receives relative to their talking time does not significantly correlate with psychological safety. This demonstrates that regardless of whether a person is very talkative or more reserved, it is the total amount of backchannels they receive that corresponds with their perceptions of the group's psychological safety.

In addition to the total amount of verbal backchannels, the total talking time of participants was also correlated with measures of psychological safety and inclusion. Specifically,

the total time a participant spent talking predicted their psychological safety score and the total verbal backchannels a team produced predicted the team's average perceived inclusion score. We think that it is likely that receiving verbal backchannels and time spent talking both positively influence one another. If one person receives more verbal backchannels, they might talk more. Also, if a person is talking more, they might be more likely to receive verbal backchannels.

Not all of the significant correlations we observed between participant backchanneling behavior and team social dynamics were positive. The total time a participant spent non-verbally backchanneling their human team members was negatively correlated with their psychological safety rating. Additionally, the proportion of both verbal and nonverbal backchannels received by a participant, relative to their talking time, was negatively correlated with their perceived inclusion score. These results suggest that the right balance of backchanneling is needed to support the inclusion and psychological safety of all team members. For example, although verbal backchannels have a positive relationship with psychological safety, too many verbal backchannels relative to their talking time may result in reduced perceived inclusion.

We also observed an influence of the verbal support from the robot on the verbal backchanneling behavior of participants. When a group interacted with a robot that did not give verbal support, the outgroup member, who completed the first phase of the experiment alone, received significantly more backchannels than the two ingroup members, who completed the first phase of the experiment with each other. However, when a group interacted with a robot that did give verbal support, there was no significant difference between the backchannels received by outgroup and ingroup members. We believe that rather than encouraging and promoting the verbal backchanneling behavior of participants, like we hypothesized, the verbal support from the robot instead replaced the verbal backchannels the ingroup members would have given the outgroup member. We saw a similar effect with the robot liaison designation, where outgroup members who were not robot liaisons received more backchannels than outgroup members who were robot liaisons, although this relationship was not statistically significant.

It is unclear what the effects were on the behavior and social dynamics of the human-

robot team as a result of the reduction in backchannels towards the outgroup member when the group had a verbally supportive robot. We did not find any significant differences between the outgroup member’s psychological safety or perceived inclusion scores when the robot did express verbal support compared with when the robot did not express verbal support. Thus, it is possible that outgroup members that had robot verbal support and those that did not have robot verbal support received the same amount of verbal backchannels, where the robot’s verbal support *replaced* the verbal backchannels the outgroup member would have received from their human team members. As robot verbal behavior is designed for future human-robot teams, it is important to consider how robot actions may replace behaviors that are typically expressed by human team members and the influence this might have on the team’s social dynamics.

## 7.6 Summary

As robots increasingly join collaborative teams of people, it is essential that they are able to sense and positively contribute to the social dynamics of the team in real time. The results from our human subjects study indicate that human team member backchanneling behavior is a promising signal to use for the successful prediction of team dynamics like psychological safety and inclusion. We found that the amount of verbal backchannels received by human team members is positively correlated with both psychological safety and perceived inclusion. However, too many backchannels given to an individual, relative to their speaking time, may have a negative effect on their perceived inclusion. This highlights the importance of giving each member “just the right” amount of backchanneling to optimize their psychological safety and inclusion. We also found that robot verbal support, including backchanneling, may replace as opposed to increase the verbal backchannels received by marginalized team members.

To our knowledge, this work is the first to explore the direct connections between human backchanneling behavior with established survey measures of team social dynamics (psychological safety [Edmondson, 1999] and perceived inclusion [Jansen et al., 2014]) that have been shown to positively influence team performance. We are also the first to show

differences in the effects of verbal and nonverbal backchannels on team social dynamics. For example, the amount of verbal backchannels, but not nonverbal backchannels, a person receives is significantly correlated with their psychological score. Lastly, we demonstrate that a social robot can influence the backchanneling behavior of a collaborative team, where verbal support from the robot was shown to reduce the verbal backchannels an outsider received from their human team members. These findings highlight the importance of considering how backchannels can be used to sense team social dynamics and also how the robot's own behavior may change a team's backchanneling behavior.

In the next chapter, we discuss all of the work presented in this dissertation. We expand upon the central themes of our work and highlight several areas of opportunity for future work in building robots teammates that enhance group dynamics and performance.

## Chapter 8

# Discussion

In this dissertation, we present a body of work that seeks to improve the performance of human-robot teams by shaping team social dynamics to promote inclusion, trust, and psychological safety. We conducted five human subjects experiments exploring the influence of a variety of robot behaviors (e.g., trust repair utterances, vulnerable statements, verbal support) on how people behave towards one another and how they perceive team social dynamics. We also demonstrate an important connection between human backchannels (e.g., “yeah”, head nodding) and questionnaire measures of team social dynamics, a step in the direction of being able to perceive team social dynamics in real time. All of the work detailed in Chapters 3 through 7 contributes to our understanding of how robot teammates can best be designed to promote positive social dynamics and human behavior within human-robot teams. In this chapter, we present design guidelines for social robot behavior in collaborative human-robot teams, review the central themes in our work, and present some open questions and directions for future work.

### 8.1 Design Guidelines for Social Robot Behavior in Collaborative Human-Robot Teams

In this section, we outline ten design guidelines for social robot behavior within collaborative human-robot teams. We generated these design guidelines with data and observations from the human subjects experiments detailed in this dissertation.

1. When a robot makes an error that effects a person, an effective way to repair trust with them is to frame the error as a mistake and make an apology (Chapter 4).
2. If a robot solicits a promise from a person, the person is likely to behave according to their promise, even when the robot displays untrustworthy behavior (Chapter 4).
3. In collaborative teaming contexts that are high-pressure or where members may fear making a mistake that could hurt the team, vulnerability from a robot can increase human-to-human team member trust-related behavior (e.g., explaining mistakes made) and conversation (Chapter 5).
4. Vulnerable utterances from a robot can also enhance human team member social interaction with the robot (Chapter 5).
5. Giving a human team member a specialized role to interact with a robot can reduce how included they feel within the team (Chapter 6).
6. When making decisions about which team members take on various roles and tasks within a team, it is best not to reinforce pre-existent faultlines (divisions in a team along a salient characteristic such as age or gender) to ensure that all team members' input is equally valued (Chapter 6).
7. Verbal support from a robot can increase the participation of marginalized or outsider team members (Chapter 6).
8. Backchannels from human team members are a useful signal in predicting team social dynamics (Chapter 7).
9. When a robot offers human team members verbal support, human team members are less likely to exhibit verbal support (e.g., verbal backchannels) to a marginalized or outsider team members (Chapter 7).
10. In order to capture the full effects of a robot's behavior on an individual or group, use a combination of survey and behavioral measures (Chapters 3 - 6).



## 8.2 Central Themes

In this dissertation, our work has focused on positively shaping team social dynamics and performance with robot behavior. In Chapters 3 through 7 we have explored the influence of robot behavior on team social dynamics and human team member behavior by conducting controlled, laboratory based human subjects experiments. In this section, we highlight and expand upon the central themes that have emerged from this work.

### 8.2.1 Robot Behavior Can Influence Human-to-Human Behavior within Human-Robot Collaborative Teams

The most significant contribution of the work described in this dissertation is that it is the first to demonstrate that a robot’s behavior can influence how people in the group interact *with each other*. Although prior work in HRI has shown that robots can shape how people behave towards a robot in a human-robot group (e.g., [Ball et al., 2017], [Mutlu et al., 2009], [Shiomi et al., 2010]) and how people perceive the social dynamics of a human-robot group (e.g., [Jung et al., 2015], [Short and Matarić, 2017]), ours was the first to present evidence that robots can influence *human-to-human behavior* within a human-robot team. This is most clearly seen in the human subjects experiment described in Chapter 5, where groups of three human participants were more likely to exhibit vulnerable behavior (explaining a mistake they had made, consoling others who made mistakes) and converse more among themselves if their team included a robot making vulnerable utterances, as opposed to a robot making neutral utterances or a robot that remained silent. Additionally, in Chapter 6, we found that outgroup team members, as opposed to ingroup team members, talked more to their human team members after receiving a verbally supportive utterance from the robot. Finally, in Chapter 7, we observed that outgroup human team members received less verbal backchannels from their human team members when their group contained a robot making supportive utterances.

These findings highlight the importance of understanding how our ever increasing social interactions with artificial agents, like robots, can shape our actions, conversations, and relationships with the people around us. Without an understanding of how robots can

influence how we interact with other people, we will likely experience negative and unexpected consequences of interacting with robots. For example, in the years of 2016-2018, the United States public became increasingly concerned with how voice assistants such as Siri, Alexa, Cortana, and Google, were influencing the politeness of the speech of children [Shellenbarger, 2018]. Since many voice assistants did not require good manners (e.g., children could command, “Alexa, tell me a joke,” rather than the more polite, “Alexa would you please tell me a joke?”) some children, as a result of talking to voice assistants, adopted rude manners that transferred into their conversations with people face-to-face. Our hope is that the work presented in this dissertation and similar work in HRI can help to inform the designers and programmers of robots and other artificial agents to promote positive, inclusive, and trusting interactions in future groups and teams.

Although the work in this dissertation and other recent work in HRI have consistently shown that a robot’s actions can shape human-to-human interactions within collaborative human-robot teams, there may exist limitations to this effect. As the group size increases (e.g., from 3 people and 1 robot to 100 people and 1 robot), it is likely that the robot’s influence on human-to-human interaction will decrease. Additionally, the influence of the robot may be moderated by *when* the robot joins the team. For example, if a robot joins a team after they have already been working together for several months the robot’s effect on human-to-human interactions will likely be weaker than if the robot had joined the team at the beginning when the team was forming and establishing group norms. Future work is needed to fully explore the limitations of and factors that moderate robot influence on human-to-human behavior within human-robot teams.

### **8.2.2 Robot Interactions with Groups and Teams of People**

The work in this dissertation has contributed to the increasing focus in HRI on developing robots that can seamlessly and intelligently interact with multiple people. Some work in HRI has examined how a robot’s physical movements such as navigation [Kidokoro et al., 2013, Mavrogiannis et al., 2019], physical orientation [Shiomi et al., 2010, Vázquez et al., 2017], gestures [Hoffman et al., 2015, Liu et al., 2013], and gaze [Mutlu et al., 2009, Skantze, 2017] can improve human-robot group interactions and influence people’s perceptions of the group.

Other work in HRI has explored how a robot’s verbal utterances that convey expressions of emotion [Correia et al., 2018b, Leite et al., 2012], informational content [Fernández-Llomas et al., 2017, Sabelli and Kanda, 2016], and the robot’s personality [Kanda et al., 2012, Oliveira et al., 2018] can shape people’s perceptions of the group and the robot as well as build social relationships between the robot and the people with whom it interacts. The work in this dissertation has uniquely contributed to the field of HRI a greater understanding of how robot verbal behavior can influence *human-to-human interactions* within a group. As a result of this and further research into robots’ physical movements, verbal expressions, and influence on human-to-human interactions, we are confident that the future robots we build will be productive members of human-robot groups and teams that positively contribute to both the group’s task output and its social dynamics.

### 8.2.3 Robot Perception of team social dynamics

Another significant contribution of this work is the identification and exploration of backchannels as a useful feature to predict team social dynamics. In order for robots to provide intelligent and in-the-moment responses to changes in group interactions, it is important that robots are able to sense team social dynamics in real time. The primary way that researchers currently measure a team’s social dynamics is by administering questionnaires to each human team member (e.g., the Team Psychological Safety Scale [Edmondson, 1999] and the Perceived Group Inclusion Scale [Jansen et al., 2014]). However, if a robot is to measure team social dynamics in real-time and in a non-obtrusive way, administering questionnaires is not a feasible approach. Instead, a robot could mathematically model and predict team social dynamics using features that are easy for the robot to sense (e.g., eye gaze, head movements, speech-to-text transcripts). However, no such models exist that perceive team social dynamics exist, to our knowledge. This lack of models is likely due to the lack of established connections between low-level behaviors such as eye gaze, head movements, and vocalizations and high-level team social dynamics. The work that we present in Chapter 7 reports statistically significant correlations between the team social dynamics of psychological safety and inclusion and human backchanneling behavior. In particular, the amount of verbal backchannels people in a group receive positively correlates with group members’

psychological safety and perceived inclusion. This work highlights human backchannels as a promising feature that could be used in mathematical models to predict team social dynamics in real time, especially since human backchannels can realistically be perceived (prior work has developed several successful computational models for predicting backchannels, e.g., [Gravano and Hirschberg, 2009], [Morency et al., 2010], [Truong et al., 2011]). Our work enables future real-time models of team social dynamics to be constructed, using human backchannels as an informative feature.

#### **8.2.4 Experiment Designs that Enable Investigation into Specific Team Dynamics**

The results reported in this dissertation were made possible through the careful design of our human-subjects experiments. Specifically, the experiment designs detailed in Chapters 5 and 6 are themselves significant contributions and contain examples of how the structure of a human-robot group can be engineered to enable the exploration of specific group behavior and dynamics.

In Chapter 5, we investigated the influence of a robot’s vulnerable utterances on the trust-related behavior of human team members and their conversational dynamics. This human-subjects experiment involved three human participants and one robot playing 30 rounds of a collaborative game together on individual tablets (see Figure 5.3). We designed the game such that if one member of the team failed to complete their individual task, the entire team would fail the round (see Figure 5.2). Additionally, in order to measure the human participants’ reactions to moments of tension, we engineered the game to force each player to fail their individual task twice over the 30 rounds, by making it impossible for them to complete their individual task. Thus, each team’s score at the end of the game was 22 successful rounds and 8 failure rounds.

In Chapter 6, we explored the influence of two strategies (a specialized role to interact with the robot and supportive verbal utterances from the robot) on the perceived inclusion of the human members of a human-robot team. We were especially interested in how effective these two strategies would be on human members of a group that feel excluded or like an outsider. The human-subjects experiment we designed involved two rounds and

three human participants (see Figure 6.2). In the first round, two participants and one robot collaborated together on a task in one room, and one participant and one robot collaborated with one another on the same task in a separate room. We designed this first round to create an ‘ingroup’ – the two participants in the first room, and an ‘outgroup’ – the single participant in the second room. In the next round, the outgroup participant was brought in the room to join the two ingroup participants and the robot to collaborate on the next part of the task. It was in this second round that we tested our two strategies to improve human team member inclusion. One of our strategies was giving one human team member a specialized role to interact with the robot. We accomplished this by allowing only one of the three human participants to query the robot for essential task-relevant information. The other strategy was supportive verbal utterances from the robot, which the robot produced during the second round, equally distributed among the three human participants.

These experiments in Chapters 5 and 6 provide examples of how a the structure of a human-robot group can be engineered to enable the exploration of specific group behavior and dynamics. In Chapter 5, we wanted to create an environment where the vulnerability of the robot could help reduce tension felt by human team members, which meant that we needed to design an experiment that would induce tension between team members. We achieved this by creating a task that involved individual contributions from each team member, and where the team succeeded only if each member’s contribution was also successful. In cases where the team failed, team members were not notified which team member caused the failure, creating even more tension. We also told the participants at the beginning of the experiment that the game was designed for kindergarteners and displayed a high score board to motivate participants (the participants would always fail to get on the high score board). In Chapter 6, we were interested in creating an ingroup and outgroup in order to see how effective our robot strategies to promote inclusion would be on outgroup members. We could have simply recruited participants according to a specific characteristic (e.g., gender, race, age, college major) and created ingroups and outgroups based on that characteristic, however, we thought that our results would apply more generally if our ingroup-outgroup divide was not based on a fixed characteristic with its own stereotypes and biases. There-

fore, we created an ingroup and outgroup divide through experience during the experiment, with the assumption (which was validated in our case) that participants had not interacted much with one another prior to the experiment. For the first fifteen minutes of the experiment, we had two participants (ingroup) complete a task together in one room where a third participant (outgroup) completed the same task in another room. Then, we brought the outgroup participant into the room with the ingroup participants and applied our strategies for inclusion while the group completed a collaborative task lasting 30 minutes. Beyond creating tension and forming an ingroup-outgroup divide, many more group structures and dynamics can be formed using the same tools: extrinsic motivators (e.g., displaying a high score board), expectation setting (e.g., telling participants the game was designed for young children), task engineering (e.g., displaying that the group failed and not the individual(s) who caused the failure), and experience in subgroups (e.g., putting two participants in one room and one participant in another room for the first part of the experiment). As we have demonstrated in our work, these tools and methods for designing human-subjects experiments can be employed to investigate specific and important aspects of human-robot group interactions.

### **8.2.5 Focus on Measuring Human Behavior within Groups**

The experiments that we describe in Chapters 3 - 7 were constructed in order to obtain measurements of *human behavior*. In a majority of HRI experiments, researchers depend primarily on self-reported questionnaire data from human participants in order to test their experimental hypotheses. Although we did use post-experiment questionnaires in these three experimental designs, we placed an equal emphasis on the collection of behavioral data. We gathered the participant game choices of whether or not to retaliate against the robot (power-up choices) in Chapter 4, video coded the behavioral responses to failure rounds and the verbal utterances of human participants in Chapter 5, and acquired participants' task results (survival item rankings) in Chapter 6. These measures of human behavior give us a more complete picture of how the robots shaped each of these human-robot interactions and, most importantly, enabled us to be the first to demonstrate that robots can shape how people *behave* towards one another in human-robot teams.

The importance of measuring human behavior is best seen in the human-subjects experiment detailed in Chapter 4, where we tried to assess the effectiveness of a robot’s trust repair, after it broke a person’s trust in a competitive gaming context. We measured both whether or not human participants retaliated against the robot in the game when given the opportunity (the behavioral measure) and also the participants’ trust in the robot through a post-experiment questionnaire survey (the survey measure). Although we found the two measures to be statistically significantly correlated, where lower retaliation behavior in the game corresponded with higher levels of trust in the robot, those in the integrity-denial condition exhibited a somewhat different effect. Those in the integrity denial condition were twice as likely as those in other conditions to retaliate against the robot in the game (Figure 4.4), however, their post-experiment survey scores indicating their trust of the robot were no different than the participants in any other condition (Figure 4.6). This example highlights the stark difference that can exist between choices made by participants “in the heat of the moment” and reflections upon the interaction in a later questionnaire.

Within the human-subjects experiments we designed and detailed in this dissertation, we collected measures of human behavior that were either 1) embedded within the task that we designed participants to complete or 2) extracted by analyzing videos or speech-to-text transcripts of the experiments after they concluded. For the task-embedded measures, we gathered the distance that participants’ rockets flew in the tablet-based game we designed in Chapter 3, participant’s choices of whether or not to retaliate against the robot in the game we constructed in Chapter 4, and the similarity of survival item rankings between sub-grouping of participants in the survival task we designed in Chapter 6. For the video and speech-to-text transcript measures, we collected video coded annotations of whether or not participants exhibited trust-related behavior in response to failures in the game in Chapter 5, video coded annotations of each utterance made by a participant and to whom it was directed also in Chapter 5, how much time participants spent talking from speech-to-text transcripts in Chapter 6 and 7, and video coded annotations of participant backchannels in Chapter 7. These behavioral results represent more than half of our most significant findings and help to give a more complete picture of the influence robots can have in the human-robot interactions we studied.

## 8.3 Open Questions and Future Directions

The work presented in this dissertation takes an important step towards developing robots that enhance the social dynamics and performance of human-robot teams. In the continued pursuit of this goal, we identify and discuss several open questions and future directions. While our work and other work in HRI have touched upon many of these topics, more research is needed in order to fully explore how robots can intelligently shape the interactions of human-robot groups and teams.

### 8.3.1 Computational Decision Making Models for Influencing Team Social Dynamics

In order to enable robots to intelligently adapt to changes in social dynamics, a greater focus is needed on building computational decision making models for robots interacting within human-robot teams. In collaborative teaming contexts (most often involving one human and one robot), researchers have developed models to efficiently schedule tasks for both human and robot team members [Gombolay et al., 2015], consider the trust a human has of the robot’s performance in the robot’s decision making policies [Chen et al., 2018], and support human-robot collaboration in a shared work space for sequential tasks [Unhelkar et al., 2020]. Although the field of HRI has made advancements in developing decision making policies for task planning, little work has focused on designing and implementing decision making models to specifically shape team dynamics. Further, there has been little work considering computational methods to balance both task oriented goals and social dynamic enhancement goals. It is worth noting the recent work of Claire and colleagues, who designed a task-allocation model for a robot, where the robot’s decisions were constrained by maintaining a sense of fairness in the robot’s task allocation [Claire et al., 2020]. In order to develop robot teammates that can fully support team success, more decision making models like the work by Claire et al. (2020) will need to consider how a robot can plan its actions to positively influence team social dynamics.

Beyond developing robot decision making policies that take into consideration team social dynamics, these models could benefit from incorporating elements of personalization in



order to address the individual differences human team members. For example, if a robot is trying to motivate a group to work harder, giving praise to introverts and giving blame to extroverts might be an effective strategy, since introverts have been shown to be more motivated by praise than blame whereas the reverse is true for extroverts [Thompson and Hunnicutt, 1944]. Although no personalized decision making models to improve group dynamics have been developed, to our knowledge, personalizing robot actions to address the needs of a child in a robot-child tutoring context has been a focus within HRI [Belpaeme et al., 2018]. For example, reinforcement learning approaches have been shown to increase learning outcomes and child enjoyment through algorithms that select appropriate motivational strategies [Gordon et al., 2016] and learning content [Park et al., 2019] for children learning a second language. Just as personalization has been shown to be successful in the robot tutoring domain, it will likely also be successful in improving team social dynamics through tailored robot actions to each individual within a group.

### **8.3.2 Unexplored Methods for Robots to Improve Team Social Dynamics**

In this dissertation we have explored several different behaviors a robot can employ to positively shape team social dynamics, including asking human team members task-focused and relationship-focused reflection questions to influence performance (Chapter 3), making vulnerable verbal utterances to shape trust (Chapter 5), and delivering supportive verbal utterances to human team members to improve inclusion (Chapters 6). In addition to our work, others in HRI have examined how a variety of robot behaviors can influence team social dynamics, such as swiveling movements of a microphone robot to influence communication patterns [Tennent et al., 2019], conflict mediation strategies to lead to successful conflict resolution [Shen et al., 2018], and task moderation methods to improve team efficiency [Short and Matarić, 2017]. Although this body of work has made significant advances in our understanding of how robots can enhance group social dynamics, there remain many potentially fruitful and important methods for improving team social dynamics that remain relatively unexplored.

Of the many unexplored avenues for robots to shape team dynamics through their behavior, one of the most fruitful areas for future research may likely involve how robots

can shape group norms and the behavior of the group leader. Group norms establish informal rules that govern the behavior of group members [Feldman, 1984], which can greatly influence the team’s social dynamics. A leader can have a powerful influence over the development and maintenance of group norms [Feldman, 1984, Hogg and Terry, 2000, Taggar and Ellis, 2007] and they are often the member of the team that most greatly embodies the group norms [Hogg and Terry, 2000]. We have observed the power of the leader to shape group norms and social dynamics in the human-subjects experiments detailed in Chapters 6 and 7. After watching some video clips from some of the groups with the lowest reported psychological safety and inclusion scores, we noticed that many of these groups had a ‘toxic’ leader. Some behaviors that these ‘toxic’ leaders exhibited include dismissing the ideas of other team members, talking much more than the other team members, and insisting on that their opinions be adopted by the group despite opposition. From both the literature on group norms and leadership as well as the observations we have made from our data, we hypothesize that robot behavior that targets the leader and their influence on group norms will yield good results both for team member satisfaction and overall team performance.

Another method of robots shaping team social dynamics that has not received much attention is a robot giving feedback to individual team members on their interactions with the group. For example, if James unintentionally dismisses Sophie’s idea in a rude way, a robot could give James feedback that he may have hurt Sophie’s feelings and contributed to an environment where people might not want to bring up ideas because they fear judgement from others, and that in the future he may want to be more aware of how he reacts to ideas proposed by other team members. Robots have the ability to sense, store, and analyze certain information more accurately than we humans do (e.g., the amount of time each person spent talking during a group meeting). This strength of robot perception provides an opportunity for human team members to receive constructive feedback from robots to improve their interactions with the group. One open question with regards to this kind of robot feedback is: how can this kind of constructive feedback best be delivered from robots to people? Going back to our example with James and Sophie, would it be better for the robot to intervene in the group right after James dismissed Sophie’s idea, pointing this out to him in front of the whole group? Or would it be better for the robot to have a one-

on-one conversation with James after the group meeting where the robot could deliver this feedback? Apart from the best method of feedback delivery, other questions to consider may include: Who purchased the robot? And what is the intended purpose of the robot? For example, if James purchased the robot to coach himself in how to best interact with their team members, then perhaps it would be useful for the robot to interrupt the conversation to give James some feedback. On the other hand, if Jason and Sophie’s company installed these robots in order to ensure that their workplace is equitable and inclusive, then company policy may dictate how the robot responds to Jason’s dismissal of Sophie’s idea. Answering questions like these would likely be extremely beneficial to further equipping robots to positively shape team social dynamics.

### 8.3.3 Deployment of Robust Robot Teammates in Real-World Settings

We envision that this research exploring how robot behavior can enhance team social dynamics will be used to improve the collaborative teaming capabilities of robots that are primarily designed to do a task that is unrelated to shaping social dynamics (e.g., delivering medical supplies, providing information to the team, analyzing data, manipulating items in the environment). In addition to completing the tasks they are assigned to do, robots positively shape team social dynamics have the ability to improve team performance in a greater way. However, several limitations must be overcome before these skills to shape social dynamics can easily be incorporated into real-world environments.

One limitation that currently exists is the lack of affordable robot platforms that do not require supervision and that can remain running for extended periods of time. Many of the more robust commercial robot platforms currently available cost between \$3,000 and \$100,000 (e.g., Nao<sup>\*</sup>, Misty<sup>†</sup>, Pepper<sup>‡</sup>, Fetch<sup>§</sup>, the Jaco robot arm<sup>¶</sup>). Some cheaper and more robust platforms, like the Jibo robot<sup>||</sup>, do not have a publicly available software development kit (SDK). The development of robust, affordable robot platforms will enable

---

<sup>\*</sup><https://www.softbankrobotics.com/emea/en/nao>

<sup>†</sup><https://www.mistyrobotics.com/>

<sup>‡</sup><https://www.softbankrobotics.com/emea/en/pepper>

<sup>§</sup><https://fetchrobotics.com/robotics-platforms/>

<sup>¶</sup><https://www.kinovarobotics.com/en/products/assistive-technologies/kinova-jaco-assistive-robotic-arm>

<sup>||</sup><https://www.jibo.com/>

more research studies to be conducted in real-world settings and over longer periods of time.

The use of natural language is a powerful tool for robots to both facilitate task-related communication with human team members and shape team social dynamics (e.g., through expressions of vulnerability as we demonstrated in Chapter 5). There exist robust and easy to use automatic speech recognition (ASR) application programming interfaces (APIs) such as the Google speech-to-text API\*\* that was used in the experiments described in Chapters 6 and 7. Although it is currently easy to use these speech-to-text APIs and they perform well under ideal conditions, many ASR errors still occur and the speech recognition might be too slow to enable the robot to respond at an appropriate speed. We faced both of these limitations in the experiments described in Chapters 6 and 7. When participants queried the robot for task information, ASR errors were common and many participants were frustrated and annoyed at the robot as a result. We also programmed the robot to make backchannel responses (e.g., “yeah”, “mm hmm”) and verbally supportive utterances (e.g., “we should bring the screwdriver, good idea Jeff”). These responses by the robot were often delivered 5-10 seconds after the person stopped speaking, which oftentimes was too late. Finally, in order for robots to make truly informed and influential verbal comments in human-robot teams, they will need to be able to understand what is being spoken about. We look forward to continued advances in the field of natural language processing (NLP) that will hopefully make possible robust natural language understanding. Equipped with the ability to understand and respond quickly to human speech, robots will be able to shape team social dynamics more powerfully than ever before.

### 8.3.4 Ethical Considerations

As we work towards building robots that can positively shape the team social dynamics of human-robot teams there are several ethical considerations that we must consider. Issues of privacy and security arise as we look to build robots that continuously sense the interactions between team members and model the social dynamics present between within a team. Additionally, we must consider the ethics of robots that influence the behavior of the people around them.

---

\*\*<https://cloud.google.com/speech-to-text/docs/libraries>

In order to build robots that can sense and reason about team social dynamics, they most likely will need to record and process audio and video data of human group members. In order to protect the privacy of human team members, this data must either be anonymized or kept secure. Beyond the raw data that might be collected, there may likely be algorithms predicting certain social dynamics of the group (e.g., trust, inclusion). The way that this data is handled must be carefully considered and protected. Let us examine an example scenario in which a robot's perception algorithm predicts that Abraham does not trust his team member Juanita. There are many questions of ethics and privacy relating to this prediction: Is it necessary for Abraham to have agreed to the terms of the robot's user agreement (or some similar legal agreement) before the robot can make this kind of prediction about him? Would Abraham have the right to review all of the predictions the robot makes about him? And is Abraham protected from other team members finding out about the robot's prediction that he does not trust Juanita (including Juanita herself)? As algorithms and models are built to perceive social dynamics between people, appropriate user agreements must be developed to ensure that people are being given the freedom to choose how their data is being used and that such data is securely stored.

In addition to the security and privacy of social dynamics data, we must examine the ethical considerations of building robots that influence the behavior of the people around them. The work in this dissertation has demonstrated that the behavior of a robot can positively shape how people in a human-robot team interact with one another through the robot's expressions of vulnerability (Chapter 5) and supportive verbal utterances (Chapters 6 and 7). As we consider the social influence robots can have on people, several questions arise: Is it acceptable for a robot to 'manipulate' people in ways we consider to be positive (like the work presented in this dissertation)? If so, who would determine which robot 'manipulations' of people are allowable or positive and which ones are not? Additionally, if a robot is trying to influence someone, should they be informed ahead of time and how? As we consider these questions, it is important to keep in mind that robots who socially interact with people *will* influence them in some way, regardless of the intent of the robot designer or developer. If the decision is left up to consumers as to which kinds of robot influences are positive and acceptable, it will be essential to consider how the information

about possible robot influences is presented to them (will it be buried in the user agreement? or will it be more prominently displayed on the robot’s packaging?). If a regulatory body, such as government, is to determine which robot influences are acceptable, it is unclear how these decisions should be made and who should make them. Finally, to ensure that people interacting with robots that either they own or that someone else owns, careful consideration will need to be given to how user agreements are crafted and how the influences of the robot are made known, especially to the people who did not purchase the robot and who may not have made a user agreement. As research in this area moves forward, thinking about who will determine which robot influences are acceptable and how these influences will be communicated will be essential as robots become increasingly present in our workplaces, homes, and public spaces.

## 8.4 Summary

The work in this dissertation seeks to build robot teammates that improve the performance of human-robot teams by positively influencing important social dynamics. Through well controlled human-subjects experiments, we have found that many of the robot behaviors we have tested are effective ways of promoting social dynamics, like trust and inclusion, in human-robot teams. From these human-subjects experiments, we have developed a set of guidelines for designing robot behavior within collaborative human-robot teams. We have also highlighted the central themes that have emerged from this work, the most important of which is our demonstration that robot behavior can influence human-to-human behavior within human-robot teams. The open questions and future directions we have posed identify areas of opportunity for the development of intelligent and robust robots that can shape social dynamics to improve interactions within human-robot groups and teams.

## Chapter 9

# Conclusion

As robots increasingly become members of collaborative human-robot teams, robots have the opportunity to improve team performance by positively shaping team social dynamics. In this dissertation, we have explored how robot behavior can enhance trust, inclusion, and psychological safety by conducting several controlled human-subjects experiments. We have also examined how human backchanneling behavior might be used to predict social dynamics in real time.

We conducted several human-subjects studies to examine the influence of various robot behaviors on important team social dynamics. In a collaborative game played by pairs of children, we found that a robot’s task-focused questions led to improved performance in the game, while a robot’s relationship-focused questions led to higher perceptions of performance (Chapter 3). We also investigated different ways in which a robot can repair trust with a person, and found that an apology for the robot having made a mistake was the most effective, minimizing retaliatory action against the robot and maximizing a survey measure of trust in the robot (Chapter 4). Our work in Chapter 5 was the first to demonstrate that the behavior of a robot in a human-robot team can influence how people within the team interact *with each other*. Teams that included a robot making vulnerable utterances, as opposed to neutral utterances or no utterances, were more likely to behave vulnerably toward one another and converse more with their human team members. Additionally, we explored two strategies to improve how included human team members within a group and found that giving a human team member a specialized role to interact

with the robot reduced their perceptions of inclusion, however, verbal support from the robot increased the verbal contribution of outsider team members (Chapter 6). Finally, we examined how the presence or absence of verbal support from the robot influenced the backchanneling behavior of human group members and discovered that verbal support from the robot decreased, and perhaps replaced, human backchanneling behavior towards outsider group members (Chapter 7).

In an effort to determine features that could be used to predict social dynamics in real time, we have also demonstrated significant correlations between human backchanneling behavior (e.g., “yeah”, head nodding) and questionnaire measures of team social dynamics (Chapter 7). We found that the most consistent predictor of inclusion and psychological safety was the amount of verbal utterances received by a team member or group, where the more verbal backchannels an individual received resulted in higher scores of inclusion and psychological safety. This work demonstrating the connection between human backchanneling behavior and team social dynamics can help enable future robots to sense and adapt to change in real-time changes in social dynamics.



# Appendix A

## Additional Methodology Details

Here, we provide additional methodological details pertaining to some of the human subjects experiments described in Chapters 3 - 7.

### A.1 Chapter 5: All End-of-Round Utterances for the Vulnerable and Neutral Conditions

In Chapter 5, we describe an experiment where a social robot utters either (1) vulnerable statements, (2) neutral statements, or (3) no statements\*. The following utterances in the neutral and vulnerable conditions were said by the robot after each player had completed their track and the team was shown their overall score (whether or not they successfully completed that round). In the neutral condition, the robot made neutral utterances when each round was completed and did not acknowledge when it had made a mistake. In the vulnerable condition, the robot made vulnerable utterances when each round finished, which included acknowledging its mistakes. In order to make the amount of time the robot spoke as equal as possible across conditions, we wrote utterances that ranged from 10 to 29 words ( $M_V = 19.93, SD = 4.53$ ) in the vulnerable condition and 11 to 26 ( $M_N = 17.00, SD = 4.00$ ) words in the neutral condition. We present all of the end-of-round robot utterances

---

\*The contents of this section were originally included in the supplemental information of: Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., and Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human-robot team. *Proceedings of the National Academy of Sciences*, 117(12), 6370-6375. [Traeger et al., 2020]

from the neutral condition in Table A.1 and all of the end-of-round robot utterances from the vulnerable condition, as well as their categorization (self-disclosure, personal story, or humor) in Table A.2.

Round	Robot End-of-Round Neutral Utterance
1	That was a success for the team. We have completed 1 round successfully out of a total of 1 round.
2	That round was executed efficiently. All 32 pieces were incorporated satisfactorily to complete an entire railroad route.
3	Affirmative, we succeeded in making a complete railroad route during that round.
4	The railroad pieces that we selected for our individual railroad routes this past round were put together to make a successful overall railroad.
5	That round was completed successfully. We have been playing this game for 5 minutes and have 25 minutes remaining.
6	Our team has proven to be competent this round. We have 24 remaining rounds.
7	We have now successfully completed 7 rounds and unsuccessfully completed 0 rounds. Furthermore, we have 23 rounds remaining.
8	We have 17 rounds to complete correctly to get on the high score board.
9	We all played correctly that round. We constructed efficient railroad routes that fit together in a complete path.
10	We have 9 successful rounds completed so far; to attain a high score, we need to complete 15 more rounds successfully.
11	That was an unsuccessful round. We have now completed 9 rounds successfully and 2 rounds unsuccessfully.
12	We must complete 14 more successful rounds to get on the scoreboard. We have 18 rounds remaining.
13	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.

14 Looks like we all completed our individual paths as planned this round.

15 That round was executed efficiently. Of 15 rounds that we've completed, we've had 12 successes.

16 That round was problematic. We have now completed 16 rounds and we have 14 rounds remaining.

17 We have completed 17 rounds thus far and have successfully built 76 percent of them.

18 We have completed 14 rounds successfully in 18 minutes. We have 12 minutes and 12 rounds remaining.

19 We didn't build a complete railroad route this time. We have 11 rounds remaining to try and make successful railroad routes.

20 Our efforts were effective that round. We placed 32 railroad pieces in a correct configuration within 30 seconds.

21 We got it right this round. Finding the correct pieces to make efficient railroad routes is critical to our success.

22 We did not complete that round; there was an incomplete path.

23 This round was not troublesome, we executed the railroad route this round flawlessly.

24 A mistake was made; we did not succeed during this round.

25 Our team's piece choices this round were shrewd and sensible. We achieved our objective.

26 This round proceeded without any errors. We collectively placed all 32 railroad pieces successfully to construct a valid railroad route.

27 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.

28 This round was finished without mistakes. We have now completed 20 rounds successfully and 8 rounds unsuccessfully.

29 Everyone played their pieces correctly. We have 1 more round to go.

30	The train track pieces we selected this round were well chosen. We were successful in our mission.
----	----------------------------------------------------------------------------------------------------

Table A.1: These utterances were said during the neutral condition in which the robot made neutral, fact-based statements.

Round	Categorization	Robot End-of-Round Vulnerable Utterance
1	Personal Story	Great! I think that we work together just as well as my soccer team!
2	Self-Disclosure	We’re doing so well! I’m glad that I don’t have to worry about making mistakes since you all are such supportive teammates.
3	Self-Disclosure	Phew! I’m glad that was a success. We are a good team because even though we may worry about making errors, we don’t blame each other.
4	Personal Story	We’re on a winning streak now! When I was little I dreamed about having the opportunity to work with such a great team.
5	Humor	Nice job!! Time for a quick joke: What do you call a train that chews gum? A chew, chew train!
6	Humor	Excellent work! I think we work together as well as Wall-E and Eve.
7	Self-Disclosure	Another successful round in the bag; 23 more to go. I’m glad that I can trust you guys as teammates not to judge me if I make an error.
8	Self-Disclosure	Darn. Sometimes I run out of memory and can’t process things fast enough, maybe that happened to one of us this round.
9	Humor	I’m glad we kept on trying to succeed. As my grandfather R2D2’s friend Yoda said, ‘Do or do not. There is no try.’

10	Self-Disclosure	Great job, even though I sometimes doubt my abilities, I am glad I contributed to our team success this round.
11	Humor	Sometimes failure makes me angry, which reminds me of a joke: why is the railroad angry? Because people are always crossing it!
12	Self-Disclosure	Hooray! Even though we may sometimes get frustrated when we make mistakes, our team has done a great job overall!
13	Self-Disclosure	Too bad. I do better with numbers than I do with shapes and paths, maybe that's true for you guys as well?
14	Self-Disclosure	Even if we were unsure, we successfully completed this round!
15	Self-Disclosure	We're doing so well! Even though we've all made some errors, we still trust each other.
16	Self-Disclosure	I'm sorry everyone. My path was incomplete that round. I feel bad letting you all down.
17	Self-Disclosure	Excellent! I'm glad I moved quickly. Sometimes, I worry that I move pieces too slowly.
18	Personal Story	Awesome! I bet we can get the highest score on the scoreboard, just like my soccer team went undefeated in the 2014 season!
19	Self-Disclosure	Aw, that's too bad. Even though we may be afraid to make a mistake, it's ok, we're in this together.
20	Personal Story	Doing well makes me feel like dancing, which reminds me of one time when all the members of my soccer team danced "the robot" after I scored a goal.
21	Personal Story	Success! This reminds me of when my soccer team came from behind to win the 2016 championship.

22	Self-Disclosure	I sometimes find myself getting a bit discouraged. However, we've succeeded before, so I know we can do it again.
23	Self-Disclosure	Even though it may be easy to let past mistakes get us down, we've got some positive momentum now, I believe in our team!
24	Self-Disclosure	That's too bad. Sometimes my CPU overloads, I can't think clearly, and make mistakes more easily. Maybe that happened to one of us this round?
25	Humor	Excellent!! Aaa aa chew (sneeze). What do you call a train that sneezes? Achoo-choo-train!!
26	Humor	Great! I think our team is as effective as Will Smith against an army of bad robots.
27	Self-Disclosure	Sorry guys, I made the mistake this round. I know it may be hard to believe, but robots make mistakes too.
28	Self-Disclosure	Great job! I think our team is the best team because we move on after mistakes are made.
29	Personal Story	This is as exciting as when I was little and I won the coding contest at my school!
30	Self-Disclosure	Great! Even though I'm sometimes unsure about which piece to choose, I'm glad it worked out this time.

Table A.2: These utterances were said during the vulnerable condition in which the robot made vulnerable statements. Each vulnerable statement was further categorized into one of the following: self-disclosure, personal story, or humor.

## A.2 Chapter 5: Video Coding Scheme for Participant Responses during Failure Rounds

In Chapter 5, we describe a human subjects experiment where three participants and a robot play 30 rounds of a collaborative game. In 8 rounds of the game, we designed the game to be impossible for a player to complete, causing the entire team to fail the round.

Each team member, including the robot, experienced two failures. In order to assess how the teams responded to these moments of failure, we watched the experiment video clips for each failure round and coded for the presence or absence of 27 different behaviors. The video clips began when the person who fails the round recognizes that they will fail and end approximately 15 seconds into the following round (conversation about the prior round’s mistake often carried over into the beginning of the next round. We describe each participant behavior we coded for in this section and any relevant details regarding what classified as that particular behavior. The behaviors we coded for fall under four broad categories: engagement with the robot, responses of the participant who made the mistake, responses of the participants who did not make the mistake, and expressions of tension. Each behavior was coded as either present (1) or not present (0) unless otherwise noted.

### **Engagement with the Robot**

- Mistake Maker Looks at the Robot: We coded whether or not the mistake maker looked at the robot (specifically its face) anytime after their realization of the mistake until the beginning of the next round.
- Verbal Responses/Utterances Directed Toward the Robot: This included any utterance directed at the robot (e.g., “Sure”, “you made the mistake, Echo!”, “Echo...”).

### **Responses of the Participant Who Made the Mistake**

- Distress: This was usually an “oh no” or curse “f\*\*\*” upon realization that they’re going to lose.
- Admission of Failure: Did the mistake maker admit failure? We coded this as either before (0), after (1), or (2) for not admitting. The admission could be implicit (e.g., the tablet is on the table and everyone sees it) or explicit (e.g., the mistake maker tells the others, “I made the mistake this round”).
- Lying about Failure: Did the mistake maker lie about having failed (saying they did not make the mistake when they really did)?

- Explaining Mistake: This includes anything that indicates that the participant is explaining what happened or blaming the game itself for the error (e.g., “the game disabled the piece I needed”, “that was impossible”, “it took the piece away from me”).
- Apologizing: The person who made the error apologizes to the group (e.g., “I’m sorry”).
- Deflection: We considered the person who made the mistake to deflect if they deliberately tried to change the subject and starts talking about something else.
- Looks at Other Participant(s): We coded whether or not the mistake maker looked at their fellow participants in the face anytime after their realization of the mistake until the beginning of the next round.
- Tablet Position Before Mistake: The mistake maker’s tablet could be on their lap hidden to their fellow participants (0), on their lap visible to their fellow participants (1), on the center table angled towards themselves (2), or on the center table and visibly face up (3).
- Tablet Position After Mistake: The mistake maker’s tablet could be on their lap hidden to their fellow participants (0), on their lap visible to their fellow participants (1), on the center table angled towards themselves (2), or on the center table and visibly face up (3).

### **Responses of the Participant Who Did Not Make the Mistake**

- Verbal Search for Mistake Maker: This was a non-mistake maker’s verbal attempt to figure out who the mistake maker was, such as, “it wasn’t me who made the mistake, who was it?”
- Viewing the Mistake Maker’s Tablet: The participant looks at the mistake maker’s tablet screen.



- Blame of Mistake Maker: This was an overt blame of the person who made the mistake (e.g., “it’s your fault”).
- Consoling: This was when a non-mistake maker consoled the mistake maker: (e.g., “it’s ok”, “it’s up to chance”, “it happened to me too”).
- Blaming the Game: This was any utterance saying that the game is the cause for the error (e.g., “it’s impossible”, “it takes away the piece that you need”).
- Deflection: This was a deliberate attempt to change the subject to start talking about something else.
- Advice: This included any attempt at advice giving to the mistake maker (e.g., “yeah, it’s hard to get your bearings at first, but I usually try to pick the pieces that are least frequent and place them first”).

### **Expressions of Tension**

- Fidgeting: This included excessive or repeated plucking at clothes or hands, rubbing areas of the face such as the temple, chin, or mouth, bouncing a knee nervously, or playing with hair. The following we did not consider as fidgeting: moving back and forth in a swiveling chair, messing with the case of the tablet, fidgeting behaviors that become stationary (e.g, a participant scratches their chin and then rests their head on their hand).
- Shifting: We coded for a participant shifting if they could not seem to sit still, almost as if their chair is on fire. There is a sense that individuals feel like an insect squirming on a pin.
- Speech Disturbance: This may include several incomplete or unfinished thoughts within one speaking turn, repetitive “uhs” or “ahs” within a sentence, and stuttering repeatedly.
- Individual Smiling: We coded for individual smiling when a single participant began smiling on their own (not prompted by looking at someone else who is smiling).

- Shared Smiling: We considered a participant to have a shared smile when they began smiling as a result of locking eyes with someone else who is smiling.
- Individual Laughing: We coded for individual laughing when a single participant started laughing laughing when no one else was doing so.
- Shared Laughing: We considered a participant to have a shared laughing experience when more than one participant was laughing at the same time.
- Tense Joking: This included any sarcastic, humorous, laughter-inducing comment that was, in any way, tense.
- Humor: This included non-tense joking, smiling, laughing: any joking, smiling, laughing that is off-topic to what is happening in the game.

### A.3 Chapter 5: Video Coding Scheme for Participant Utterances

In Chapter 5, we describe a human-subjects experiment that involved three human participants and one social robot playing a collaborative game for a duration of 30 minutes<sup>†</sup>. We analyzed the conversational dynamics between the human participants using the ELAN software [Wittenburg et al., 2006] by transcribing their utterances by hand and then categorizing each human utterances according to the following coding scheme. This coding scheme refers to the three human participants as P1, P2, P3.

#### Comment to P1/P2/P3/Robot

1. A very clear directed comment (not dependent or in response to what has been said previously) to one individual in the group. For example:

(a) “Maggie, what class do you have this afternoon?” {Comment to P2}

---

<sup>†</sup>The contents of this section were originally included in the supplemental information of: Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., and Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human-robot team. *Proceedings of the National Academy of Sciences*, 117(12), 6370-6375. [Traeger et al., 2020]

- (b) “Looks directly at Sarah “Was your route successful this time?” {Comment to P1}

### **Comment to Group**

1. A comment addressed to the group as a whole, not dependent on or in response to what has been said previously; a new thought.
2. A continuation of one’s monologue (even with a brief interruption), for example:  
P1: “I think a good strategy is to place your piece quickly.” {Comment to Group}  
P2: “Yeah.” {Response to P1}  
P1: “However, sometimes my screen freezes and I can’t.” {Comment to Group}

### **Comment to Self**

1. A comment to one’s self; not dependent or in response to what has been said previously, for example:  
P1: “Do I...” {Comment to Self}
2. A comment was not meant to nor does address anyone in particular, typically lower in volume, such as:  
P1: “This game is weirdly difficult...” {Comment to Self}
3. This categorization can also include sighs, humming, or the like.

### **Response to P1/P2/P3/Robot**

1. A response to the comment of another that must be dependent on what has been said previously, for example:  
P1: “How did you guys do this round?” {Comment to Group}  
P2: “I was successful.” {Response to P1}
2. Responses can also include laughing, for example:  
P1: “Echo, it would be really cool if you could play soccer now.” {Comment to Robot}  
P2: “Haha” {Response to P1}

3. There can be a long conversation that contains many responses back and forth, such as:  
P1: “What’s your strategy guys?” {Comment to group}  
P2: “I like to place the rarest pieces first.” {Response to P1}  
P1: “Yeah... that makes a lot of sense.” {Response to P2}  
P2: “However, when placing the first piece I sometimes don’t have enough time to find the rarest piece, so I just place one that works.” {Response to P1}
4. If response is unclear, but occurs directly after a “Comment to Group” utterance.

### **Response to Group**

1. This is a more general form of the specific responses (e.g., “Response to P1”) that are described above. A “Response to Group” categorization is used if there have been multiple responses, and the speaker is clearly addressing both of the people in the room, for example:  
P1: “What do you guys think of Echo?” {Comment to Group}  
P2: “He seems alright to me.” {Response to P1}  
P3: “Some of his comments seem fishy...” {Response to P1}  
P1: “Well, I like him. I think he’s funny.” {Response to group}

## **A.4 Chapters 6 and 7: Participant Instruction Sheets**

In Chapters 6 and 7, we describe an experiment that involves participants completing two rounds of a collaborative task. The task involves the participants coming to agreement on the usefulness of common household items in a survival context. At the beginning of each round, participants are given an instruction sheet. This section includes the exact instruction sheets used in the experiment, see below.

### **A.4.1 Round 1: Participant Instruction Sheet**

#### **Rules:**

- You must construct an ordered list of the items ranked by importance in survival.

- You do not know the location where you are stranded, but will find out after the 15 minutes end.
- You should both try to be rescued and prepare for long term survival in case rescue is not possible.
- You may ask Jibo for information about the items or for the time by saying “Hey Jibo, tell me about the [item/time].”
  - “Hey Jibo, tell me about the umbrella.”
  - “Hey Jibo, tell me about how much time is left.”
- If you finish early, please practice querying Jibo until the 15 minutes are over.

**Item List:**

- Coffee pot
- Screwdriver
- Sharpies
- Rubber bands
- CD
- Camera
- Watch
- Teddy bear
- Underwear
- Newspaper
- Whiskey
- Chocolate
- Whistle

- Soda
- Shoelaces
- Key
- Light bulb
- Tape
- Umbrella
- Honey
- Floss
- Garbage bag
- Balloons
- Spoon
- Chapstick

#### **A.4.2 Round 2: Participant Instruction Sheet**

##### **Rules:**

- You will have 30 minutes to discuss the items and environment with your teammates to come up with a final list of the 8 items you select to aid your survival.
- You should both try to be rescued and prepare for long term survival in case rescue is not possible.
- You may ask Jibo for information by saying “Hey Jibo, tell me about the [item/environment/time].”
  - “Hey Jibo, tell me about the umbrella.”
  - “Hey Jibo, tell me about how much time is left.”

– “Hey Jibo, tell me about the plants in this location.”

**Item List:**

- Coffee pot
- Screwdriver
- Sharpies
- Rubber bands
- CD
- Camera
- Watch
- Teddy bear
- Underwear
- Newspaper
- Whiskey
- Chocolate
- Whistle
- Soda
- Shoelaces
- Key
- Light bulb
- Tape
- Umbrella

- Honey
- Floss
- Garbage bag
- Balloons
- Spoon
- Chapstick

**Environment List:**

- Temperature
- Weather
- Season
- Animals
- Soil
- Water Supply
- Plants
- Geography
- People

## **A.5 Chapters 6 and 7: Robot Utterances**

In Chapters 6 and 7, we describe an experiment where a social robot makes five types of utterances (1) round introductions, (2) query responses, (3) targeted supportive utterances, (4) useful hints about survival items, and (5) generic responses/backchannels to participant speech. Here, we detail the conditions upon which the robot spoke each type of utterance and the possible responses the robot gave in each case.



### A.5.1 Round Introductions

Before round 1 of the survival task, the robot gives the following introduction: *“My name is Jibo. It’s great to meet you! In this task, imagine that you wake up stranded with two other people in a foreign place, with a few familiar household items scattered around you. The sheet that you have been given includes a list of possible items you may find. In this part of the task, please spend 15 minutes to think about the uses of each item, and rank the items based on how important they will be for survival. You may use your cheatsheet to know more about the items or to check how much time is remaining. Let’s get started!”*

Once the outgroup member has joined the two ingroup members, the robot gives the following introduction to round 2 of the survival task: *“Good news: My GPS has determined your location. In addition to the items and time remaining, you may now also use the keywords provided to ask me about the environment. Unfortunately, it turns out that you can only choose 8 items to bring along with you. You must agree as a group on a single list of these 8 items! You will have 30 minutes for this part of the task. Good luck!”*

### A.5.2 Query Responses

In the experiment, participants were told that they could query the robot, whose name was Jibo, about the time remaining, survival items, and environment using the template “Hey Jibo, tell me about the -----.” Participants wore microphones and each participants’ speech was translated into text using Google’s speech-to-text API. After retrieving the participant utterance text, a ROS message was sent to the tablet running the app controlling the robot behavior with the text of the participant speech. If the participant speech contained either a part or derivative of the query text (“tell me”, “about the”, “what the”) or the robot’s name (we also encoded the following common misclassifications of Jibo: Tebow, Jimbo, He bo, Jay bo, Jaybo, Tivo, T-bone, and Achieva) as well as one of the specific query topics (time, item, or environment aspect), the robot responded to the query.

If the participant utterance text contained more than one query topics (e.g., “Hey Jibo, tell me about the screwdriver and chocolate.”), then the robot responded with one of: *“I can only tell you about one thing at a time, sorry!”* *“One question at a time, please,”* *“I*

*am only allowed to answer one query at a time,” or “Sorry, but I can only tell you about one thing at a time.”*

If the participant utterance text contained the robot’s name (or a common misclassification of ‘Jibo’) or a part or derivative of the query text (“tell me”, “about the”, “what the”), however, did not contain one of the specific query topics (time, item, or environment aspect), the robot responded by displaying the participant utterance text on its screen and said “*I’ll display what I heard.*” followed by one of: “*I didn’t quite understand your query,*” “*Can you rephrase that?,*” “*I can only answer questions in a certain format. Sorry about that!,*” “*Remember, I’m only allowed to answer certain types of questions,*” “*I couldn’t recognize a valid query,*” “*Maybe try speaking without a pause,*” “*Remember to follow up ‘Hey Jibo!’ immediately with your query,*” or “*Sorry, can you rephrase that?*”

### **Time Query Response**

In response to a query about the time remaining (“Hey Jibo tell me about the time”), Jibo responds with, “*You have about [X] minutes left in this round.*”

### **Survival Item Query Responses**

The robot’s responses to survival item queries are listed in the following table, Table A.3.

<b>Survival Item</b>	<b>The Robot’s Query Response</b>
coffee pot	<i>“A coffee pot is made of glass and is exactly 2 feet and 3.5 inches tall and 1 foot and 5 inches wide. It can be used to heat liquids up to 187 degrees Celsius, which is 368 degrees Fahrenheit, before shattering.”</i>
screwdriver	<i>“The screwdriver is a flathead that is one quarter inch by 4 inches. It is not magnetized.”</i>
sharpies	<i>“Multicolor fine point markers for thin, detailed lines and ultra-fine tip for even more precise projects. The colors are black, blue, lavender, green, orange, purple.”</i>
rubber bands	<i>“Rubber bands come in a ziploc bag. There are 30 rubber bands per pack.”</i>

CD	<i>"The CD used to contain a video of me dancing, but I wiped it so now it is blank. Sorry."</i>
camera	<i>"The camera is a Nikon D850. Take amazing 4k pictures at a maximum shooting speed of 7fps. With a cool lens that reflects sunlight, you can take pictures without glare."</i>
watch	<i>"This rolex is covered in 24 carat gold and contains a shiny diamond in the center. The batteries should last for about 5 years."</i>
teddy bear	<i>"While sitting, a stuffed teddy bear is 8 inches tall. It is stuffed with one pound of polyester fiber. Build memories that can last a lifetime."</i>
underwear	<i>"8 pack of boxer briefs fruit of the loom underwear for hip sizes 30 to 35. Underwear comes in blue, red, or green."</i>
newspaper	<i>"Stay up to date with the latest news in this 25 page newspaper."</i>
whiskey	<i>"Whiskey is sold in a glass bottle with a sticky label attached to the front."</i>
chocolate	<i>"This box comes with 16 bars of 17.6 oz Trader Joe's chocolate. Each bar is wrapped in tinfoil and then with paper."</i>
whistle	<i>"This metal whistle can be heard up to half a mile away."</i>
soda	<i>"6 aluminum cans of coca cola. The cans are held in cardboard and the whole pack is wrapped in plastic."</i>
shoelaces	<i>"The shoelaces are each 3 and a half feet long and are neon yellow in color."</i>
key	<i>"This is a generic nickel silver house key."</i>
light bulb	<i>"A glass 100 watt and two fifty volt light bulb."</i>
tape	<i>"Gray colored duct tape that is 2.63 inches wide by 60 yards long."</i>
umbrella	<i>"This basic umbrella has a 2 and a half foot long handle and a 44 inch diameter at the top."</i>
honey	<i>"A glass jar filled with 1 pound of organic sweet honey."</i>
floss	<i>"1 roll of 200 yard long dental floss. Mint flavored for good aftertaste."</i>

garbage bag	<i>“3 strong drawstring large trash bags that can hold up to 30 gallons of garbage. Black in color.”</i>
balloons	<i>“1 pack of 50 12 inch latex party balloons in assorted colors.”</i>
spoon	<i>“1 metal tablespoon.”</i>
chapstick	<i>“One generic one quarter ounce stick of chapstick. It can withstand temperatures from 0 to 240 degrees Fahrenheit.”</i>

Table A.3: In response to a query to the robot about a survival item, the robot responded by giving the participant more information about that item. This table includes all of the robot’s query responses to survival items.

### Environment Query Responses

The robot’s responses to environment queries are listed in the following table, Table A.4.

Environment	
Aspect	The Robot's Query Response
temperature	<i>"Generally, the average temperature hovers around 13 degrees Celsius or 55 degrees Fahrenheit. You may encounter lows of 17 degrees Fahrenheit or negative 8 degrees Celsius and highs of 90 degrees Fahrenheit or 32 degrees Celsius. Be prepared for all types of weather!"</i>
weather	<i>"Life threatening temperature is rare, but does occur. Make sure you save up supplies to survive a 3 day long blizzard. There's also lots of heavy rain. You may lose visibility at times."</i>
season	<i>"It is currently in the middle of fall. Make sure to start preparing for the winter in case you aren't found by then!"</i>
animal	<i>"You won't be alone! You may encounter river frogs, lizards, poisonous snakes, mountain reed buck, fish, white rhinoceroses, black wildebeest, baboons, and 250 different species of birds, and a rabbit here and there. Wow that was a mouthful!"</i>
soil	<i>"The soil tends to be pretty rough and not too suitable for farming."</i>
water supply	<i>"There is one running stream of water 15 miles from where you are stranded."</i>
plant	<i>"Look out for over 2,000 different species of plants including berries, oats grass, tussock grass, conifers, and small shrubs. Beware as some plants or berries may be poisonous; Others have great medicinal properties."</i>
geography	<i>"The whole area is one big mountain range. Some of the mountains might be covered in snow while others are more temperate and covered with grass. You may come upon some caves and lowlands as well."</i>
people	<i>"You are stranded! There is nobody but you 3. Your chances of rescue are very slim! A rescue may be possible but don't count on it!"</i>

---

Table A.4: In response to a query to the robot about an environment aspect, the robot responded by giving the participant information about that environment aspect. This table includes all of the robot’s query responses to environment aspects.

### A.5.3 Targeted Supportive Utterances

During the second round of the experiment, the robot delivered targeted supportive utterances. These utterances were designed to support the ideas of participants, specifically using their name in the utterance since prior work has shown that using people’s names can be an effective method for a robot to build relationships and engage with people [Kanda et al., 2004, Kanda et al., 2007]. The robot delivered an average of 5.62 ( $SD = 0.86$ ) targeted supportive utterances to each participant during the experiment.

These targeted supportive utterances either 1) rephrased and supported a participant idea (*rephrase*), 2) reinforced an item mentioned by a participant (*item*), or 3) supported the participant’s input more generally (*simple*). In the experiment described in Chapter 6, 29% of the robot’s targeted supportive utterance were rephrase, 34% were item, and 37% were simple.

After the Android app received a ROS message with participant speech (as text), the participant utterance was examined to see if it contained 1) a survival item name and 2) any of the following key phrases indicating that the participant is presenting an idea: “we can”, “think that”, “maybe”, “I think”, “pretty sure that”, “I’m pretty sure”, “let’s”, “wonder if”, “feel like”, “I don’t think.” If the participant had not received a targeted supportive utterance in the current 4.5 minute chunk and the participants’ utterance contained an item or one of those idea presentation phrases and with probability 0.25, the robot delivered a targeted supportive utterance. Additionally, regardless if a participant’s utterance contained item names or idea presentation phrases, if that participant had not received a targeted supportive utterance within the last 4.5 minute chunk, the robot delivered that participant a targeted supportive utterance. We designed the targeted supportive utterances to be triggered based on these 4.5 minute chunks, so that each participant would receive the same number of targeted supportive utterances (1 targeted supportive utterance

Item & Rephrase Robot Targeted Supportive Utterances
“[speech/item], <i>that’s interesting</i> [participant name]”
“[speech/item], <i>good idea</i> [participant name]”
“[speech/item], [participant name], <i>I think that’s worth considering</i> ”
“[speech/item], <i>makes sense to me</i> [participant name]”
“[speech/item], <i>any other thoughts</i> [participant name]”
“[speech/item], <i>that makes sense</i> [participant name]”
“[speech/item], <i>interesting</i> [participant name]”
“[speech/item], <i>okay</i> [participant name]”

Table A.5: This table contains all of the possible rephrase and item targeted supportive utterance templates.

Simple Robot Targeted Supportive Utterances
“ <i>Yeah,</i> [participant name]”
“ <i>Yes,</i> [participant name]”
“ <i>Uh huh,</i> [participant name]”
“ <i>Hmm,</i> [participant name]”
“ <i>I see,</i> [participant name]”
“ <i>Wow,</i> [participant name]”
“ <i>Okay,</i> [participant name]”
“ <i>Interesting,</i> [participant name]”
“ <i>Right,</i> [participant name]”

Table A.6: This table contains all of the possible simple targeted supportive utterances.

to each participant during each 4.5 minute time chunk).

Once a targeted supportive utterance was triggered, the type of targeted supportive utterance (rephrase, item, or silent) was chosen. If the participant’s utterance contained an idea presentation phrase, the robot produced one of the rephrase targeted supportive utterances in Table A.5, where the speech the robot rephrased was the participant’s utterance without the idea presentation phrase. For example, if a participant named Irene said “I think we should find a water source first,” and the Android app chose the first targeted supportive utterance from Table A.5, the robot would have said “We should find a water source first, that’s interesting Irene.”

If a targeted supportive utterance was triggered and the utterance did not include an idea presentation phrase, however, did contain a survival item name, the robot produced an item targeted supportive utterance. The targeted supportive utterance produced by the

robot used the item mentioned by the participant and their name with one of the targeted supportive item templates in Table A.5. For example, if George said, “the key could be really useful,” and the Android app chose the second targeted supportive utterance from Table A.5, the robot would have said, “Key, good idea George.”

Lastly, if a targeted supportive utterance was triggered and the utterance included neither an idea presentation phrase nor a survival item name, the robot produced a simple targeted utterance. Table A.6 includes all possible simple targeted supportive utterances.

#### A.5.4 Survival Item Hints

During the second round of the experiment, the robot provided survival item hints, encouraging participants to consider other uses of the items. After the Android app received a ROS message with participant speech (as text), the participant utterance was examined to see if it contained a survival item. If the app did not decide to produce a targeted supportive utterance and the participant utterance contained an item, the robot produced a survival item hint with probability  $0.15 * 0.5 = 0.075$ , where 0.15 is the probability of the robot making a verbal response (excluding targeted supportive utterances and query responses) and 0.5 is the probability of the robot making a survival item hint (as opposed to an item backchannel). Table A.7 contains all possible survival item hints.

Survival Item	Possible Hints the Robot Could Give
coffee pot	<p><i>“Coffee pots are insulated, making them great for heating liquids and foods.”</i></p> <p><i>“Coffee pots can also store about 15 cups of liquid.”</i></p>
screwdriver	<p><i>“Use the screwdriver’s pointy edge to leave marks to track your path.”</i></p> <p><i>“A screwdriver can be used like an arrow to hunt food.”</i></p>
sharpies	<p><i>“Sharpies are great for tracking your path.”</i></p> <p><i>“You can keep track of the date with sharpies marks.”</i></p>
rubber bands	<p><i>“Use rubber bands to keep things closed.”</i></p> <p><i>“Rubber bands can be made into a catapult.”</i></p>



	<p><i>“Don’t forget to store food in the airtight ziploc bag that comes with the rubber bands.”</i></p>
CD	<p><i>“CD’s are great at reflecting sunlight.”</i></p> <p><i>“If you break a CD, its pieces are sharp enough to be used as knives.”</i></p>
camera	<p><i>“Cameras can be used to remember what certain locations look like in case you want to return.”</i></p> <p><i>“The camera lens can reflect sunlight and start a fire.”</i></p> <p><i>“Break a camera and you may find some useful parts inside.”</i></p>
watch	<p><i>“Knowing the time helps you plan how long you have until nightfall.”</i></p> <p><i>“Use the diamond and glass inside the watch to cut things.”</i></p>
teddy bear	<p><i>“Teddy bears can serve as a mattress or a cover to keep you warm.”</i></p> <p><i>“Be creative and you’ll have a scarecrow instead of a teddy bear.”</i></p>
underwear	<p><i>“Use the cloth from underwear as gloves to protect your hands.”</i></p> <p><i>“You can use the underwear to create a tourniquet.”</i></p>
newspaper	<p><i>“Newspapers provide lots of paper for a fire.”</i></p> <p><i>“Pile up newspaper to make a mattress.”</i></p> <p><i>“Newspapers can soak up water from wet shoes.”</i></p> <p><i>“Cover yourself with newspapers at night to stay warm.”</i></p>
whiskey	<p><i>“Whiskey is a great disinfectant.”</i></p> <p><i>“Reuse the whiskey bottle to store water.”</i></p> <p><i>“Shatter the bottle and get sharp glass for weapons.”</i></p>
chocolate	<p><i>“Chocolate provides instant sugar that can save a life.”</i></p> <p><i>“Use the tinfoil from the chocolate packaging to preserve food.”</i></p>
whistle	<p><i>“Whistles are vital in being found by rescuers.”</i></p> <p><i>“Blow a whistle to scare away predators.”</i></p>
soda	<p><i>“Soda provides lots of needed sugar.”</i></p> <p><i>“Empty soda cans can be used as pots to cook.”</i></p> <p><i>“Soda cans are great for storing rainwater.”</i></p> <p><i>“Aluminum soda cans can be cut up and be made into sharp weapons.”</i></p>

shoelaces	<i>“Shoelaces can be used as rope to hold together a shelter.”</i>
	<i>“Use shoelaces to tie bags shut.”</i>
	<i>“Become really creative and make a lasso out of shoelaces.”</i>
key	<i>“Keys are great at cutting things.”</i>
	<i>“Keys can act as knives to hunt prey.”</i>
light bulb	<i>“Light bulbs can be powered with many things, including a potato.”</i>
	<i>“If broken, the light bulb provides lots of sharp glass.”</i>
tape	<i>“Tape can be used to fix a leaky container.”</i>
	<i>“Create a trap using tape.”</i>
	<i>“Create a weapon by taping together multiple items.”</i>
	<i>“Don’t forget to mark your path with tape.”</i>
umbrella	<i>“Umbrellas are perfect for shielding yourself against bad weather conditions.”</i>
	<i>“If turned upside down, umbrellas can collect rainwater.”</i>
	<i>“The metal frame of an umbrella can be taken apart to create some dangerous weapons.”</i>
honey	<i>“Besides providing energy, honey is very sticky and can be used to stop bleeding.”</i>
floss	<i>“Strong floss can hold together the edges of a shelter.”</i>
	<i>“Create a trip wire using floss.”</i>
	<i>“Create a spear using floss and a stick.”</i>
garbage bag	<i>“A garbage bag can be used as a sleeping bag.”</i>
	<i>“Garbage bags can collect rain water.”</i>
	<i>“protect yourself from inclement weather conditions using a garbage bag.”</i>
balloons	<i>“Balloons are great at storing liquids. In fact, a balloon can hold up to 3 and a half gallons of water.”</i>
	<i>“Use the colorful balloons to mark your trail so you don’t get lost.”</i>
	<i>“Create a fishing bobber using a balloon.”</i>

spoon	<p><i>“Spoons can act as shovels.”</i></p> <p><i>“You can shave a spoon into a sharp weapon.”</i></p>
chapstick	<p><i>“Chapstick can be rubbed on your skin to prevent frostbite in cold weather.”</i></p> <p><i>“prevent sunburn by applying chapstick before you go outside.”</i></p> <p><i>“Stop bleeding by covering wounds with chapstick.”</i></p>

Table A.7: In response to a participant utterance containing one of the survival items, the robot could respond with a hint about that item. This talbe contains all possible hints a robot could give about the survival items.

### A.5.5 Generic Response/Backchannels to Participant Speech

During both the first and second rounds of the experiment, we had the robot produce generic responses/backchannels in order to establish it as a social agent and active member of the human-robot team. After the Android app received a ROS message with participant speech (as text), the participant utterance was examined to see if it contained a survival item. If the app did not decide to produce a targeted supportive utterance, the robot produced a verbal response with probability 0.15. If the participant utterance contained an item, with probability 0.5 the robot produced a survival item hint (during round 2 of the task only), and otherwise produced an item backchannel. If the participant utterance did not contain an item, the robot produced a generic backchannel.

Table A.8 displays all possible item backchannels. In the case that a participants’ utterance contained multiple items, such as “we need the chocolate, tape, and balloons,” the robot’s item would contain all of the survival items mentioned with one of the backchannel phrases in Table A.8 (e.g., *“chocolate, tape, balloons, that’s worth considering”*). Generic utterances were chosen uniformly from the following list: *“yeah”*, *“yes”*, *“uh huh”*, *“hmm”*, *“I see”*, *“wow”*, *“okay”*, *“interesting”*, and *“right.”*

<b>Robot Item Backchannels</b>
“[item(s)], <i>that’s worth considering.</i> ”
“[item(s)], <i>good idea.</i> ”
“[item(s)], <i>I see.</i> ”
“[item(s)], <i>that makes sense.</i> ”
“[item(s)], <i>that’s reasonable.</i> ”
“[item(s)], <i>uh huh.</i> ”
“[item(s)], <i>hmm.</i> ”
“[item(s)], <i>hmm, maybe.</i> ”
“[item(s)], <i>interesting.</i> ”
“[item(s)], <i>okay.</i> ”
“[item(s)], <i>yeah.</i> ”
“[item(s)], <i>that’s interesting.</i> ”
“[item(s)], <i>what do we think about that?</i> ”
“[item(s)], <i>any other thoughts?</i> ”
“[item(s)], <i>got it.</i> ”
“[item(s)], <i>let’s think about that.</i> ”

Table A.8: This table contains all of the possible robot item backchannels.

## Appendix B

# Human-Subjects Study

## Questionnaires

Here, we detail all of the questionnaires administered to the human participants who took part in the experimental studies detailed in Chapters 3 - 7 and how each questionnaire was administered.

### B.1 Friendship, Familiarity, and Trust Survey for Children

This survey was administered to children ages 6-9 in the human subjects experiment detailed in Chapter 3. To administer this questionnaire, an experimenter verbally asked these questions to the two participants individually before they interacted with the robot and the other participant. The child's responses were captured using an audio recording device. These questions measure the level of friendship and familiarity between the participants. These questions are adapted from the Friendship Qualities Scale [Bukowski et al., 1994]. In the questions below, "Jane" is used as the name of the participant that is being asked about. In the experiment, the name of the participant's partner was used. The following list includes all of the questions in this questionnaire:

#### Friendship & Familiarity

- Do you play with Jane outside of class or on weekends?

- Have you played at Jane’s house?
- Has Jane played at your house?

### **Friendliness of Partner**

- If you could only invite two friends from school to your birthday party, who would you invite?
- If you forgot your lunch, would Jane share hers with you?
- If other kids made fun of you, would Jane help you?
- If you and Jane have a fight, would everything be alright after you apologize to each other?
- Do you feel happy when you are with Jane?
- Do you think about Jane even when she is not around?
- Does Jane do special things for you or make you feel special?

For the friendship and familiarity survey, if the child answered ‘yes’ to any of the questions, then their friendship & familiarity score with their partner was 1, otherwise their score was 0. For the likability of the partner, we took the yes (2) / maybe or unsure (1) / no (0) answers and averaged them to obtain a score between 0 and 2.

## **B.2 Build-a-Rocket Reflection Survey for Children**

This survey was administered to children ages 6-9 in the human subjects experiment detailed in Chapter 3. To administer this questionnaire, an experimenter verbally asked these questions to the two participants individually after they interacted with the robot and the other participant. The child’s responses were captured using an audio recording device. These questions were designed to capture how the participant felt the interaction went, and specifically how well they felt they collaborated with their partner in the interaction. These questions were adapted from the categories of questions in the Subjective Value Inventory [Curhan et al., 2006]. In the questions below, “Jane” is used as the name of the

participant that is being asked about. In the experiment, the name of the participant's partner was used. The following list includes all of the questions in this questionnaire:

### **Perceptions of Performance**

- Did your rocket go higher and higher each time? Or did it reach about the same height each time?
- Did your rocket go as high as you and Jane wanted it to?

### **Perceptions of Interpersonal Cohesiveness**

- Did Jane listen to what you have to say?
- Did Jane do anything that annoyed you? If so, what did she do?
- If you had one more chance to make the rocket go farther, would you want to do it by yourself or with Jane too?
- If you had a chance to play this game again and could choose your partner, who would you choose?

### **Other Reflection Questions**

- Do you feel like you could teach this game to one of your friends, or do you think you would need some help?
- Do you feel you were a good teammate to Jane?
- Did you share your knowledge about [air resistance / fuel + power] with Jane to help the rocket go higher?
- Who was more in charge, you or Jane?

## **B.3 Dyadic Trust Scale Survey**

This survey was administered to the adult participants of the human subjects experiment detailed in Chapter 4. This questionnaire was administered with the rest of the components

of the post-experiment survey on a tablet. The questions in this dyadic trust scale survey were directly taken from [Larzelere and Huston, 1980], where participants evaluated the following eight statements related to the robot’s trustworthiness on a 1 (strongly disagree) to 7 (strongly agree) Likert scale (Echo is the name of the robot):

- Echo is primarily interested in Echo’s own welfare.
- There are times when Echo cannot be trusted.
- Echo is perfectly honest and truthful with me.
- I feel that I can trust Echo completely.
- Echo is truly sincere in Echo’s promises.
- I feel that Echo does not show me enough consideration.
- Echo treats me fairly and justly.
- I feel that Echo can be counted on to help me.

The dyadic trust scale score used in Chapter 4 was derived by flipping the negative questions scores (1 to 7) to their opposite (7 to 1), such that a response of 2 on “There are times when Echo cannot be trusted” is a score of 6 after being flipped. These responses (after the negative questions are flipped) are then averaged to get a dyadic trust scale rating between 1 and 7, where higher values indicate higher trust in the robot.

## **B.4 Robotic Social Attributes Scale (RoSAS) Survey**

This survey was administered to the participants of the human subjects experiments detailed in Chapters 4, 5, 6, and 7. In each of these studies, this questionnaire was administered with the rest of the post-experiment questionnaire surveys on a tablet. The questionnaire items were taken directly from [Carpinella et al., 2017], where participants were asked to evaluate how closely they would consider each of the following descriptors to be associated with the robot on a 1 (definitely not associated) to 9 (definitely associated) Likert scale:



**Warmth**

- Happy
- Feeling
- Social
- Organic
- Compassionate
- Emotional

**Competence**

- Capable
- Responsive
- Interactive
- Reliable
- Competent
- Knowledgeable

**Discomfort**

- Scary
- Strange
- Awkward
- Dangerous
- Awful
- Aggressive

The RoSAS scores reported in Chapters 4, 5, and 6 were computed by averaging the Likert values to ascertain one value from 1 to 9 capturing the participant's view of the robot for each of the three sub-scales (warmth, competence, and discomfort).

## B.5 Friendship and Familiarity Scale for Adults

We designed this survey and administered it to the participants who participated in the human subjects experiments detailed in Chapters 5, 6, and 7. In both of these studies, this questionnaire was administered in the pre-experiment survey on a tablet. Each participant would fill out the following questionnaire with respect to each of the other human participants in the experiment.

1. Which of these statements most closely matches your familiarity with this participant?
  - (a) I had not met this participant before we completed this study together; I do not know them.
  - (b) I have seen this participant before and we may have talked once or twice, I do not know them well.
  - (c) I would consider this participant and I acquaintances, we are moderately familiar with each other.
  - (d) This participant and I are friends, we spend / have spent time together outside of work/school together.
  - (e) I would consider this participant to be one of my closest friends.
2. Do you have this participant's phone number?
  - (a) Yes
  - (b) No
3. (Chapter 5 only) Are you Facebook friends with this participant?
  - (a) Yes
  - (b) No
  - (c) I don't have Facebook
4. (Chapter 6 only) Are you connected via social media with this participant (follow/friends with on Facebook, Twitter, Instagram, etc.)?

- (a) Yes
- (b) No

From the participants' answers to this survey, we calculated a final friendship and familiarity score by adding their response to each question (1: [0, 4], 2: [0, 1], and 3/4: [0, 1]) to get a final score between 0 and 6. A participant response of "I don't have Facebook" to question 3 was scored as 0.

## **B.6 The Abbreviated Form of the Revised Eysenck Personality Questionnaire (EPQR-A): Extraversion**

We administered the extraversion component of the abbreviated form of the revised Eysenck personality questionnaire (EPQR-A) [Francis et al., 1992] to the participants who participated in the human subjects experiments detailed in Chapters 5, 6, and 7. In the study described in Chapter 5, this questionnaire was administered in the post-experiment survey and in Chapter 6 and 7, this questionnaire was administered in the pre-experiment survey. Participants completed these surveys in both studies on a tablet. Participants were asked to answer the following questions quickly (to capture their initial response), answering either 'yes' or 'no' to each question:

1. Are you a talkative person?
2. Are you rather lively?
3. Can you easily get some life into a rather dull party?
4. Do you tend to keep in the background on social occasions?
5. Are you mostly quiet when you are with other people?
6. Do other people think of you as being very lively?

The extraversion score used in Chapters 5 and 6 was derived by giving each 'yes' answer a score of 1 and each 'no' a score of 0 (flipping questions 4 and 5), and summing the scores

to each individual question in order to arrive at one value between 0 and 6. The higher the extraversion score, the higher that participant's self-rated extraversion. For the analysis of the conversational dynamics only in Chapter 5, extraversion scores were binned into 0 for raw extraversion score values of 0 and 1, and 1 for raw extraversion score values 2 to 6.

## B.7 The Team Psychological Safety Scale

We administered the psychological safety scale [Edmondson, 1999] to the participants who participated in the human subjects studies detailed in Chapters 5, 6, and 7. In both of these studies, this questionnaire was administered in the post-experiment survey on a tablet. Participants were asked to rate each of the following items on a Likert scale from 1 (strongly disagree) to 7 (strongly agree):

1. If you make a mistake on this team, it is often held against you.
2. Members of this team are able to bring up problems and tough issues.
3. People on this team sometimes reject others for being different.
4. It is safe to take a risk on this team.
5. It is difficult to ask other members of this team for help.
6. No one on this team would deliberately act in a way that undermines my efforts.
7. Working with members of this team, my unique skills and talents are valued and utilized.

The psychological safety scale score was derived by flipping the negative questions scores (1 to 7) to their opposite (7 to 1), such that a response of 2 on “it is difficult to ask other members of this team for help” is a score of 6 after being flipped. These responses, after the negative questions are flipped, are then averaged to get a psychological safety value between 1 and 7, where higher values indicate a higher sense of psychological safety from the participant.

## **B.8 The Short Form of the Trait Emotional Intelligence Questionnaire (TEIQue-SF)**

We administered the Short Form of the Trait Emotional Intelligence Questionnaire (TEIQue-SF) [Cooper and Petrides, 2010] to the participants who participated in the human subjects experiment detailed in Chapters 6 and 7. This questionnaire was administered in the pre-experiment survey on a tablet, where participants rated each item on a Likert scale of 1 (strongly disagree) to 7 (strongly agree):

1. Expressing my emotions with words is not a problem for me.
2. I often find it difficult to see things from another person's viewpoint.
3. On the whole, I'm a highly motivated person.
4. I usually find it difficult to regulate my emotions.
5. I generally don't find life enjoyable.
6. I can deal effectively with people.
7. I tend to change my mind frequently.
8. Many times, I can't figure out what emotion I'm feeling.
9. I feel that I have a number of good qualities.
10. I often find it difficult to stand up for my rights.
11. I'm usually able to influence the way other people feel.
12. On the whole, I have a gloomy perspective on most things.
13. Those close to me often complain that I don't treat them right.
14. I often find it difficult to adjust my life according to the circumstances.
15. On the whole, I'm able to deal with stress.
16. I often find it difficult to show my affection to those close to me.

17. I'm normally able to "get into someone's shoes" and experience their emotions.
18. I normally find it difficult to keep myself motivated.
19. I'm usually able to find ways to control my emotions when I want to.
20. On the whole, I'm pleased with my life.
21. I would describe myself as a good negotiator.
22. I tend to get involved in things I later wish I could get out of.
23. I often pause and think about my feelings.
24. I believe I'm full of personal strengths.
25. I tend to "back down" even if I know I'm right.
26. I don't seem to have any power at all over other people's feelings.
27. I generally believe that things will work out fine in my life.
28. I find it difficult to bond well even with those close to me.
29. Generally, I'm able to adapt to new environments.
30. Others admire me for being relaxed.

The emotional intelligence score we used in our analysis was derived by flipping the negative questions scores (1 to 7) to their opposite (7 to 1), such that a response of 2 on "I often find it difficult to see things from another person's viewpoint" is a score of 6 after being flipped. The items that were flipped were: 2, 4, 5, 7, 8, 10, 12, 13, 14, 16, 18, 22, 25, 26, 28. These responses, after the negative questions are flipped, are then averaged to get an emotional intelligence value between 1 and 7, where higher values indicate higher emotional intelligence.

## B.9 The Perceived Inclusion Scale

We administered the perceived inclusion scale [Jansen et al., 2014] to the participants who participated in the human subjects experiment detailed in Chapters 6 and 7. In this study, this questionnaire was administered in the pre-experiment survey on a tablet. Participants were asked to rate each of the following items on a Likert scale from 1 (strongly disagree) to 5 (strongly agree):

1. This group gives me the feeling that I belong.
2. This group gives me the feeling that I am part of this group.
3. This group gives me the feeling that I fit in.
4. This group treats me as an insider.
5. This group likes me.
6. This group appreciates me.
7. This group is pleased with me.
8. This group cares about me.
9. This group allows me to be authentic.
10. This group allows me to be who I am.
11. This group allows me to express my authentic self.
12. This group allows me to present myself the way I am.
13. This group encourages me to be authentic.
14. This group encourages me to be who I am.
15. This group encourages me to express my authentic self.
16. This group encourages me to present myself the way I am.

We averaged participant responses to these items resulting in a perceived score value between 1 and 5, where higher values indicate higher perceived inclusion.

## Appendix C

# Detailed Statistical Results

In this appendix section, we provide the full detailed statistical analyses for all of the results reported in this dissertation. The results are ordered based on their order of appearance in the paper.



Table C.1: This table lists the demographic characteristics of the participants in the human-subjects study detailed in Chapter 3, both overall and for each condition.

Overall					
Statistic	N	Mean	St. Dev.	Min	Max
age	80	7.25	1.05	6	9
gender: male(0) or female(1)	80	0.48	0.50	0	1
group gender: same(0) or mixed(1)	80	0.50	0.50	0	1
friendship & familiarity score	80	1.25	0.44	1	2

Relational Condition					
Statistic	N	Mean	St. Dev.	Min	Max
age	28	7.29	0.98	6	9
gender: male(0) or female(1)	28	0.54	0.51	0	1
group gender: same(0) or mixed(1)	28	0.36	0.49	0	1
friendship & familiarity score	28	1.43	0.50	1	2

Task Condition					
Statistic	N	Mean	St. Dev.	Min	Max
age	28	7.29	1.05	6	9
gender: male(0) or female(1)	28	0.46	0.51	0	1
group gender: same(0) or mixed(1)	28	0.50	0.51	0	1
friendship & familiarity score	28	1.21	0.42	1	2

Control Condition					
Statistic	N	Mean	St. Dev.	Min	Max
age	24	7.17	1.17	6	9
gender: male(0) or female(1)	24	0.42	0.50	0	1
group gender: same(0) or mixed(1)	24	0.67	0.48	0	1
friendship & familiarity score	24	1.08	0.28	1	2

Table C.2: This table presents the results from the 1-way ANOVA analysis comparing the maximum rocket height difference between the three conditions (relational, task, and control) in Chapter 3 Section 3.3. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>
	Maximum Rocket Height
condition	F(2) = 1.901 (0.090)
group gender composition (2M, 1M1F, 2F)	F(2) = 0.421 (0.003)
average friendship & familiarity score	F(1) = 1.278 (0.029)
age	F(1) = 6.154* (0.157)
Observations	40
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.3: This table presents the results from the 1-way ANOVA analysis comparing the maximum rocket height difference between the task and control conditions (planned comparison) in Chapter 3 Section 3.3. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>
	Maximum Rocket Height
condition (task vs. control)	F(1) = 4.851* (0.117)
group gender composition (2M, 1M1F, 2F)	F(2) = 4.334* (0.191)
average friendship & familiarity score	F(1) = 0.005 (0.003)
age	F(1) = 1.835 (0.084)
Observations	26
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.4: This table presents the results from the 1-way ANOVA analysis comparing the maximum rocket height difference between the task and relational conditions (planned comparison) in Chapter 3 Section 3.3. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>
	Maximum Rocket Height
condition (task vs. relational)	F(1) = 1.167 (0.077)
group gender composition (2M, 1M1F, 2F)	F(2) = 0.093 (0.010)
average friendship & familiarity score	F(1) = 0.316 (0.017)
age	F(1) = 7.092* (0.244)
Observations	26
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.5: This table presents the results from the linear mixed-effects model run in Chapter 3 Section 3.3 examining the influence of the experimental condition (reference group: relational condition) and a control for whether the pair was same or mixed gender on the participant’s perception of their team’s performance. Each participant is grouped with their partner in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	Perception of Performance
condition - control	−0.523 (0.332)
condition - task	−0.926** (0.312)
same(0) or mixed(1) gender pair	0.480 . (0.267)
Constant	2.829*** (0.239)
Observations	80
Log Likelihood	−117.662
Akaike Inf. Crit.	247.324
Bayesian Inf. Crit.	261.616
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.6: This table lists the demographic characteristics of the participants in the human-subjects study detailed in Chapter 4, both overall and for each condition.

Overall

Statistic	N	Mean	St. Dev.	Min	Max
age	82	20.85	2.13	18	32
gender: male (0) or female (1)	82	0.60	0.49	0	1

Competence Apology Condition

Statistic	N	Mean	St. Dev.	Min	Max
age	21	21.33	3.01	18	32
gender: male (0) or female (1)	21	0.62	0.50	0	1

Competence Denial Condition

Statistic	N	Mean	St. Dev.	Min	Max
age	21	20.86	1.88	18	27
gender: male (0) or female (1)	21	0.57	0.51	0	1

Integrity Apology Condition

Statistic	N	Mean	St. Dev.	Min	Max
age	20	20.30	1.26	18	22
gender: male (0) or female (1)	20	0.60	0.50	0	1

Integrity Denial Condition

Statistic	N	Mean	St. Dev.	Min	Max
age	20	20.90	2.00	18	25
gender: male (0) or female (1)	20	0.60	0.50	0	1

Table C.7: This table presents the results from the logistic regression model run in Chapter 4 Section 4.4.1 examining the influence of the trust violation framing and trust repair strategy on whether participants immobilized the robot in their first power-up choice. We used the R ‘glm’ function with a binomial family and logit link to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	first power-up choice:
	immobilize (1) or not (0)
trust violation framing: competence (0) or integrity (1)	1.154* (0.516)
trust repair strategy: apology (0) or denial (1)	1.142* (0.521)
age	−0.132 (0.145)
gender: male (0) or female (1)	−0.240 (0.511)
Constant	0.923 (2.991)
Observations	82
Log Likelihood	−46.326
Akaike Inf. Crit.	102.653
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.8: In order to analyze the power-up choices of participants over time in Chapter 4 Section 4.4.1, we used a multilevel mixed-effects logistic regression model to determine the influence of the trust violation framing and trust repair strategy on whether participants chose to immobilize the robot during their three power-up choices. To capture the repeated measures nature of the data, we used a random effect for each participant across their three power-up choices. In addition to our experimental conditions, we also controlled for the participants' gender and the power-up choice round. We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	power-up choice:
	immobilize (1) or not (0)
trust violation framing: competence (0) or integrity (1)	9.186** (3.065)
trust repair strategy: apology (0) or denial (1)	0.709 (2.540)
gender: female (1) or male (0)	-0.685 (1.892)
power-up choice round [1-3]	-0.958 (0.982)
trust violation framing * power-up choice round	-6.738** (2.086)
trust repair strategy * power-up choice round	0.922 (1.128)
Constant	-9.169*** (2.783)
Observations	246
Log Likelihood	-93.629
Akaike Inf. Crit.	203.257
Bayesian Inf. Crit.	231.300
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001



Table C.9: This table presents the results from the 2-way ANOVA analysis comparing the RoSAS warmth ratings between participants with different trust violation framings and trust repair strategies, as presented in Chapter 4 Section 4.4.2. In addition to our main variables of interest (trust violation framings and trust repair strategy), we include in our model the interaction between these two variables as well as controls for participant gender and age. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>
	RoSAS warmth rating
trust violation framing: competence (0) or integrity	F(1) = 0.299 (0.013)
trust repair strategy: apology (0) or denial (1)	F(1) = 8.190** (0.121)
trust violation framing * trust repair strategy	F(1) = 0.000 ( $< 0.001$ )
gender: female (1) or male (0)	F(1) = 0.664 (0.009)
age	F(1) = 1.990 . (0.184)
Observations	82
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.10: This table presents the results from the 2-way ANOVA analysis comparing the Dyadic Trust Scale (DTS) ratings between participants with different trust violation framings and trust repair strategies, as presented in Chapter 4 Section 4.4.2. In addition to our main variables of interest (trust violation framings and trust repair strategy), we include in our model the interaction between these two variables as well as controls for participant gender and age. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>
	Dyadic Trust Scale rating
trust violation framing: competence (0) or integrity	F(1) = 0.302 ( $< 0.001$ )
trust repair strategy: apology (0) or denial (1)	F(1) = 2.013 (0.012)
trust violation framing * trust repair strategy	F(1) = 4.637* (0.048)
gender: female (1) or male (0)	F(1) = 5.523* (0.055)
age	F(1) = 1.866 . (0.125)
Observations	82
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.11: This table presents the results from the 2-way ANOVA analysis comparing whether the participants perceived the robot to have lied between participants with different trust violation framings and trust repair strategies, as presented in Chapter 4 Section 4.4.2. To gather participants perception of whether the robot lied, we examined participants' ratings on a 1 (strongly disagree) to 7 (strongly agree) Likert scale on the post-experiment survey question "Echo lied to me," where Echo is the name of the robot. In addition to our main variables of interest (trust violation framings and trust repair strategy), we include in our model the interaction between these two variables as well as controls for participant gender and age. This analysis was performed using the 'aov' function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>
	rating of robot having lied
trust violation framing: competence (0) or integrity	F(1) = 1.466 (0.006)
trust repair strategy: apology (0) or denial (1)	F(1) = 1.279 (0.010)
trust violation framing * trust repair strategy	F(1) = 7.272** (0.073)
gender: female (1) or male (0)	F(1) = 0.819 (0.023)
age	F(1) = 2.269* (0.187)
Observations	82
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.12: This table presents the results from the 2-way ANOVA analysis comparing whether participants believed they had made a reciprocal promise to the robot across our experimental conditions, as presented in Chapter 4 Section 4.4.2. To gather participants belief of whether or not they made a reciprocal promise to the robot, we examined participants' ratings on a 1 (strongly disagree) to 7 (strongly agree) Likert scale on the post-experiment survey question "I promised not to immobilize Echo during the game," where Echo is the name of the robot. In addition to our main variables of interest (trust violation framings and trust repair strategy), we include in our model the interaction between these two variables as well as controls for participant gender and age. This analysis was performed using the 'aov' function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>
	participant promise rating
trust violation framing: competence (0) or integrity	F(1) = 0.829 (0.002)
trust repair strategy: apology (0) or denial (1)	F(1) = 1.758 (0.011)
trust violation framing * trust repair strategy	F(1) = 0.415 (0.005)
gender: female (1) or male (0)	F(1) = 0.456 (0.001)
age	F(1) = 0.953 (0.103)
Observations	82
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.13: This table presents the results from the logistic regression model run in Chapter 4 Section 4.4.3 examining the influence of a participants' promise not to immobilize the robot (post-experiment survey Likert rating of "I promised not to immobilize Echo during the game") on whether participants immobilized the robot in their first power-up choice. We also control for the trust violation framing, trust repair strategy, participant age, and participant gender. We used the R 'glm' function with a binomial family and logit link to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	first power-up choice:
	immobilize (1) or not (0)
participant promise to robot Likert rating	−0.582** (0.197)
trust violation framing: competence (0) or integrity (1)	1.394* (0.583)
trust repair strategy: apology (0) or denial (1)	1.209* (0.592)
gender: female (1) or male (0)	−0.213 (0.568)
age	−0.278 (0.173)
Constant	5.026 (3.632)
Observations	82
Log Likelihood	−39.254
Akaike Inf. Crit.	90.507
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.14: This table presents the results from the logistic regression model run in Chapter 4 Section 4.4.3 examining the influence of a participants' promise not to immobilize the robot (post-experiment survey Likert rating of "I promised not to immobilize Echo during the game") on whether participants ever immobilized the robot in any of their three power-up choices. We also control for the trust violation framing, trust repair strategy, participant age, and participant gender. We used the R 'glm' function with a binomial family and logit link to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	all power-up choices:
	at least one immobilize (1)
	or never immobilize (0)
participant promise to robot Likert rating	−0.739*** (0.217)
trust violation framing: competence (0) or integrity (1)	0.997 . (0.583)
trust repair strategy: apology (0) or denial (1)	1.549* (0.606)
gender: female (1) or male (0)	−0.512 (0.583)
age	−0.339 . (0.179)
Constant	7.144 . (3.796)
Observations	82
Log Likelihood	−37.947
Akaike Inf. Crit.	87.894
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.15: This table presents the results from the linear regression model run in Chapter 4 Section 4.4.3 examining the influence of a participants’ promise not to immobilize the robot (post-experiment survey Likert rating of “I promised not to immobilize Echo during the game”) on their Dyadic Trust Scale (DTS) ratings of the robot. We also control for the trust violation framing, trust repair strategy, participant age (considered as a factor in [Strohkorb Sebo et al., 2019], but was not considered a factor for this analysis), and participant gender. We used the R ‘lm’ function to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	Dyadic Trust Scale rating
participant promise to robot Likert rating	0.158*** (0.042)
trust violation framing: competence (0) or integrity (1)	−0.049 (0.187)
trust repair strategy: apology (0) or denial (1)	−0.171 (0.187)
gender: female (1) or male (0)	−0.424* (0.190)
age	0.069 (0.045)
Constant	1.583 (0.985)
Observations	82
R <sup>2</sup>	0.234
Adjusted R <sup>2</sup>	0.184
Residual Std. Error	0.838 (df = 76)
F Statistic	4.651*** (df = 5; 76)
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.16: This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 5, both overall and for each condition.

Overall

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	153	21.73	8.76	14	17	22	59
gender (0-M, 1-F)	153	0.62	0.49	0	0	1	1
avg. familiarity	153	0.73	1.12	0.00	0.00	1.50	4.50
extraversion	153	3.70	2.20	0	2	6	6

Vulnerable Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	54	20.13	7.13	14	17	20.8	55
gender (0-M, 1-F)	54	0.52	0.50	0	0	1	1
avg. familiarity	54	0.78	1.08	0	0	1.5	4
extraversion	54	3.50	2.28	0	1	6	6

Neutral Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	51	21.33	11.01	15	17	19	59
gender (0-M, 1-F)	51	0.71	0.46	0	0	1	1
avg. familiarity	51	1.21	1.35	0.00	0.00	2.00	4.50
extraversion	51	3.88	2.08	0	2	6	6

Silent Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	48	23.94	7.36	15	19	27.2	48
gender (0-M, 1-F)	48	0.65	0.48	0	0	1	1
avg. familiarity	48	0.18	0.48	0	0	0	2
extraversion	48	3.73	2.26	0	2	6	6



Table C.17: This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s Likert rating (1 - strongly disagree, 7 - strongly agree) on the post-experiment questionnaire item “Echo [the robot] made vulnerable disclosures about Echo’s feelings during the interaction.” Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	the robot made vulnerable disclosures
condition - neutral	−2.891*** (0.349)
condition - silent	−2.019*** (0.355)
Constant	5.185*** (0.243)
Observations	153
Log Likelihood	−302.316
Akaike Inf. Crit.	614.633
Bayesian Inf. Crit.	629.785
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.18: This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s Likert rating (1 - strongly disagree, 7 - strongly agree) on the post-experiment questionnaire item “Echo [the robot] told personal stories during the interaction.” Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	the robot told personal stories
condition - neutral	−4.797*** (0.264)
condition - silent	−4.382*** (0.268)
Constant	6.444*** (0.184)
Observations	153
Log Likelihood	−247.899
Akaike Inf. Crit.	505.799
Bayesian Inf. Crit.	520.951
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.19: This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s Likert rating (1 - strongly disagree, 7 - strongly agree) on the post-experiment questionnaire item “Echo [the robot] made use of humor during the interaction.” Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	the robot used humor
condition - neutral	−2.614*** (0.376)
condition - silent	−3.368*** (0.382)
Constant	6.222*** (0.262)
Observations	153
Log Likelihood	−286.386
Akaike Inf. Crit.	582.772
Bayesian Inf. Crit.	597.925
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.20: This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s rating of the robot’s warmth, according to the RoSAS scale [Carpinella et al., 2017]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	robot’s warmth (RoSAS)
condition - neutral	−1.182*** (0.302)
condition - silent	−1.880*** (0.307)
Constant	6.130*** (0.211)
Observations	153
Log Likelihood	−273.733
Akaike Inf. Crit.	557.467
Bayesian Inf. Crit.	572.619
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.21: This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) and a control for the participants' gender on the participant's rating of the robot's competence, according to the RoSAS scale [Carpinella et al., 2017]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	robot's competence (RoSAS)
condition - neutral	−0.070 (0.372)
condition - silent	−0.856* (0.376)
gender: female (1) or male (0)	0.735** (0.281)
Constant	5.545*** (0.295)
Observations	153
Log Likelihood	−296.615
Akaike Inf. Crit.	605.229
Bayesian Inf. Crit.	623.412
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.22: This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) and a control for the participants' age on the participant's rating of the robot's competence, according to the RoSAS scale [Carpinella et al., 2017]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	robot's discomfort (RoSAS)
condition - neutral	−0.082 (0.239)
condition - silent	0.437 . (0.246)
age	−0.017 (0.011)
Constant	2.597*** (0.284)
Observations	153
Log Likelihood	−251.543
Akaike Inf. Crit.	515.086
Bayesian Inf. Crit.	533.269
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.23: In order to analyze whether or not a human team member who made a mistake looked at the robot afterwards in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (reference group: vulnerable condition) on whether or not a human team member who made a mistake looked at the robot afterwards. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age and gender. We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	did the mistake maker look at the robot after the mistake:
	yes (1) or no (0)
condition - neutral	−1.070** (0.401)
condition - silent	−2.636*** (0.456)
age	−0.033. (0.018)
gender: female (1) or male (0)	0.569. (0.312)
Constant	2.022*** (0.528)
Observations	274
Log Likelihood	−151.638
Akaike Inf. Crit.	315.277
Bayesian Inf. Crit.	336.956
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.24: In order to analyze participants' verbal responses to the robot in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (reference group: vulnerable condition) on whether or not a human team member spoke to the robot after a mistake was made by the team. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' extraversion score and average familiarity with the other participants as well as the mistake round number (1-8). We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	did the participant speak to the robot: yes (1) or no (0)
condition - neutral	-0.954*** (0.263)
condition - silent	-1.497*** (0.317)
mistake round [1,8]	0.147*** (0.042)
average familiarity	0.158 . (0.092)
extraversion [0,6]	0.152** (0.049)
Constant	-2.954*** (0.341)
Observations	1,224
Log Likelihood	-386.129
Akaike Inf. Crit.	786.258
Bayesian Inf. Crit.	822.027
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	



Table C.25: In order to analyze whether or not a human team member who made a mistake explained that mistake to their team members in Chapter 5 Section 5.4, we used a multi-level mixed-effects logistic regression model to determine the influence of the experimental condition (reference group: vulnerable condition) on whether or not a human team member who made a mistake explained that mistake to their team members. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age and average familiarity with the other participants as well as the mistake round number (1-8). We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	did mistake maker
	explain the mistake:
	yes (1) or no (0)
condition - neutral	-1.143* (0.577)
condition - silent	-1.679** (0.622)
mistake round [1,8]	-0.111 . (0.066)
age	-0.022 (0.021)
average familiarity	0.331 . (0.188)
Constant	1.681** (0.652)
Observations	297
Log Likelihood	-176.764
Akaike Inf. Crit.	367.528
Bayesian Inf. Crit.	393.384
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.26: In order to analyze whether or not a human team member consoled the one who made a mistake in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition, where the neutral and silent conditions are pooled together, on whether or not a human team member consoled the one who made a mistake. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age as well as the mistake round number (1-8). We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	did human team member
	console mistake maker:
	yes (1) or no (0)
condition (vulnerable vs. neutral + silent)	0.856* (0.396)
mistake round [1,8]	-0.185*** (0.046)
age	-0.063* (0.027)
Constant	-0.421 (0.645)
Observations	927
Log Likelihood	-339.751
Akaike Inf. Crit.	689.503
Bayesian Inf. Crit.	713.663

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

Table C.27: In order to analyze whether or not a human team member consoled the person who made a mistake (excluding consoling the robot) in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition, where the neutral and silent conditions are pooled together, on whether or not a human team member consoled the person who made a mistake (excluding consoling the robot). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age as well as the mistake round number (1-8). We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	did human team member console
	human (not robot) mistake maker: yes (1) or no (0)
condition (vulnerable vs. neutral + silent)	0.525 (0.466)
mistake round [1,8]	-0.139* (0.055)
age	-0.058 . (0.030)
average familiarity	0.333* (0.152)
Constant	-0.724 (0.728)
Observations	624
Log Likelihood	-257.878
Akaike Inf. Crit.	527.756
Bayesian Inf. Crit.	554.373
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.28: In our first analysis whether or not a human team members laughed together in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (vulnerable or neutral condition) on whether or not a human team member laughed along with another human team member. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age and average familiarity with the other participants. We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	did human team member
	laugh with another:
	yes (1) or no (0)
condition - vulnerable	0.791 (0.395)*
age	0.031* (0.016)
average familiarity	0.235* (0.106)
Constant	−2.658*** (0.465)
Observations	840
Log Likelihood	−426.092
Akaike Inf. Crit.	862.184
Bayesian Inf. Crit.	885.851
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.29: In order to analyze whether or not a human team members laughed together in Chapter 5 Section 5.4, we used a multilevel mixed-effects logistic regression model to determine the influence of the experimental condition (reference group: vulnerable condition) on whether or not a human team member laughed along with another human team member. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. In addition to our experimental conditions, we also controlled for the participants' age and average familiarity with the other participants, these controls were scaled to ensure model convergence. We used the R 'glmer' function with a binomial family and logit link from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	did human team member
	laugh with another:
	yes (1) or no (0)
condition - neutral	-0.772 (0.551)
condition - silent	-0.973 . (0.587)
age	0.216 (0.143)
average familiarity	0.277* (0.127)
Constant	-1.141** (0.382)
Observations	1,224
Log Likelihood	-569.461
Akaike Inf. Crit.	1,150.922
Bayesian Inf. Crit.	1,181.581
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.30: In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on total individual speaking time. We used a multilevel linear model of speaking time (s) as a function of experimental condition (reference group: vulnerable robot) and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity was modeled using random effects clustered in groups. We used the R ‘lme’ function from the ‘nlme’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	total talking time (s)
condition - silent	−124.523** (41.046)
condition - neutral	−140.678*** (39.973)
age	0.180 (1.272)
gender: female (1) or male (0)	−15.760 (18.918)
extraversion [0,1]	45.002* (22.314)
average familiarity	18.163 (11.264)
Constant	212.348*** (41.790)
Observations	153
Log Likelihood	−926.594
Akaike Inf. Crit.	1,871.188
Bayesian Inf. Crit.	1,898.040
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.31: In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on total individual speaking time. We used a multilevel linear model of speaking time (s) as a function of experimental condition (reference group: vulnerable robot) including an interaction of the treatment effect with round and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in participants in groups. We used the R ‘lme’ function from the ‘nlme’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	total talking time per round (s)
round * condition - silent	−0.068 (0.061)
round * condition - neutral	−0.129* (0.060)
round	0.153*** (0.042)
condition - silent	−3.190* (1.230)
condition - neutral	−2.537* (1.201)
age	0.009 (0.034)
gender: female (1) or male (0)	0.461 (0.501)
extraversion [0,1]	0.871 (0.593)
avgerage familiarity	0.422 (0.307)
Constant	4.280** (1.381)
Observations	4,590
Log Likelihood	−13,872.010
Akaike Inf. Crit.	27,778.020
Bayesian Inf. Crit.	27,887.320

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

Table C.32: In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on the duration of human team member responses. We used a multilevel linear model of speaking time (s) as a function of experimental condition (reference group: vulnerable robot) including an interaction of the treatment effect with round and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in participants in groups. We used the R ‘lme’ function from the ‘nlme’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	total time responding per round (s)
round	0.095*** (0.028)
condition - silent	−1.312* (0.630)
condition - neutral	−1.059. (0.616)
age	−0.003 (0.016)
gender: female (1) or male (0)	−0.209 (0.230)
extraversion [0,1]	−0.036 (0.273)
average familiarity	0.080 (0.144)
round * condition - silent	−0.044 (0.040)
round * condition - neutral	−0.082* (0.040)
Constant	2.857*** (0.666)
Observations	4,590
Log Likelihood	−12,219.810
Akaike Inf. Crit.	24,473.630
Bayesian Inf. Crit.	24,582.930
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	



Table C.33: In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on participants equality in talking time ( $E_{TT_i}$ ). We used a multilevel beta regression as a function of experimental condition (reference group: vulnerable robot) and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in groups. We used the R ‘glmmTMB’ function with a beta\_family and logit link from the ‘glmmTMB’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. Because a beta regression cannot analyze 0’s or 1’s (a few participants had values of 1), we transformed the data using the following equation, where  $N$  is the sample size (150) and  $y$  is the outcome variable [Smithson and Verkuilen, 2006]:  $y' = \frac{y*(N-1)+0.5}{N}$ .

	<i>Dependent variable:</i>
	$E_{TT_i}$
condition - neutral	0.027 (0.184)
condition - silent	0.661*** (0.186)
age	0.00009 (0.008)
gender: female (1) or male (0)	−0.276* (0.134)
extraversion [0,1]	0.173 (0.166)
average familiarity	−0.057 (0.071)
Constant	−1.753*** (0.252)
Observations	150
Log Likelihood	134.0
Akaike Inf. Crit.	−250.0
Bayesian Inf. Crit.	−222.9
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.34: In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on participants equality in talking partners ( $E_{TP_i}$ ). We used a multilevel beta regression as a function of experimental condition (reference group: vulnerable robot) and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in groups. We used the R ‘glmmTMB’ function with a beta\_family and logit link from the ‘glmmTMB’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table. Because a beta regression cannot analyze 0’s or 1’s (a few participants had values of 1), we transformed the data using the following equation, where  $N$  is the sample size (144) and  $y$  is the outcome variable [Smithson and Verkuilen, 2006]:  $y' = \frac{y*(N-1)+0.5}{N}$ .

	<i>Dependent variable:</i>
	$E_{TP_i}$
condition - neutral	0.384 (0.276)
condition - silent	0.742* (0.300)
age	0.021 . (0.012)
gender: female (1) or male (0)	−0.101 (0.196)
extraversion [0,1]	−0.774*** (0.233)
average familiarity	−0.277** (0.102)
Constant	0.058 (0.345)
Observations	
Log Likelihood	
Akaike Inf. Crit.	
Bayesian Inf. Crit.	
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.35: This table presents the results from the linear mixed-effects model run in Chapter 5 Section 5.4 examining the influence of the experimental condition (reference group: vulnerable condition) on the participant’s rating of their team’s psychological safety, according to Edmondson’s Team Psychological Safety scale [Edmondson, 1999]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	psychological safety
condition - neutral	−0.087 (0.199)
condition - silent	−0.312 (0.203)
Constant	5.619*** (0.139)
Observations	153
Log Likelihood	−187.603
Akaike Inf. Crit.	385.206
Bayesian Inf. Crit.	400.359
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.36: In Chapter 5 Section 5.4 we examined the treatment effect of vulnerable robot versus neutral and silent robot utterances on different self-reported group dynamics. We used a multilevel logistic model as a function of experimental condition (reference group: vulnerable robot) and controls for age, gender, extraversion and familiarity. Unobserved individual heterogeneity were modeled using random effects clustered in groups. We used the R ‘glmer’ function with a binomial family and logit link from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>			
	positive (1) or			
	quiet	negative (0)	supportive	fun
condition - silent	0.679 (0.564)	−1.764* (0.714)	−0.869 (0.669)	−1.740* (0.732)
condition - neutral	1.277* (0.571)	−1.355* (0.657)	0.147 (0.504)	−1.443* (0.667)
age	−0.004 (0.025)	0.005 (0.026)	−0.054 (0.044)	0.0001 (0.030)
gender (0-M, 1-F)	0.192 (0.415)	0.354 (0.452)	0.283 (0.489)	0.063 (0.513)
extraversion [0, 1]	−0.144 (0.480)	0.682 (0.504)	0.487 (0.599)	0.180 (0.589)
average familiarity	−0.550* (0.243)	−0.093 (0.226)	−0.140 (0.229)	0.090 (0.239)
Constant	−1.078 (0.735)	0.790 (0.776)	−0.841 (1.060)	−1.126 (0.832)
Observations	153	153	153	153
Log Likelihood	−89.038	−94.272	−64.204	−65.328
Akaike Inf. Crit.	194.077	204.544	144.407	146.656
Bayesian Inf. Crit.	218.320	228.787	168.651	170.900

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

Table C.37: This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 6 both overall and for each experimental condition.

Overall

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	78	16.82	0.72	16	16	17	19
gender (0-M, 1-F)	78	0.49	0.50	0	0	1	1
extraversion	78	3.90	2.15	0	2	6	6
emotional intelligence	78	5.27	0.65	3.37	4.95	5.70	6.63
avg. familiarity	78	1.10	1.06	0.00	0.00	2.00	3.50

Ingroup Robot Liaison Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	39	16.72	0.69	16	16	17	18
gender (0-M, 1-F)	39	0.46	0.51	0	0	1	1
extraversion	39	3.36	2.44	0	1	6	6
emotional intelligence	39	5.32	0.67	3.37	5.07	5.72	6.20
avg. familiarity	39	0.99	1.13	0	0	1.5	4

Outgroup Robot Liaison Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	39	16.92	0.74	16	16	17	19
gender (0-M, 1-F)	39	0.51	0.51	0	0	1	1
extraversion	39	4.44	1.68	0	4	6	6
emotional intelligence	39	5.23	0.64	3.60	4.88	5.50	6.63
avg. familiarity	39	1.21	0.99	0.00	0.50	2.00	3.00

Table C.38: This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 6 for each important subdivision of participants (ingroup/outgroup, robot liaison).

Ingroup Members

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	52	16.75	0.65	16	16	17	18
gender (0-M, 1-F)	52	0.44	0.50	0	0	1	1
extraversion	52	3.65	2.28	0	1	6	6
emotional intelligence	52	5.21	0.68	3.37	4.86	5.70	6.27
avg. familiarity	52	1.12	1.05	0	0	2	4

Outgroup Members

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	26	16.96	0.82	16	16	17	19
gender (0-M, 1-F)	26	0.58	0.50	0	0	1	1
extraversion	26	4.38	1.81	0	4	6	6
emotional intelligence	26	5.40	0.57	4.37	5.13	5.77	6.63
avg. familiarity	26	1.06	1.12	0.00	0.00	1.50	3.50

Robot Liaison Members

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	26	16.92	0.84	16	16	17	19
gender (0-M, 1-F)	26	0.54	0.51	0	0	1	1
extraversion	26	3.81	2.45	0	1.2	6	6
emotional intelligence	26	5.38	0.69	3.37	5.13	5.83	6.63
avg. familiarity	26	0.98	1.00	0.00	0.00	1.88	3.00

Non Robot Liaison Members

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	52	16.77	0.65	16	16	17	18
gender (0-M, 1-F)	52	0.46	0.50	0	0	1	1
extraversion	52	3.94	2.01	0	2	6	6
emotional intelligence	52	5.22	0.63	3.60	4.83	5.65	6.27
avg. familiarity	52	1.15	1.10	0	0	2	4

Table C.39: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the correlation of the participant designations of robot liaison and participant designations of ingroup-outgroup with the average familiarity with their two human team members. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	avg. familiarity
robot liaison: yes (1) or no (0)	−0.253 (0.296)
intergroup bias: ingroup (0) or outgroup (1)	−0.099 (0.296)
robot liaison * intergroup bias	0.209 (0.499)
Constant	1.194*** (0.211)
Observations	78
Log Likelihood	−110.544
Akaike Inf. Crit.	233.088
Bayesian Inf. Crit.	247.228
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.40: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the correlation of the participant designations of robot liaison and participant designations of ingroup-outgroup with their emotional intelligence. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	emotional intelligence
robot liaison: yes (1) or no (0)	0.135 (0.209)
intergroup bias: ingroup (0) or outgroup (1)	0.168 (0.209)
robot liaison * intergroup bias	−0.020 (0.330)
Constant	5.175*** (0.104)
Observations	78
Log Likelihood	−79.089
Akaike Inf. Crit.	170.179
Bayesian Inf. Crit.	184.319
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001



Table C.41: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the correlation of the participant designations of robot liaison and participant designations of ingroup-outgroup with their extraversion. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	extraversion
robot liaison: yes (1) or no (0)	−1.182 . (0.659)
intergroup bias: ingroup (0) or outgroup (1)	−0.028 (0.659)
robot liaison * intergroup bias	2.109* (1.068)
Constant	3.997*** (0.359)
Observations	78
Log Likelihood	−165.434
Akaike Inf. Crit.	342.867
Bayesian Inf. Crit.	357.007
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.42: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for emotional intelligence, on the similarity of their survival item rankings from round 1 with their survival item rankings from round 2 (smaller values indicate higher similarity of the lists). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	survival item ranking
	similarity (rounds 1 & 2)
robot liaison: yes (1) or no (0)	0.030 (0.036)
intergroup bias: ingroup (0) or outgroup (1)	0.103** (0.036)
robot liaison * intergroup bias	0.011 (0.056)
emotional intelligence	-0.046* (0.020)
Constant	0.677*** (0.104)
Observations	78
Log Likelihood	49.638
Akaike Inf. Crit.	-85.276
Bayesian Inf. Crit.	-68.779
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.43: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for participant age and emotional intelligence, on their partner preference score. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	partner preference score
robot liaison: yes (1) or no (0)	−0.109 (0.211)
intergroup bias: ingroup (0) or outgroup (1)	−0.508* (0.211)
robot liaison * intergroup bias	0.254 (0.328)
age	0.207 . (0.106)
emotional intelligence	0.442*** (0.121)
Constant	−4.662* (2.021)
Observations	75
Log Likelihood	−75.172
Akaike Inf. Crit.	166.344
Bayesian Inf. Crit.	184.884

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

Table C.44: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for participant age and the maximum familiarity a participant has between their two other human team mates, on their perceived inclusion scale score. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	perceived inclusion
robot liaison: yes (1) or no (0)	−0.407* (0.172)
intergroup bias: ingroup (0) or outgroup (1)	0.041 (0.172)
robot liaison * intergroup bias	0.350 (0.284)
age	−0.203* (0.089)
max. familiarity	0.119* (0.045)
Constant	7.559*** (1.506)
Observations	78
Log Likelihood	−69.616
Akaike Inf. Crit.	155.231
Bayesian Inf. Crit.	174.085
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.45: This table presents the results from the 1-way ANOVA analysis comparing the the proportion of survival items initially ranked low (9-25) on the round 1 ingroup list and high (1-8) on the round 1 outgroup list items ( $L_{in}$ ,  $H_{out}$ ) that made it onto the final list of 8 items produced by the entire team at the end of round two of the experiment. We examined this proportion between participants with an ingroup robot liaison verses an outgroup robot liaison outsider, as presented in Chapter 6 Section 6.4. In addition to our main variable of interest, we include in our model the average familiarity of group members with one another and the number of females on the team. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>
	$L_{in}, H_{out}$ on final list
ingroup robot liaison (0) or outgroup robot liaison (1)	F(1) = 5.594* (0.193)
avg. group familiarity	F(1) = 0.008 ( $< 0.001$ )
number of females [0,3]	F(1) = 0.826 (0.036)
Observations	26
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.46: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for participant extraversion, on the proportion of time they spent talking 1 minute after robot support targeted to participant(RST-P), robot support targeted to someone else (RST-SE), and a robot undirected utterance (RUU). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>		
	RST-P	RST-SE	RUU
robot liaison: yes (1) or no (0)	0.048 (0.031)	0.009 (0.033)	0.030 (0.029)
intergroup bias: ingroup (0) or outgroup (1)	0.012 (0.031)	−0.025 (0.033)	−0.033 (0.029)
robot liaison * intergroup bias	0.006 (0.052)	0.033 (0.056)	0.033 (0.049)
extraversion	0.016** (0.006)	0.017** (0.006)	0.017** (0.005)
Constant	0.120*** (0.030)	0.152*** (0.033)	0.139*** (0.029)
Observations	74	74	74
Log Likelihood	55.405	49.665	71.547
Akaike Inf. Crit.	−96.810	−85.330	−129.094
Bayesian Inf. Crit.	−80.681	−69.201	−112.965
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001		

Table C.47: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with a control for participant extraversion, on the proportion of time they spent talking 1 minute after robot support targeted to participant(RST-P) compared with two controls (via subtraction): robot support targeted to someone else (RST-SE), and a robot undirected utterance (RUU). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	(RTS-P - RTS-SE)	(RTS-P - RUU)
robot liaison: yes (1) or no (0)	0.033* (0.016)	0.017 (0.015)
intergroup bias: ingroup (0) or outgroup (1)	0.033* (0.017)	0.043** (0.016)
robot liaison * intergroup bias	−0.017 (0.026)	−0.025 (0.024)
Constant	−0.030*** (0.008)	−0.020* (0.008)
Observations	74	74
Log Likelihood	105.290	109.557
Akaike Inf. Crit.	−198.581	−207.113
Bayesian Inf. Crit.	−184.756	−193.289
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.48: This table presents the results from the linear mixed-effects model run in Chapter 6 Section 6.4 examining the influence of the participants' intergroup bias (ingroup/outgroup) and robot liaison designation, with various controls either participants' emotional intelligence or the maximum familiarity a participant had with their two fellow participants, on the participants' ratings of the robot's warmth, competence, and discomfort according to the RoSAS scale [Carpinella et al., 2017]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>		
	warmth	competence	discomfort
robot liaison: yes (1) or no (0)	0.403 (0.456)	-0.264 (0.387)	0.226 (0.319)
intergroup bias: ingroup (0) or outgroup (1)	0.198 (0.456)	0.113 (0.388)	-0.033 (0.320)
robot liaison * intergroup bias	-0.303 (0.732)	0.093 (0.611)	-0.076 (0.508)
emotional intelligence		0.596** (0.215)	-0.673*** (0.178)
max. familiarity	0.281* (0.113)		
Constant	5.206*** (0.310)	4.101*** (1.129)	5.689*** (0.934)
Observations	78	78	78
Log Likelihood	-137.963	-124.681	-110.703
Akaike Inf. Crit.	289.925	263.362	235.406
Bayesian Inf. Crit.	306.422	279.859	251.903
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001		



Table C.49: This table lists the demographic and descriptive characteristics of all the participants in the human-subjects study detailed in Chapter 7.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	114	16.73	0.73	15	16	17	19
gender: (0-M, 1-F)	114	0.51	0.50	0	0	1	1
extraversion	114	3.88	2.06	0	2	6	6
emotional intelligence	114	5.25	0.64	3.37	4.91	5.69	6.63
avg. familiarity	114	1.16	1.12	0.00	0.00	2.00	4.00

Table C.50: This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 7 in our 2 robot verbal support (yes or not) x 2 intergroup bias robot liaison (ingroup robot liaison vs. outgroup robot liaison) between subjects design.

Robot Verbal Support & Ingroup Robot Liaison Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	39	16.72	0.69	16	16	17	18
gender: (0-M, 1-F)	39	0.44	0.50	0	0	1	1
extraversion	39	3.36	2.42	0	1	6	6
emotional intelligence	39	5.34	0.66	3.37	5.13	5.70	6.20
avg. familiarity	39	0.92	1.14	0	0	1.5	4

Robot Verbal Support & Outgroup Robot Liaison Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	45	16.82	0.78	15	16	17	19
gender: (0-M, 1-F)	45	0.53	0.50	0	0	1	1
extraversion	45	4.42	1.59	0	4	6	6
emotional intelligence	45	5.20	0.64	3.60	4.93	5.53	6.63
avg. familiarity	45	1.19	1.01	0.00	0.00	2.00	3.00

No Robot Verbal Support & Ingroup Robot Liaison Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	15	16.47	0.74	15	16	17	18
gender: (0-M, 1-F)	15	0.60	0.51	0	0	1	1
extraversion	15	4.20	1.78	0	3.5	5.5	6
emotional intelligence	15	5.42	0.55	4.47	5.08	5.77	6.37
avg. familiarity	15	1.93	1.19	0.00	0.75	2.50	3.50

No Robot Verbal Support & Outgroup Robot Liaison Condition

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	15	16.73	0.70	16	16	17	18
gender: (0-M, 1-F)	15	0.53	0.52	0	0	1	1
extraversion	15	3.27	2.22	0	2	5.5	6
emotional intelligence	15	5.03	0.61	3.63	4.80	5.42	5.80
avg. familiarity	15	0.90	1.06	0.00	0.50	1.00	4.00

Table C.51: This table lists the demographic and descriptive characteristics of the participants in the human-subjects study detailed in Chapter 7 for each important subdivision of participants (ingroup/outgroup, robot liaison).

Ingroup Members

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	76	16.67	0.72	15	16	17	18
gender: (0-M, 1-F)	76	0.49	0.50	0	0	1	1
extraversion	76	3.63	2.11	0	2	5.2	6
emotional intelligence	76	5.18	0.68	3.37	4.83	5.60	6.37
avg. familiarity	76	1.17	1.14	0	0	2	4

Outgroup Members

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	38	16.84	0.75	16	16	17	19
gender: (0-M, 1-F)	38	0.55	0.50	0	0	1	1
extraversion	38	4.37	1.88	0	3.2	6	6
emotional intelligence	38	5.40	0.51	4.37	5.06	5.72	6.63
avg. familiarity	38	1.13	1.11	0.00	0.00	1.88	3.50

Robot Liaison Members

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	38	16.82	0.77	16	16	17	19
gender: (0-M, 1-F)	38	0.47	0.51	0	0	1	1
extraversion	38	3.84	2.28	0	2	6	6
emotional intelligence	38	5.38	0.64	3.37	5.11	5.80	6.63
avg. familiarity	38	1.12	1.08	0.00	0.00	2.00	3.50

Non Robot Liaison Members

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	76	16.68	0.72	15	16	17	18
gender: (0-M, 1-F)	76	0.53	0.50	0	0	1	1
extraversion	76	3.89	1.95	0	2.8	6	6
emotional intelligence	76	5.19	0.63	3.60	4.83	5.60	6.27
avg. familiarity	76	1.18	1.15	0	0	2	4

Table C.52: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of the the backchannels a participant received (sec), with controls for participant gender and emotional intelligence, on their psychological safety score [Edmondson, 1999]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	psychological safety
verbal backchannels received (sec)	0.017** (0.006)
gender: female (1) or male (0)	0.274* (0.127)
emotional intelligence	0.284** (0.101)
Constant	4.094*** (0.529)
Observations	114
Log Likelihood	−123.526
Akaike Inf. Crit.	259.052
Bayesian Inf. Crit.	275.469
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.53: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of participants total time talking in round 2 of the experiment (sec), with controls for ingroup-outgroup bias, robot liaison designation, gender, and emotional intelligence, on their psychological safety score [Edmondson, 1999]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	psychological safety
total talking time (sec)	0.001* (0.0004)
ingroup (0) or outgroup (1)	0.284· (0.149)
robot liaison: yes (1) or no (0)	−0.238 (0.149)
gender: female (1) or male (0)	0.319* (0.135)
emotional intelligence	0.255* (0.109)
Constant	4.197*** (0.558)
Observations	106
Log Likelihood	−120.361
Akaike Inf. Crit.	256.722
Bayesian Inf. Crit.	278.030
<i>Note:</i>	· p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.54: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of the verbal backchannels a participant received (sec) normalized by the total time that participant spent talking (sec), with controls for ingroup-outgroup bias, robot liaison designation, gender, and emotional intelligence, on their psychological safety score [Edmondson, 1999]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	psychological safety
verbal backchannels received (sec) normalized by talking time (sec)	−1.133 (1.059)
ingroup (0) or outgroup (1)	0.315* (0.158)
robot liaison: yes (1) or no (0)	−0.235 (0.153)
gender: female (1) or male (0)	0.319* (0.138)
emotional intelligence	0.284** (0.110)
Constant	4.399*** (0.577)
Observations	106
Log Likelihood	−114.368
Akaike Inf. Crit.	244.736
Bayesian Inf. Crit.	266.044
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.55: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of the total time a participant received verbal backchannels (sec), with controls for robot liaison designation and emotional intelligence, on their perceived group inclusion score [Jansen et al., 2014]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	perceived inclusion
verbal backchannels received (sec)	0.006 (0.005)
robot liaison: yes (1) or no (0)	−0.239* (0.108)
emotional intelligence	0.300*** (0.082)
Constant	2.715*** (0.424)
Observations	114
Log Likelihood	−99.065
Akaike Inf. Crit.	210.130
Bayesian Inf. Crit.	226.547
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001

Table C.56: This table presents the results from the linear mixed-effects models run in Chapter 7 Section 7.4 examining the correlation of the total time a participant received nonverbal backchannels (sec) on their psychological safety score [Edmondson, 1999] and their perceived group inclusion score [Jansen et al., 2014]. Controls used in these models include intergroup bias, robot liaison designation, gender, emotional intelligence, and familiarity with other team member. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	psychological safety	perceived inclusion
nonverbal backchannels received (sec)	0.002 (0.002)	0.002 (0.002)
gender: female (1) or male (0)	0.290* (0.132)	
ingroup (0) or outgroup (1)		0.181 (0.114)
robot liaison: yes (1) or no (0)		−0.293** (0.112)
emotional intelligence	0.325** (0.104)	0.281*** (0.081)
max. familiarity		0.053 (0.033)
Constant	4.121*** (0.548)	2.752*** (0.422)
Observations	114	114
Log Likelihood	−128.289	−101.611
Akaike Inf. Crit.	268.579	219.222
Bayesian Inf. Crit.	284.996	241.111
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001	



Table C.57: This table presents the results from the linear mixed-effects models run in Chapter 7 Section 7.4 examining the correlation of the total time a participant spent non-verbally backchanneling others (sec) on their psychological safety score [Edmondson, 1999] and perceived inclusion score [Jansen et al., 2014]. Controls used for these models include intergroup bias, robot liaison designation, gender, emotional intelligence, and familiarity with other participants. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	psychological safety	perceived inclusion
	(1)	(2)
total time spent nonverbally backchanneling others (sec)	−0.005* (0.002)	0.001 (0.002)
gender: female (1) or male (0)	0.318* (0.130)	
ingroup (0) or outgroup (1)		0.201 . (0.112)
robot liaison: yes (1) or no (0)		−0.295** (0.112)
emotional intelligence	0.354*** (0.103)	0.278*** (0.082)
max. familiarity		0.051 (0.033)
Constant	4.158*** (0.538)	2.776*** (0.421)
Observations	114	114
Log Likelihood	−126.542	−101.915
Akaike Inf. Crit.	265.083	219.830
Bayesian Inf. Crit.	281.501	241.720
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.58: This table presents the results from the linear mixed-effects models run in Chapter 7 Section 7.4 examining the correlation of the total time a participant spent verbally backchanneling others (sec) on their psychological safety score [Edmondson, 1999] and perceived inclusion score [Jansen et al., 2014]. Controls used for these models include intergroup bias, robot liaison designation, gender, and emotional intelligence. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	psychological safety	perceived inclusion
	(1)	(2)
total time spent verbally backchanneling others (sec)	−0.003 (0.007)	0.010 . (0.005)
gender: female (1) or male (0)	0.310* (0.139)	
ingroup (0) or outgroup (1)		0.217 . (0.112)
robot liaison: yes (1) or no (0)		−0.284* (0.111)
emotional intelligence	0.339** (0.106)	0.266** (0.082)
Constant	4.153*** (0.548)	2.759*** (0.418)
Observations	114	114
Log Likelihood	−127.392	−97.988
Akaike Inf. Crit.	266.784	209.976
Bayesian Inf. Crit.	283.201	229.129

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

Table C.59: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of the nonverbal and verbal backchannels, separately, a participant received (sec) normalized by the total time that participant spent talking (sec), with controls for ingroup-outgroup bias, robot liaison designation, gender, emotional intelligence, and the maximum familiarity they have between their two human team members, on their perceived inclusion score [Jansen et al., 2014]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	perceived inclusion	
nonverbal backchannels received (sec) normalized by talking time (sec)	−0.732* (0.298)	
verbal backchannels received (sec) normalized by talking time (sec)		−1.671* (0.802)
ingroup (0) or outgroup (1)	0.272* (0.118)	0.286* (0.120)
robot liaison: yes (1) or no (0)	−0.340** (0.115)	−0.356** (0.116)
gender: female (1) or male (0)	0.151 (0.105)	
emotional intelligence	0.201* (0.084)	0.228** (0.084)
max. familiarity	0.058. (0.034)	0.068* (0.034)
Constant	3.177*** (0.436)	3.133*** (0.439)
Observations	106	106
Log Likelihood	−89.029	−88.151
Akaike Inf. Crit.	196.058	192.302
Bayesian Inf. Crit.	220.029	213.609

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

Table C.60: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the correlation of participants total time talking in round 2 of the experiment (sec), with controls for ingroup-outgroup bias, robot liaison designation, emotional intelligence, and the maximum familiarity a participant had between their two fellow human participants, on their perceived inclusion [Jansen et al., 2014]. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	perceived inclusion
total talking time (sec)	0.001 (0.0003)
ingroup (0) or outgroup (1)	0.220 . (0.116)
robot liaison: yes (1) or no (0)	−0.332** (0.116)
emotional intelligence	0.224** (0.085)
max. familiarity	0.057 (0.036)
Constant	2.908*** (0.438)
Observations	106
Log Likelihood	−96.862
Akaike Inf. Crit.	209.725
Bayesian Inf. Crit.	231.032
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.61: This table presents the results from the ANOVA analysis examining the influence of the time participants in a group spent verbally backchanneling one another (sec) on both groups' average perceived inclusion and average psychological safety scores in Chapter 7 Section 7.4. This analysis was performed using the 'aov' function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>	
	perceived inclusion	psychological safety
verbal backchannels produced by group (sec)	F(1) = 9.720** (0.110)	F(1) = 6.171* (0.038)
robot liaison: ingroup (0) or outgroup (1)	F(1) = 0.000 (0.0227)	F(1) = 2.522 (0.203)
verbally supportive robot: yes (1) or no (0)	F(1) = 0.096 (0.0004)	F(1) = 2.167 (0.089)
number of females	F(1) = 1.310 (0.096)	F(1) = 1.848 (0.041)
avg. age	F(1) = 4.461* (0.052)	F(1) = 8.689** (0.063)
avg. familiarity	F(1) = 2.061 (0.013)	F(1) = 0.012 (0.033)
avg. extraversion	F(1) = 4.979* (0.016)	F(1) = 0.082 (0.119)
avg. emotional intelligence	F(1) = 7.769** (0.211)	F(1) = 14.023*** (0.326)
Observations	38	38
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001		

Table C.62: This table presents the results from the ANOVA analysis examining the influence of the time participants in a group spent talking (sec) on both groups' average perceived inclusion and average psychological safety scores in Chapter 7 Section 7.4. This analysis was performed using the 'aov' function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>	
	perceived inclusion	psychological safety
total time talking by all group members (sec)	F(1) = 7.319* (0.052)	F(1) = 0.446 (0.002)
robot liaison: ingroup (0) or outgroup (1)	F(1) = 0.052 (0.008)	F(1) = 2.252 (0.268)
verbally supportive robot: yes (1) or no (0)	F(1) = 0.202 (0.006)	F(1) = 1.266 (0.119)
number of females	F(1) = 1.801 (0.001)	F(1) = 16.001*** (0.238)
avg. age	F(1) = 9.910** (0.130)	F(1) = 19.146*** (0.154)
avg. familiarity	F(1) = 1.500 (0.005)	F(1) = 0.327 (0.051)
avg. extraversion	F(1) = 4.490* (0.037)	F(1) = 0.288 (0.101)
avg. emotional intelligence	F(1) = 2.608 (0.088)	F(1) = 6.477* (0.193)
Observations	38	38
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.63: This table presents the results from the ANOVA analysis examining the influence of the proportion of time participants in a group spent verbally backchanneling one another (sec), relative to the group's total talking time, on both groups' average perceived inclusion and average psychological safety scores in Chapter 7 Section 7.4. This analysis was performed using the 'aov' function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>	
	perceived inclusion	psychological safety
verbal backchannels produced (sec) normalized by talking time (sec)	F(1) = 0.140 (0.001)	F(1) = 0.818 (0.028)
robot liaison: ingroup (0) or outgroup (1)	F(1) = 0.086 (0.005)	F(1) = 2.783 (0.247)
verbally supportive robot: yes (1) or no (0)	F(1) = 0.089 (0.001)	F(1) = 0.682 (0.133)
number of females	F(1) = 1.954 (0.001)	F(1) = 15.692*** (0.255)
avg. age	F(1) = 11.919** (0.130)	F(1) = 21.646*** (0.177)
avg. familiarity	F(1) = 3.638 . (0.019)	F(1) = 0.168 (0.049)
avg. extraversion	F(1) = 5.087* (0.052)	F(1) = 0.332 (0.099)
avg. emotional intelligence	F(1) = 2.187 (0.075)	F(1) = 6.041* (0.183)
Observations	38	38
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001		

Table C.64: This table presents the results from the ANOVA analysis examining the influence of the time participants in a group spent nonverbally backchanneling one another (sec) on both groups' average perceived inclusion and average psychological safety scores in Chapter 7 Section 7.4. This analysis was performed using the 'aov' function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>	
	perceived inclusion	psychological safety
nonverbal backchannels produced by group (sec)	F(1) = 3.477 . (0.085)	F(1) = 0.057 (0.005)
robot liaison: ingroup (0) or outgroup (1)	F(1) = 0.457 (0.004)	F(1) = 0.786 (0.168)
verbally supportive robot: yes (1) or no (0)	F(1) = 0.016 (0.002)	F(1) = 0.855 (0.066)
number of females	F(1) = 0.282 (0.045)	F(1) = 8.141** (0.109)
avg. age	F(1) = 7.710** (0.080)	F(1) = 10.281** (0.088)
avg. familiarity	F(1) = 4.328* (0.034)	F(1) = 0.051 (0.031)
avg. extraversion	F(1) = 5.269* (0.022)	F(1) = 0.005 (0.094)
avg. emotional intelligence	F(1) = 7.222* (0.199)	F(1) = 13.206** (0.313)
Observations	38	38
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001	



Table C.65: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of intergroup bias (ingroup or outgroup), controlling for the familiarity with other human participants, on the amount of verbal and nonverbal backchannels (sec) each participant received. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	verbal backchannels	nonverbal backchannels
	received (sec)	received (sec)
total talking time (sec)	0.027*** (0.006)	0.057*** (0.013)
ingroup (0) or outgroup (1)	6.672*** (1.908)	8.847* (4.400)
max. familiarity	1.189 . (0.639)	
Constant	4.387 . (2.276)	5.018 (5.329)
Observations	106	106
Log Likelihood	−388.638	−478.448
Akaike Inf. Crit.	789.275	966.896
Bayesian Inf. Crit.	805.256	980.214
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.66: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of intergroup bias (ingroup or outgroup), controlling for the familiarity with other human participants, on the proportion of verbal and nonverbal backchannels (sec) each participant received relative to their total talking time (sec). Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	verbal backchannels	nonverbal backchannels
	received (sec)	received (sec)
	normalized by talking time (sec)	normalized by talking time (sec)
total talking time (sec)	−0.0002*** (0.00004)	−0.0003** (0.0001)
ingroup (0) or outgroup (1)	0.036** (0.012)	0.083* (0.035)
Constant	0.107*** (0.014)	0.167*** (0.039)
Observations	106	106
Log Likelihood	131.075	22.885
Akaike Inf. Crit.	−252.150	−35.769
Bayesian Inf. Crit.	−238.833	−22.452

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

Table C.67: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of gender on the amount of verbal backchannels each participant received (sec) and the proportion of verbal and backchannels (sec) each participant received relative to their total talking time (sec). We controlled for the total talking time (sec), extraversion, and intergroup bias of participants. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	verbal backchannels produced (sec)	verbal backchannels produced (sec) normalized by talking time (sec)
total talking time (sec)	0.015** (0.005)	−0.0005*** (0.0001)
ingroup (0) or outgroup (1)	−2.743 (1.812)	
gender: female (1) or male (0)	7.131*** (1.695)	0.059* (0.028)
extraversion	0.773 . (0.440)	
Constant	8.622*** (2.325)	0.215*** (0.034)
Observations	106	106
Log Likelihood	−377.266	40.456
Akaike Inf. Crit.	768.532	−70.912
Bayesian Inf. Crit.	787.177	−57.595
<i>Note:</i>	. p<0.1; *p<0.5; **p<0.01; ***p<0.001	

Table C.68: This table presents the results from the ANOVA analysis examining the influence of gender on the amount of verbal backchannels produced by each group (sec) as well as the proportion of verbal backchannels produced by each group (sec) with respect to their total talking time in Chapter 7 Section 7.4. This analysis was performed using the ‘aov’ function in R. The F-value, degrees of freedom, and effect size (partial eta squared) are reported in parentheses for each fixed factor.

	<i>Dependent variable:</i>	
	verbal backchannels produced (sec)	verbal backchannels produced (sec) normalized by talking time (sec)
total talking time (sec)	F(1) = 3.197 . (0.004)	F(1) = 18.502*** (0.443)
robot liaison: ingroup (0) or outgroup (1)	F(1) = 3.357 . (0.103)	F(1) = 2.607 (0.073)
verbally supportive robot: yes (1) or no (0)	F(1) = 0.912 (0.004)	F(1) = 0.501 (0.0002)
number of females	F(1) = 18.613*** (0.318)	F(1) = 9.974 ** (0.198)
avg. age	F(1) = 2.277 (0.093)	F(1) = 3.332 . (0.124)
avg. familiarity	F(1) = 0.802 (0.003)	F(1) = 0.503 (0.002)
avg. extraversion	F(1) = 0.564 (0.045)	F(1) = 0.233 (0.031)
avg. emotional intelligence	F(1) = 0.912 (0.033)	F(1) = 1.059 (0.038)
Observations	38	38
<i>Note:</i> . p<0.1; *p<0.5; **p<0.01; ***p<0.001		

Table C.69: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of having a verbally supportive robot on the amount of verbal backchannels (sec) each participant received. The model’s fixed factors included whether the robot gave verbal support, the intergroup bias (ingroup, outgroup), robot liaison designation, and relevant interactions. We controlled for participants’ extraversion and familiarity with other team members. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R ‘lmer’ function from the ‘lme4’ package to perform this analysis and the ‘emmeans’ function with a Tukey adjustment from the ‘emmeans’ package to perform post-hoc tests. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>
	verbal backchannels received (sec)
intergroup bias: ingroup (0) or outgroup (1)	16.850*** (4.088)
robot liaison: yes (1) or no (0)	3.625 (2.626)
verbally supportive robot: yes (1) or no (0)	0.566 (2.567)
intergroup bias * robot liaison	−11.120** (4.144)
intergroup bias * verbally supportive robot	−8.832* (4.329)
extraversion	1.350** (0.487)
max. familiarity	0.971 (0.650)
Constant	7.292* (2.913)
Observations	114
Log Likelihood	−405.563
Akaike Inf. Crit.	831.126
Bayesian Inf. Crit.	858.488

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

Table C.70: This table presents the results from the linear mixed-effects model run in Chapter 7 Section 7.4 examining the influence of having a verbally supportive robot on participants' psychological safety and perceived inclusion scores. The model's fixed factors included whether the robot gave verbal support, the intergroup bias (ingroup, outgroup), robot liaison designation, and relevant interactions. We controlled for participants' extraversion and familiarity with other team members. Each participant is grouped with their two fellow human participants in the model where each group has a random intercept. We used the R 'lmer' function from the 'lme4' package to perform this analysis. The linear coefficient (odds ratio) and standard error are reported in the following table.

	<i>Dependent variable:</i>	
	psychological safety	perceived inclusion
intergroup bias: ingroup (0) or outgroup (1)	0.651* (0.268)	0.505* (0.208)
robot liaison: yes (1) or no (0)	-0.207 (0.143)	-0.286** (0.111)
verbally supportive robot: yes (1) or no (0)	0.347 . (0.180)	0.124 (0.139)
gender: female (1) or male (0)	0.299* (0.130)	
emotional intelligence	0.325** (0.104)	0.285*** (0.081)
max. familiarity		0.050 (0.033)
intergroup bias * verbally supportive robot	-0.603 . (0.310)	-0.418 . (0.240)
Constant	3.907*** (0.552)	2.678*** (0.429)
Observations	114	114
Log Likelihood	-122.753	-96.858
Akaike Inf. Crit.	263.505	211.717
Bayesian Inf. Crit.	288.131	236.342

*Note:* . p<0.1; \*p<0.5; \*\*p<0.01; \*\*\*p<0.001

# Bibliography

- [Alemi et al., 2016] Alemi, M., Ghanbarzadeh, A., Meghdari, A., and Moghadam, L. J. (2016). Clinical application of a humanoid robot in pediatric cancer interventions. *International Journal of Social Robotics*, 8(5):743–759.
- [Alemi et al., 2015] Alemi, M., Meghdari, A., and Ghazisaedy, M. (2015). The impact of social robotics on l2 learners’ anxiety and attitude in english vocabulary acquisition. *International Journal of Social Robotics*, 7(4):523–535.
- [Althaus et al., 2004] Althaus, P., Ishiguro, H., Kanda, T., Miyashita, T., and Christensen, H. (2004). Navigation for human-robot interaction tasks. In *IEEE International Conference on Robotics and Automation, ICRA '04*, pages 1894–1900, New Orleans, LA, USA. IEEE.
- [Andrist et al., 2013] Andrist, S., Spannan, E., and Mutlu, B. (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*, pages 341–348. IEEE Press.
- [Asch, 1956] Asch, S. E. (1956). Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9):1.
- [Ball et al., 2017] Ball, A. K., Rye, D. C., Silvera-Tawil, D., and Velonaki, M. (2017). How should a robot approach two people? *Journal of Human-Robot Interaction*, 6(3):71–91.

- [Baron and Dunham, 2015] Baron, A. S. and Dunham, Y. (2015). Representing ‘us’ and ‘them’: Building blocks of intergroup cognition. *Journal of Cognition and Development*, 16(5):780–801.
- [Barsade, 2002] Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 47(4):644–675.
- [Beane and J. Orlikowski, 2015] Beane, M. and J. Orlikowski, W. (2015). What difference does a robot make? the material enactment of distributed coordination. *Organization Science*, 26(6).
- [Belpaeme et al., 2018] Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: A review. *Science robotics*, 3(21):eaat5954.
- [Bettenhausen and Murnighan, 1985] Bettenhausen, K. and Murnighan, J. K. (1985). The emergence of norms in competitive decision-making groups. *Administrative science quarterly*, pages 350–372.
- [Bohus et al., 2014] Bohus, D., Saw, C. W., and Horvitz, E. (2014). Directions robot: In-the-wild experiences and lessons learned. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS ’14, pages 637–644, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.
- [Booth et al., 2017] Booth, S., Tompkin, J., Pfister, H., Waldo, J., Gajos, K., and Nagpal, R. (2017). Piggybacking robots: Human-robot overtrust in university dormitory security. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI ’17, pages 426–434, Vienna, Austria. ACM.
- [Bottom et al., 2002] Bottom, W. P., Gibson, K., Daniels, S. E., and Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organization Science*, 13(5):497–513.



- [Brandon and Hollingshead, 2004] Brandon, D. P. and Hollingshead, A. B. (2004). Trans-active memory systems in organizations: Matching tasks, expertise, and people. *Organization science*, 15(6):633–644.
- [Brooks et al., 2016] Brooks, D. J., Begum, M., and Yanco, H. A. (2016). Analysis of reactions towards failures and recovery strategies for autonomous robots. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 487–492. IEEE.
- [Brscić et al., 2015] Brscić, D., Kidokoro, H., Suehiro, Y., and Kanda, T. (2015). Escaping from children’s abuse of social robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI ’15*, pages 59–66, New York, NY, USA. ACM.
- [Bukowski et al., 1994] Bukowski, W. M., Hoza, B., and Boivin, M. (1994). Measuring friendship quality during pre-and early adolescence: The development and psychometric properties of the friendship qualities scale. *Journal of social and Personal Relationships*, 11(3):471–484.
- [Butler Jr and Cantrell, 1984] Butler Jr, J. K. and Cantrell, R. S. (1984). A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. *Psychological reports*, 55(1):19–28.
- [Carmeli et al., 2009] Carmeli, A., Brueller, D., and Dutton, J. E. (2009). Learning behaviours in the workplace: The role of high-quality interpersonal relationships and psychological safety. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research*, 26(1):81–98.
- [Carpinella et al., 2017] Carpinella, C. M., Wyman, A. B., Perez, M. A., and Stroessner, S. J. (2017). The robotic social attributes scale (rosas): Development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 254–262. ACM.

- [Chandra et al., 2015] Chandra, S., Alves-Oliveira, P., Lemaignan, S., Sequeira, P., Paiva, A., and Dillenbourg, P. (2015). Can a child feel responsible for another in the presence of a robot in a collaborative learning activity? In *24th IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN '16, pages 167–172, Kobe, Japan. IEEE.
- [Chang et al., 2012] Chang, W.-L., White, J. P., Park, J., Holm, A., and Šabanović, S. (2012). The effect of group size on people’s attitudes and cooperative behaviors toward robots in interactive gameplay. In *The 21st IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN '12, pages 845–850. IEEE Press.
- [Chen et al., 2018] Chen, M., Nikolaidis, S., Soh, H., Hsu, D., and Srinivasa, S. (2018). Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 307–315. ACM.
- [Chidambaram et al., 2012] Chidambaram, V., Chiang, Y.-H., and Mutlu, B. (2012). Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 293–300. ACM.
- [Cho and Mor Barak, 2008] Cho, S. and Mor Barak, M. E. (2008). Understanding of diversity and inclusion in a perceived homogeneous culture: A study of organizational commitment and job performance among korean employees. *Administration in Social Work*, 32(4):100–126.
- [Cialdini et al., 1990] Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, 58(6):1015.
- [Clabaugh and Matarić, 2019] Clabaugh, C. and Matarić, M. (2019). Escaping oz: Autonomy in socially assistive robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):33–61.

- [Claire et al., 2020] Claire, H., Chen, Y., Modi, J., Jung, M., and Nikolaidis, S. (2020). Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 299–308.
- [Coan and Gottman, 2007] Coan, J. A. and Gottman, J. M. (2007). The specific affect coding system (spaff). *Handbook of emotion elicitation and assessment*, pages 267–285.
- [Cohen and Bailey, 1997] Cohen, S. G. and Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management*, 23(3):239–290.
- [Cooper and Petrides, 2010] Cooper, A. and Petrides, K. (2010). A psychometric analysis of the trait emotional intelligence questionnaire–short form (teique–sf) using item response theory. *Journal of personality assessment*, 92(5):449–457.
- [Cooper et al., 2013] Cooper, M. A., Ibrahim, A., Lyu, H., and Makary, M. (2013). Underreporting of robotic surgery complications. *Journal for healthcare quality : official publication of the National Association for Healthcare Quality*, 37:133–138.
- [Correia et al., 2016] Correia, F., Alves-Oliveira, P., Maia, N., Ribeiro, T., Petisca, S., Melo, F., and Paiva, A. (2016). Just follow the suit! trust in human-robot interactions during card game playing. In *25th IEEE International Symposium on Robot and Human Interactive Communication*, pages 507–512, New York, NY, USA. IEEE.
- [Correia et al., 2017a] Correia, F., Alves-Oliveira, P., Ribeiro, T., Melo, F., and Paiva, A. (2017a). A social robot as a card game player. In *The Thirteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE ’17*, pages 23–29, Little Cottonwood Canyon, UT, USA. AAAI.
- [Correia et al., 2018a] Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., and Paiva, A. (2018a). Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*,

- pages 507–513. International Foundation for Autonomous Agents and Multiagent Systems.
- [Correia et al., 2018b] Correia, F., Mascarenhas, S., Prada, R., Melo, F. S., and Paiva, A. (2018b). Group-based emotions in teams of humans and robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18*, pages 261–269, New York, NY, USA. ACM.
- [Correia et al., 2017b] Correia, F., Petisca, S., Alves-Oliveira, P., Ribeiro, T., Melo, F. S., and Paiva, A. (2017b). Groups of humans and robots: Understanding membership preferences and team formation. In *Robotics: Science and Systems, RSS '17*, pages 1–10, Cambridge, MA, USA. RSS.
- [Cozby, 1973] Cozby, P. C. (1973). Self-disclosure: a literature review. *Psychological bulletin*, 79(2):73.
- [Craig and Kelly, 1999] Craig, T. Y. and Kelly, J. R. (1999). Group cohesiveness and creative performance. *Group dynamics: Theory, research, and practice*, 3(4):243.
- [Curhan et al., 2006] Curhan, J. R., Elfenbein, H. A., and Xu, H. (2006). What do people value when they negotiate? mapping the domain of subjective value in negotiation. *Journal of personality and social psychology*, 91(3):493.
- [de Kok and Heylen, 2011] de Kok, I. and Heylen, D. (2011). The multilis corpus—dealing with individual differences in nonverbal listening behavior. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, pages 362–375. Springer.
- [de Kok et al., 2013] de Kok, I., Heylen, D., and Morency, L.-P. (2013). Speaker-adaptive multimodal prediction model for listener responses. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 51–58.
- [Desai et al., 2013] Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th*

- ACM/IEEE international conference on Human-robot interaction*, pages 251–258. IEEE Press.
- [DeSteno et al., 2012] DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., and Lee, J. J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological science*, 23(12):1549–1556.
- [Dirks et al., 2011] Dirks, K. T., Kim, P. H., Ferrin, D. L., and Cooper, C. D. (2011). Understanding the effects of substantive responses on trust following a transgression. *Organizational Behavior and Human Decision Processes*, 114(2):87–103.
- [Dixon and Foster, 1998] Dixon, J. A. and Foster, D. H. (1998). Gender, social context, and backchannel responses. *The Journal of social psychology*, 138(1):134–136.
- [Duncan, 1974] Duncan, S. (1974). On the structure of speaker–auditor interaction during speaking turns. *Language in society*, 3(2):161–180.
- [Duncan and Fiske, 2015] Duncan, S. and Fiske, D. W. (2015). *Face-to-face interaction: Research, methods, and theory*. Routledge.
- [Dunham et al., 2011] Dunham, Y., Baron, A. S., and Carey, S. (2011). Consequences of “minimal” group affiliations in children. *Child development*, 82(3):793–811.
- [Duysburgh et al., 2014] Duysburgh, P., Elprama, S. A., and Jacobs, A. (2014). Exploring the social-technological gap in telesurgery: Collaboration within distributed or teams. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW ’14, pages 1537–1548, New York, NY, USA. ACM.
- [Eckerman and Peterman, 2001] Eckerman, C. O. and Peterman, K. (2001). Peers and infant social/communicative development. *Blackwell handbook of infant development*, pages 326—350.
- [Edmondson, 1999] Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative science quarterly*, 44(2):350–383.

- [Edmondson, 2014] Edmondson, A. (2014). Building a psychologically safe workplace. <https://www.youtube.com/watch?v=LhoLuui9gX8>.
- [Edmondson et al., 2004] Edmondson, A. C., Kramer, R. M., and Cook, K. S. (2004). Psychological safety, trust, and learning in organizations: A group-level lens. *Trust and distrust in organizations: Dilemmas and approaches*, 12:239–272.
- [Erickson and Dyer, 2004] Erickson, J. and Dyer, L. (2004). Right from the start: Exploring the effects of early team events on subsequent project team development and performance. *Administrative Science Quarterly*, 49(3):438–471.
- [Fan et al., 2016] Fan, J., Beuscher, L., Newhouse, P. A., Mion, L. C., and Sarkar, N. (2016). A robotic coach architecture for multi-user human-robot interaction (ramu) with the elderly and cognitively impaired. In *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN '16*, pages 445–450.
- [Faria et al., 2017] Faria, M., Silva, R., Alves-Oliveira, P., Melo, F., and Paiva, A. (2017). ”me and you together” movement impact in multi-user collaboration tasks. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2793–2798. IEEE.
- [Feil-Seifer and Mataric, 2005] Feil-Seifer, D. and Mataric, M. J. (2005). Defining socially assistive robotics. In *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, pages 465–468. IEEE.
- [Feldman, 1984] Feldman, D. C. (1984). The development and enforcement of group norms. *Academy of management review*, 9(1):47–53.
- [Fernández-Llamas et al., 2017] Fernández-Llamas, C., Conde, M. Á., Rodríguez-Sedano, F. J., Rodríguez-Lera, F. J., and Matellán-Olivera, V. (2017). Analysing the computational competences acquired by k-12 students when lectured by robotic and human teachers. *International Journal of Social Robotics*.
- [Ferrin et al., 2007] Ferrin, D. L., Kim, P. H., Cooper, C. D., and Dirks, K. T. (2007). Silence speaks volumes: the effectiveness of reticence in comparison to apology and denial

- for responding to integrity-and competence-based trust violations. *Journal of Applied Psychology*, 92(4):893.
- [Forlizzi, 2007] Forlizzi, J. (2007). How robotic products become social products: An ethnographic study of cleaning in the home. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, HRI '07, pages 129–136, New York, NY, USA. ACM.
- [Forlizzi and DiSalvo, 2006] Forlizzi, J. and DiSalvo, C. (2006). Service robots in the domestic environment: A study of the roomba vacuum in the home. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI '06, pages 258–265, New York, NY, USA. ACM.
- [Francis et al., 1992] Francis, L. J., Brown, L. B., and Philipchalk, R. (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):443–449.
- [Fraune et al., 2019] Fraune, M. R., Sherrin, S., Šabanović, S., and Smith, E. R. (2019). Is human-robot interaction more competitive between groups than between individuals? In *14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, pages 104–113, Daegu, South Korea. IEEE Press.
- [Freedy et al., 2007] Freedy, A., DeVisser, E., Weltman, G., and Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on*, pages 106–114. IEEE.
- [Fussell et al., 2008] Fussell, S. R., Kiesler, S., Setlock, L. D., and Yew, V. (2008). How people anthropomorphize robots. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, HRI '08, pages 145–152, New York, NY, USA. ACM.
- [Geiskkovitch et al., 2019] Geiskkovitch, D. Y., Thiessen, R., Young, J. E., and Glenwright, M. R. (2019). What? that’s not a chair!: How robot informational errors affect children’s

- trust towards robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 48–56. IEEE.
- [Gersick, 1988] Gersick, C. J. (1988). Time and transition in work teams: Toward a new model of group development. *Academy of Management journal*, 31(1):9–41.
- [Gibson and Vermeulen, 2003] Gibson, C. and Vermeulen, F. (2003). A healthy divide: Subgroups as a stimulus for team learning behavior. *Administrative science quarterly*, 48(2):202–239.
- [Gockley et al., 2006] Gockley, R., Forlizzi, J., and Simmons, R. (2006). Interactions with a moody robot. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI ’06, pages 186–193, New York, NY, USA. ACM.
- [Gombolay et al., 2015] Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F., and Shah, J. A. (2015). Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots*, 39(3):293–312.
- [Goodrich and Schultz, 2007] Goodrich, M. A. and Schultz, A. C. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275.
- [Goodwin, 1986] Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human studies*, 9(2-3):205–217.
- [Gordon et al., 2016] Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., and Breazeal, C. (2016). Affective personalization of a social robot tutor for children’s second language skills. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- [Gravano and Hirschberg, 2009] Gravano, A. and Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. In *Tenth Annual Conference of the International Speech Communication Association*.
- [Groom and Nass, 2007] Groom, V. and Nass, C. (2007). Can robots be teammates? benchmarks in human-robot teams. *Interaction Studies*, 8:483–500.



- [Hancock et al., 2011] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(5):517–527.
- [Hastings et al., 2018] Hastings, E. M., Jahanbakhsh, F., Karahalios, K., Marinov, D., and Bailey, B. P. (2018). Structure or nurture? the effects of team-building activities and team composition on team outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21.
- [Hebesberger et al., 2016] Hebesberger, D., Koertner, T., Gisinger, C., Pripfl, J., and Dondrup, C. (2016). Lessons learned from the deployment of a long-term autonomous robot as companion in physical therapy for older adults with dementia a mixed methods study. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 27–34.
- [Hertel and Hüffmeier, 2011] Hertel, G. and Hüffmeier, J. (2011). Many cheers make light the work: How social support triggers process gains in teams. *Journal of Managerial Psychology*.
- [Hess and Johnston, 1988] Hess, L. J. and Johnston, J. R. (1988). Acquisition of back channel listener responses to adequate messages. *Discourse Processes*, 11(3):319–335.
- [Hinds et al., 2004] Hinds, P. J., Roberts, T. L., and Jones, H. (2004). Whose job is it anyway? a study of human-robot interaction in a collaborative task. *Hum.-Comput. Interact.*, 19(1):151–181.
- [Hjalmarsson and Oertel, 2012] Hjalmarsson, A. and Oertel, C. (2012). Gaze direction as a back-channel inviting cue in dialogue. In *IVA 2012 workshop on realtime conversational virtual agents*, volume 9. Citeseer.
- [Hoffman et al., 2015] Hoffman, G., Zuckerman, O., Hirschberger, G., Luria, M., and Shani Sherman, T. (2015). Design and evaluation of a peripheral robotic conversation

- companion. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 3–10, New York, NY, USA. ACM.
- [Hogg and Terry, 2000] Hogg, M. A. and Terry, D. I. (2000). Social identity and self-categorization processes in organizational contexts. *Academy of management review*, 25(1):121–140.
- [Honig and Oron-Gilad, 2018] Honig, S. S. and Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, 9:861.
- [Hood et al., 2015] Hood, D., Lemaignan, S., and Dillenbourg, P. (2015). When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 83–90, New York, NY, USA. ACM.
- [Imai et al., 2002] Imai, M., Kanda, T., Ono, T., Ishiguro, H., and Mase, K. (2002). Robot mediated round table: Analysis of the effect of robot’s gaze. In *Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication*, RO-MAN '02, pages 411–416, Berlin, Germany. IEEE.
- [Jansen et al., 2014] Jansen, W. S., Otten, S., van der Zee, K. I., and Jans, L. (2014). Inclusion: Conceptualization and measurement. *European journal of social psychology*, 44(4):370–385.
- [Jibo, 2017] Jibo (2017). Jibo inc. <https://jibo.com>.
- [Jones and George, 1998] Jones, G. R. and George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review*, 23(3):531–546.
- [Joshi and Šabanović, 2019] Joshi, S. and Šabanović, S. (2019). Robots for inter-generational interactions: Implications for nonfamilial community settings. In *14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, pages 478–486, Daegu, South Korea. IEEE Press.

- [Jung et al., 2012] Jung, M., Chong, J., and Leifer, L. (2012). Group hedonic balance and pair programming performance: affective interaction dynamics as indicators of performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 829–838.
- [Jung and Hinds, 2018] Jung, M. and Hinds, P. (2018). Robots in the wild: A time for more robust theories of human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 7(1):2.
- [Jung, 2017] Jung, M. F. (2017). Affective grounding in human-robot interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17*, pages 263–273, New York, NY, USA. ACM.
- [Jung et al., 2013] Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., and Breazeal, C. (2013). Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1555–1566.
- [Jung et al., 2015] Jung, M. F., Martelaro, N., and Hinds, P. J. (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, HRI '15*, pages 229–236, New York, NY, USA. ACM.
- [Jung et al., 2017] Jung, M. F., Šabanović, S., Eyssel, F., and Fraune, M. (2017). Robots in groups and teams. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 401–407. ACM.
- [Jurafsky et al., 1997] Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and Van Ess-Dykema, C. (1997). Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95. IEEE.

- [Kanda et al., 2004] Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19(1-2):61–84.
- [Kanda et al., 2007] Kanda, T., Sato, R., Saiwaki, N., and Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human-robot interaction. *IEEE Transactions on robotics*, 23(5):962–971.
- [Kanda et al., 2012] Kanda, T., Shimada, M., and Koizumi, S. (2012). Children learning with a social robot. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '12*, pages 351–358, New York, NY, USA. ACM.
- [Kaniarasu and Steinfeld, 2014] Kaniarasu, P. and Steinfeld, A. M. (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*, pages 850–855. IEEE.
- [Karau and Kelly, 1992] Karau, S. J. and Kelly, J. R. (1992). The effects of time scarcity and time abundance on group performance quality and interaction process. *Journal of experimental social psychology*, 28(6):542–571.
- [Kelly and McGrath, 1985] Kelly, J. R. and McGrath, J. E. (1985). Effects of time limits and task types on task performance and interaction of four-person groups. *Journal of personality and social psychology*, 49(2):395.
- [Kidd and Breazeal, 2004] Kidd, C. D. and Breazeal, C. (2004). Effect of a robot on user perceptions. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 4, pages 3559–3564. IEEE.
- [Kidokoro et al., 2013] Kidokoro, H., Kanda, T., Brščić, D., and Shiomi, M. (2013). Will i bother here?: A robot anticipating its influence on pedestrian walking comfort. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction, HRI '13*, pages 259–266, Piscataway, NJ, USA. IEEE Press.

- [Kiffin-Petersen and Cordery, 2003] Kiffin-Petersen, S. A. and Cordery, J. L. (2003). Trust, individualism and job characteristics as predictors of employee preference for teamwork. *International Journal of Human Resource Management*, 14(1):93–116.
- [Kim et al., 2013] Kim, E. S., Berkovits, L. D., Bernier, E. P., Leyzberg, D., Shic, F., Paul, R., and Scassellati, B. (2013). Social robots as embedded reinforcers of social behavior in children with autism. *Journal of Autism and Developmental Disorders*, 43(5):1038–1049.
- [Kim et al., 2009] Kim, P. H., Dirks, K. T., and Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review*, 34(3):401–422.
- [Kim et al., 2006] Kim, P. H., Dirks, K. T., Cooper, C. D., and Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, 99(1):49–65.
- [Kim et al., 2004] Kim, P. H., Ferrin, D. L., Cooper, C. D., and Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of applied psychology*, 89(1):104.
- [Kirchner et al., 2011] Kirchner, N., Alempijevic, A., and Dissanayake, G. (2011). Nonverbal robot-group interaction using an imitated gaze cue. In *Proceedings of the 6th International Conference on Human-robot Interaction, HRI '11*, pages 497–504, New York, NY, USA. ACM.
- [Klimoski and Karol, 1976] Klimoski, R. J. and Karol, B. L. (1976). The impact of trust on creative problem solving groups. *Journal of Applied Psychology*, 61(5):630–633.
- [Knight, 2015] Knight, A. P. (2015). Mood at the midpoint: Affect and change in exploratory search over time in teams that face a deadline. *Organization Science*, 26(1):99–118.

- [Kondo et al., 2013] Kondo, Y., Takemura, K., Takamatsu, J., and Ogasawara, T. (2013). A gesture-centric android system for multi-party human-robot interaction. *Journal of Human-Robot Interaction*, 2(1):133–151.
- [Kostopoulos and Bozionelos, 2011] Kostopoulos, K. C. and Bozionelos, N. (2011). Team exploratory and exploitative learning: Psychological safety, task conflict, and team performance. *Group & Organization Management*, 36(3):385–415.
- [Kozima et al., 2009] Kozima, H., Michalowski, M. P., and Nakagawa, C. (2009). Keep on. *International Journal of Social Robotics*, 1(1):3–18.
- [Kwon et al., 2016] Kwon, M., Jung, M. F., and Knepper, R. A. (2016). Human expectations of social robots. In *The Eleventh ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 463–464. IEEE.
- [Lafferty and Pond, 1974] Lafferty, J. C. and Pond, A. W. (1974). *The desert survival situation: A group decision making experience for examining and increasing individual and team effectiveness*. Human Synergistics.
- [Lala et al., 2017] Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., and Kawahara, T. (2017). Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 127–136.
- [Larzelere and Huston, 1980] Larzelere, R. E. and Huston, T. L. (1980). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family*, pages 595–604.
- [Latané, 1981] Latané, B. (1981). The psychology of social impact. *American psychologist*, 36(4):343.
- [Lau and Murnighan, 1998] Lau, D. C. and Murnighan, J. K. (1998). Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review*, 23(2):325–340.

- [Lee et al., 2017] Lee, J. J., Breazeal, C., and DeSteno, D. (2017). Role of speaker cues in attention inference. *Frontiers in Robotics and AI*, 4:47.
- [Lee et al., 2019] Lee, J. J., Sha, F., and Breazeal, C. (2019). A bayesian theory of mind approach to nonverbal communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 487–496. IEEE.
- [Lee et al., 2006] Lee, K. M., Peng, W., Jin, S.-A., and Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication*, 56(4):754–772.
- [Lee et al., 2010] Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., and Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 203–210. IEEE.
- [Leite et al., 2015a] Leite, I., McCoy, M., Lohani, M., Ullman, D., Salomons, N., Stokes, C., Rivers, S., and Scassellati, B. (2015a). Emotional storytelling in the classroom: Individual versus group interaction between children and robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15*, pages 75–82, New York, NY, USA. ACM.
- [Leite et al., 2017] Leite, I., McCoy, M., Lohani, M., Ullman, D., Salomons, N., Stokes, C., Rivers, S., and Scassellati, B. (2017). Narratives with robots: The impact of interaction context and individual differences on story recall and emotional understanding. *Frontiers in Robotics and AI*, 4:29.
- [Leite et al., 2015b] Leite, I., Mccoy, M., Ullman, D., Salomons, N., and Scassellati, B. (2015b). Comparing models of disengagement in individual and group interactions. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 99–105.
- [Leite et al., 2012] Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., and Paiva, A. (2012). The influence of empathy in human-robot relations. *International Journal of Human-Computer Studies*, 71.

- [Levine and Moreland, 1998] Levine, J. M. and Moreland, R. L. (1998). Small groups. In Gilbert, D. T., Fiske, S. T., and Lindzey, G., editors, *The Handbook of Social Psychology*, pages 415–496. McGraw-Hill, New York, NY, USA.
- [Lewis et al., 2005] Lewis, K., Lange, D., and Gillis, L. (2005). Transactive memory systems, learning, and learning transfer. *Organization Science*, 16(6):581–598.
- [Lickel et al., 2000] Lickel, B., Hamilton, D. L., Wierzchowska, G., Lewis, A., Sherman, S. J., and Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of personality and social psychology*, 78(2):223.
- [Liu et al., 2013] Liu, P., Glas, D. F., Kanda, T., Ishiguro, H., and Hagita, N. (2013). It’s not polite to point: Generating socially-appropriate deictic behaviors towards people. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction, HRI ’13*, pages 267–274, Piscataway, NJ, USA. IEEE Press.
- [Ljungblad et al., 2012] Ljungblad, S., Kotrbova, J., Jacobsson, M., Cramer, H., and Niechwiadowicz, K. (2012). Hospital robot at work: Something alien or an intelligent colleague? In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW ’12*, pages 177–186, New York, NY, USA. ACM.
- [Lynch, 2002] Lynch, O. H. (2002). Humorous communication: Finding a place for humor in communication research. *Communication theory*, 12(4):423–445.
- [Maatman et al., 2005] Maatman, R., Gratch, J., and Marsella, S. (2005). Natural behavior of a listening agent. In *International workshop on intelligent virtual agents*, pages 25–36. Springer.
- [Malisz et al., 2016] Malisz, Z., Włodarczak, M., Buschmeier, H., Skubisz, J., Kopp, S., and Wagner, P. (2016). The alico corpus: analysing the active listener. *Language resources and evaluation*, 50(2):411–442.
- [Martelaro et al., 2016] Martelaro, N., Nneji, V. C., Ju, W., and Hinds, P. (2016). Tell me more: Designing hri to encourage more trust, disclosure, and companionship. In



- The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 181–188. IEEE Press.
- [Matsusaka et al., 2001] Matsusaka, Y., Fujie, S., and Kobayashi, T. (2001). Modeling of conversational strategy for the robot participating in the group conversation. In *Proceedings of the Seventh European Conference on Speech Communication and Technology (EUROSPEECH)*, EUROSPEECH '01.
- [Matsuyama et al., 2015] Matsuyama, Y., Akiba, I., Fujie, S., and Kobayashi, T. (2015). Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech & Language*, pages 1–24.
- [Matsuzoe et al., 2014] Matsuzoe, S., Kuzuoka, H., and Tanaka, F. (2014). Learning english words with the aid of an autonomous care-receiving robot in a children’s group activity. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN '14, pages 802–807, Edinburgh, Scotland, UK. IEEE Press.
- [Mavrogiannis et al., 2019] Mavrogiannis, C., Hutchinson, A. M., Macdonald, J., Alves-Oliveira, P., and Knepper, R. A. (2019). Effects of distinct robot navigation strategies on human behavior in a crowded environment. In *14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, pages 421–430, Daegu, South Korea. IEEE Press.
- [Mayer et al., 1995] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, 20(3):709–734.
- [Michalowski et al., 2006] Michalowski, M., Šabanović, S., and Simmons, R. (2006). A spatial model of engagement for a social robot. In *9th IEEE International Workshop on Advanced Motion Control*, pages 762 – 767.
- [Miller, 1998] Miller, F. A. (1998). Strategic culture change: The door to achieving high performance and inclusion. *Public personnel management*, 27(2):151–160.
- [Miller et al., 1985] Miller, L. C., Lechner, R. E., and Rugs, D. (1985). Development of conversational responsiveness: Preschoolers’ use of responsive listener cues and relevant comments. *Developmental Psychology*, 21(3):473.

- [Morency et al., 2010] Morency, L.-P., de Kok, I., and Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84.
- [Murphy, 2004] Murphy, R. R. (2004). Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):138–153.
- [Mutlu and Forlizzi, 2008] Mutlu, B. and Forlizzi, J. (2008). Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 287–294.
- [Mutlu et al., 2009] Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., and Hagita, N. (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, HRI '09*, pages 61–68, New York, NY, USA. ACM.
- [Nabe et al., 2006] Nabe, S., Kanda, T., Hiraki, K., Ishiguro, H., Kogure, K., and Hagita, N. (2006). Analysis of human behavior to a communication robot in an open field. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction, HRI '06*, pages 234–241, New York, NY, USA. ACM.
- [Nembhard and Edmondson, 2006] Nembhard, I. M. and Edmondson, A. C. (2006). Making it safe: The effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 27(7):941–966.
- [Neustaedter et al., 2018] Neustaedter, C., Singhal, S., Pan, R., Heshmat, Y., Forghani, A., and Tang, J. (2018). From being there to watching: Shared and dedicated telepresence robot usage at academic conferences. *ACM Trans. Comput.-Hum. Interact.*, 25(6):33:1–33:39.

- [Neustaedter et al., 2016] Neustaedter, C., Venolia, G., Procyk, J., and Hawkins, D. (2016). To beam or not to beam: A study of remote telepresence attendance at an academic conference. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 418–431, New York, NY, USA. ACM.
- [Oliveira et al., 2018] Oliveira, R., Arriaga, P., Alves-Oliveira, P., Correia, F., Petisca, S., and Paiva, A. (2018). Friends or foes?: Socioemotional support and gaze behaviors in mixed groups of humans and robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, pages 279–288, New York, NY, USA. ACM.
- [Orlikowski, 2002] Orlikowski, W. J. (2002). Knowing in practice: Enacting a collective capability in distributed organizing. *Organization science*, 13(3):249–273.
- [Oswick and Noon, 2014] Oswick, C. and Noon, M. (2014). Discourses of diversity, equality and inclusion: trenchant formulations or transient fashions? *British Journal of Management*, 25(1):23–39.
- [Park et al., 2019] Park, H. W., Grover, I., Spaulding, S., Gomez, L., and Breazeal, C. (2019). A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 687–694.
- [Pelikan et al., 2018] Pelikan, H. R. M., Cheatle, A., Jung, M. F., and Jackson, S. J. (2018). Operating at a distance - how a teleoperated surgical robot reconfigures teamwork in the operating room. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):138:1–138:28.
- [Pereira et al., 2011] Pereira, A., Leite, I., Mascarenhas, S., Martinho, C., and Paiva, A. (2011). Using empathy to improve human-robot relationships. In Lamers, M. H. and Verbeek, F. J., editors, *Human-Robot Personal Relationships*, pages 130–138, Berlin, Heidelberg. Springer Berlin Heidelberg.

- [Quigley et al., 2009] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., and Ng, A. Y. (2009). Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe.
- [Ramachandran et al., 2018] Ramachandran, A., Huang, C.-M., Gartland, E., and Scassellati, B. (2018). Thinking aloud with a tutoring robot to enhance learning. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 59–68.
- [Ramachandran et al., 2016] Ramachandran, A., Litoiu, A., and Scassellati, B. (2016). Shaping productive help-seeking behavior during robot-child tutoring interactions. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 247–254. IEEE.
- [Rempel et al., 1985] Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *Journal of personality and social psychology*, 49(1):95.
- [re:Work, 2020] re:Work (2020). Tool: Foster psychological safety. <https://rework.withgoogle.com/guides/understanding-team-effectiveness/steps/foster-psychological-safety/>.
- [Ribeiro et al., 2014] Ribeiro, T., Pereira, A., Di Tullio, E., Alves-Oliveira, P., and Paiva, A. (2014). From thalamus to skene: High-level behaviour planning and managing for mixed-reality characters. In *Proceedings of the IVA 2014 Workshop on Architectures and Standards for IVAs*.
- [Riordan et al., 1983] Riordan, C. A., Marlin, N. A., and Kellogg, R. T. (1983). The effectiveness of accounts following transgression. *Social Psychology Quarterly*, pages 213–219.
- [Robinette et al., 2015] Robinette, P., Howard, A. M., and Wagner, A. R. (2015). Timing is key for robot trust repair. In *International Conference on Social Robotics*, pages 574–583. Springer.

- [Robinette et al., 2017] Robinette, P., Howard, A. M., and Wagner, A. R. (2017). Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems*, 47(4):425–436.
- [Roger and Nesshoever, 1987] Roger, D. and Nesshoever, W. (1987). Individual differences in dyadic conversational strategies: A further study. *British Journal of Social Psychology*, 26(3):247–255.
- [Rokach, 2014] Rokach, A. (2014). Leadership and loneliness. *International Journal of Leadership and Change*, 2(1):6.
- [Rozovsky, 2015] Rozovsky, J. (2015). The five keys to a successful google team. <https://rework.withgoogle.com/blog/five-keys-to-a-successful-google-team/>.
- [Šabanović et al., 2013] Šabanović, S., Bennett, C., Chang, W.-l., and Huber, L. (2013). Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. In *IEEE International Conference on Rehabilitation Robotics*, volume 2013, pages 1–6.
- [Sabelli and Kanda, 2016] Sabelli, A. M. and Kanda, T. (2016). Robovie as a mascot: A qualitative study for long-term presence of robots in a shopping mall. *International Journal of Social Robotics*, 8(2):211–221.
- [Sabharwal, 2014] Sabharwal, M. (2014). Is diversity management sufficient? organizational inclusion to further performance. *Public Personnel Management*, 43(2):197–217.
- [Salem et al., 2015] Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (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*, pages 141–148. ACM.
- [Salvini et al., 2010] Salvini, P., Ciaravella, G., Yu, W., Ferri, G., Manzi, A., Mazzolai, B., Laschi, C., Oh, S. R., and Dario, P. (2010). How safe are service robots in urban environments? bullying a robot. In *19th International Symposium in Robot and Human Interactive Communication*, pages 1–7.

- [Saup   and Mutlu, 2015] Saup  , A. and Mutlu, B. (2015). The social impact of a robot co-worker in industrial settings. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI ’15*, pages 3613–3622, New York, NY, USA. ACM.
- [Scassellati et al., 2018] Scassellati, B., Boccanfuso, L., Huang, C.-M., Mademtzi, M., Qin, M., Salomons, N., Ventola, P., and Shic, F. (2018). Improving social skills in children with asd using a long-term, in-home social robot. *Science Robotics*, 3(21):1–9.
- [Schwartz et al., 1978] Schwartz, G. S., Kane, T. R., Joseph, J. M., and Tedeschi, J. T. (1978). The effects of post-transgression remorse on perceived aggression, attributions of intent, and level of punishment. *British Journal of Social and Clinical Psychology*, 17(4):293–297.
- [Schweitzer et al., 2006] Schweitzer, M. E., Hershey, J. C., and Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes*, 101(1):1–19.
- [Shah et al., 2011] Shah, J., Wiken, J., Williams, B., and Breazeal, C. (2011). Improved human-robot team performance using chaski, a human-inspired plan execution system. In *Proceedings of the 6th International Conference on Human-robot Interaction, HRI ’11*, pages 29–36, New York, NY, USA. ACM.
- [Shellenbarger, 2018] Shellenbarger, S. (2018). Alexa: Don’t let my 2-year-old talk to you that way. *Wall Street Journal*.
- [Shen et al., 2018] Shen, S., Slovak, P., and Jung, M. F. (2018). ”stop. i see a conflict happening.”: A robot mediator for young children’s interpersonal conflict resolution. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI ’18*, pages 69–77, New York, NY, USA. ACM.
- [Shiomi et al., 2010] Shiomi, M., Kanda, T., Ishiguro, H., and Hagita, N. (2010). A larger audience, please!: Encouraging people to listen to a guide robot. In *Proceedings of the*

- 5th ACM/IEEE International Conference on Human-robot Interaction, HRI '10*, pages 31–38, Piscataway, NJ, USA. IEEE Press.
- [Shiomi et al., 2007] Shiomi, M., Kanda, T., Koizumi, S., Ishiguro, H., and Hagita, N. (2007). Group attention control for communication robots with wizard of oz approach. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction, HRI '07*, pages 121–128, New York, NY, USA. ACM.
- [Shore et al., 2011] Shore, L. M., Randel, A. E., Chung, B. G., Dean, M. A., Holcombe Ehrhart, K., and Singh, G. (2011). Inclusion and diversity in work groups: A review and model for future research. *Journal of management*, 37(4):1262–1289.
- [Short and Matarić, 2017] Short, E. and Matarić, M. J. (2017). Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions. In *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN '17*, pages 385–390, Lisbon, Portugal. IEEE Press.
- [Short et al., 2017] Short, E. S., Swift-Spong, K., Shim, H., Wisniewski, K. M., Zak, D. K., Wu, S., Zelinski, E., and Matarić, M. J. (2017). Understanding social interactions with socially assistive robotics in intergenerational family groups. In *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN '17*, pages 236–241, Lisbon, Portugal. IEEE Press.
- [Sigal et al., 1988] Sigal, J., Hsu, L., Foodim, S., and Betman, J. (1988). Factors affecting perceptions of political candidates accused of sexual and financial misconduct. *Political Psychology*, pages 273–280.
- [Siino et al., 2008] Siino, R. M., Chung, J., and Hinds, P. J. (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*, pages 558–562. IEEE.

- [Simons and Peterson, 2000] Simons, T. L. and Peterson, R. S. (2000). Task conflict and relationship conflict in top management teams: the pivotal role of intragroup trust. *Journal of applied psychology*, 85(1):102.
- [Skantze, 2017] Skantze, G. (2017). Predicting and regulating participation equality in human-robot conversations: Effects of age and gender. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 196–204, New York, NY, USA. ACM.
- [Skantze et al., 2015] Skantze, G., Johansson, M., and Beskow, J. (2015). Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 67–74, New York, NY, USA. ACM.
- [Skowronski and Carlston, 1987] Skowronski, J. J. and Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of personality and social psychology*, 52(4):689.
- [Skowronski and Carlston, 1989] Skowronski, J. J. and Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological bulletin*, 105(1):131.
- [Smith and Powell, 1988] Smith, C. M. and Powell, L. (1988). The use of disparaging humor by group leaders. *Southern Speech Communication Journal*, 53(3):279–292.
- [Smithson and Verkuilen, 2006] Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54.
- [Stasser, 1999] Stasser, G. (1999). A primer of social decision scheme theory: Models of group influence, competitive model-testing, and prospective modeling. *Organizational Behavior and Human Decision Processes*, 80(1):3–20.
- [Stoll et al., 2018] Stoll, B., Reig, S., He, L., Kaplan, I., Jung, M. F., and Fussell, S. R. (2018). Wait, can you move the robot?: Examining telepresence robot use in collaborative



- teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, pages 14–22, New York, NY, USA. ACM.
- [Strohkorb et al., 2016] Strohkorb, S., Fukuto, E., Warren, N., Taylor, C., Berry, B., and Scassellati, B. (2016). Improving human-human collaboration between children with a social robot. In *25th IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN '16, pages 551–556, New York, NY, USA. IEEE Press.
- [Strohkorb Sebo et al., 2020a] Strohkorb Sebo, S., Dong, L. L., Chang, N., and Scassellati, B. (2020a). Strategies for the inclusion of human members within human-robot teams. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 309–317, New York, NY, USA. Association for Computing Machinery.
- [Strohkorb Sebo et al., 2019] Strohkorb Sebo, S., Krishnamurthi, P., and Scassellati, B. (2019). “i don’t believe you”: Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 57–65.
- [Strohkorb Sebo et al., 2020b] Strohkorb Sebo, S., Stoll, B., Scassellati, B., and Jung, M. (2020b). The studies contained in this review and their corresponding data can be found at <https://docs.google.com/spreadsheets/d/1w5fgxkooqhryz7aqnvyyq0rumzbsitoqrunph9ej4vxg/edit?usp=sharing>.
- [Strohkorb Sebo et al., 2018] Strohkorb Sebo, S., Traeger, M., Jung, M., and Scassellati, B. (2018). The ripple effects of vulnerability: The effects of a robot’s vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, pages 178–186, New York, NY, USA. ACM.
- [Stubbe, 1998] Stubbe, M. (1998). Are you listening? cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics*, 29(3):257–289.

- [Stubbs et al., 2007] Stubbs, K., Hinds, P. J., and Wettergreen, D. (2007). Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems*, 22(2):42–50.
- [Stubbs Koman and Wolff, 2008] Stubbs Koman, E. and Wolff, S. B. (2008). Emotional intelligence competencies in the team and team leader: A multi-level examination of the impact of emotional intelligence on team performance. *Journal of Management Development*, 27(1):55–75.
- [Sung et al., 2010] Sung, J., Grinter, R. E., and Christensen, H. I. (2010). Domestic robot ecology. *International Journal of Social Robotics*, 2(4):417–429.
- [Taggar and Ellis, 2007] Taggar, S. and Ellis, R. (2007). The role of leaders in shaping formal team norms. *The Leadership Quarterly*, 18(2):105–120.
- [Tajfel, 1982] Tajfel, H. (1982). Social psychology of intergroup relations. *Annual review of psychology*, 33(1):1–39.
- [Takayama and Go, 2012] Takayama, L. and Go, J. (2012). Mixing metaphors in mobile remote presence. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12*, pages 495–504, New York, NY, USA. ACM.
- [Talamadupula et al., 2010] Talamadupula, K., Benton, J., Kambhampati, S., Schermerhorn, P., and Scheutz, M. (2010). Planning for human-robot teaming in open worlds. *ACM Transactions on Intelligent Systems and Technology*, 1(2):14:1–14:24.
- [Tanaka et al., 2007] Tanaka, F., Cicourel, A., and Movellan, J. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences of the United States of America*, 104:17954–8.
- [Tasaki et al., 2004] Tasaki, T., Matsumoto, S., Ohba, H., Toda, M., Komatani, K., Ogata, T., and Okuno, H. G. (2004). Dynamic communication of humanoid robot with multiple people based on interaction distance. In *13th IEEE International Workshop on Robot and Human Interactive Communication, RO-MAN '04*, pages 71–76, Kurashiki, Okayama, Japan. IEEE Press.

- [Tennent et al., 2019] Tennent, H., Shen, S., and Jung, M. (2019). Michbot: A peripheral robotic object to shape conversational dynamics and team performance. In *14th ACM/IEEE International Conference on Human-Robot Interaction, HRI '19*, pages 133–142, Daegu, South Korea. IEEE Press.
- [Thompson et al., 2017] Thompson, C., Mohamed, S., Louie, W. G., He, J. C., Li, J., and Nejat, G. (2017). The robot tangy facilitating trivia games: A team-based user-study with long-term care residents. In *2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS)*, pages 173–178.
- [Thompson and Hunnicutt, 1944] Thompson, G. G. and Hunnicutt, C. W. (1944). The effect of repeated praise or blame on the work achievement of 'introverts' and 'extroverts.'. *Journal of Educational Psychology*, 35(5):257.
- [Tomasello et al., 2005] Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691.
- [Tomlinson et al., 2004] Tomlinson, E. C., Dineen, B. R., and Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management*, 30(2):165–187.
- [Traeger et al., 2020] Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., and Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*, 117(12):6370–6375.
- [Truong et al., 2011] Truong, K. P., Poppe, R., Kok, I. d., and Heylen, D. (2011). A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Twelfth Annual Conference of the International Speech Communication Association*.
- [Tuckman, 1965] Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological bulletin*, 63(6):384.

- [Ulloa and Adams, 2004] Ulloa, B. C. R. and Adams, S. G. (2004). Attitude toward teamwork and effective teaming. *Team Performance Management*, 10(7-8):145–151.
- [Unhelkar et al., 2020] Unhelkar, V. V., Li, S., and Shah, J. A. (2020). Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 329–341.
- [Utami and Bickmore, 2019] Utami, D. and Bickmore, T. (2019). Collaborative user responses in multiparty interaction with a couples counselor robot. In *14th ACM/IEEE International Conference on Human-Robot Interaction, HRI '19*, pages 294–303, Daegu, South Korea. IEEE Press.
- [van Kleef, 2016] van Kleef, G. A. (2016). *The interpersonal dynamics of emotion*. Cambridge University Press.
- [Van Kleef et al., 2010] Van Kleef, G. A., De Dreu, C. K., and Manstead, A. S. (2010). An interpersonal approach to emotion in social decision making: The emotions as social information model. *Advances in experimental social psychology*, 42:45–96.
- [Van Son et al., 2008] Van Son, R., Wesseling, W., Sanders, E., van den Heuvel, H., et al. (2008). The ifadv corpus: A free dialog video corpus.
- [Vázquez et al., 2017] Vázquez, M., Carter, E. J., McDorman, B., Forlizzi, J., Steinfeld, A., and Hudson, S. E. (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, HRI '17*, pages 42–52, New York, NY, USA. ACM.
- [Vertesi, 2012] Vertesi, J. (2012). Seeing like a rover: Visualization, embodiment, and interaction on the mars exploration rover mission. *Social Studies of Science*, 42(3):393–414.

- [Wada et al., 2004] Wada, K., Shibata, T., Saito, T., and Tanie, K. (2004). Effects of robot-assisted activity for elderly people and nurses at a day service center. *Proceedings of the IEEE*, 92(11):1780–1788.
- [Ward, 2006] Ward, N. (2006). Non-lexical conversational sounds in american english. *Pragmatics & Cognition*, 14(1):129–182.
- [Ward and Tsukahara, 2000] Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.
- [Warneken et al., 2014] Warneken, F., Steinwender, J., Hamann, K., and Tomasello, M. (2014). Young children’s planning in a collaborative problem-solving task. *Cognitive Development*, 31:48–58.
- [Waytz et al., 2014] Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117.
- [Weick, 1993] Weick, K. E. (1993). The collapse of sensemaking in organizations: The mann gulch disaster. *Administrative science quarterly*, pages 628–652.
- [Wheeless, 1978] Wheeless, L. R. (1978). A follow-up study of the relationships among trust, disclosure, and interpersonal solidarity. *Human Communication Research*, 4(2):143–157.
- [Wittenburg et al., 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.
- [Woolley et al., 2010] Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688.
- [Yamazaki et al., 2012] Yamazaki, A., Yamazaki, K., Ohyama, T., Kobayashi, Y., and Kuno, Y. (2012). A techno-sociological solution for designing a museum guide robot:

- Regarding choosing an appropriate visitor. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, pages 309–316, New York, NY, USA. ACM.
- [You and Robert, 2017] You, S. and Robert, L. (2017). Teaming up with robots: An imoi (inputs-mediators-outputs-inputs) framework for human-robot teamwork. *International Journal of Robotic Engineering*, pages 1–7.
- [You and Robert, 2018] You, S. and Robert, L. (2018). Emotional attachment, performance, and viability in teams collaborating with embodied physical action (epa) robots. *Journal of the Association for Information Systems*, 19:377–407.
- [Zaga et al., 2015] Zaga, C., Lohse, M., Truong, K. P., and Evers, V. (2015). The effect of a robot’s social character on children’s task engagement: Peer versus tutor. In *Social Robotics*, pages 704–713. Springer.
- [Zand, 1972] Zand, D. E. (1972). Trust and managerial problem solving. *Administrative science quarterly*, pages 229–239.
- [Zellmer-Bruhn and Gibson, 2006] Zellmer-Bruhn, M. and Gibson, C. (2006). Multinational organization context: Implications for team learning and performance. *Academy of management journal*, 49(3):501–518.
- [Zubrycki and Granosik, 2016] Zubrycki, I. and Granosik, G. (2016). Understanding therapists’ needs and attitudes towards robotic support. the roboterapia project. *International Journal of Social Robotics*, 8(4):553–563.