Human Perception of Robot Failure and Explanation

Huy Quyen Ngo

CMU-RI-TR-24-12
May 2024

School of Computer Science
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Aaron Steinfeld, Chair
Henny Admoni
Nikolas Martelaro
Roshni Kaushik

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

Keywords: Non-verbal Cues and Expressiveness, User-centered Design of Robots

Abstract

In recent years, researchers have extensively used non-verbal gestures, such as head and arm movements, to express the robot's intentions and capabilities to humans. Inspired by past research, we investigated how different explanation modalities can aid human understanding and perception of how robots communicate failures and provide explanations during block pick-and-place tasks. Through an in-person experiment, we studied four modes of explanations: Head, Head & Arm, Head & Image Projection, and Head & Speech. They were used to explain four types of failures: Out Of Reach, Object Size, Grasp Failure, and Perception Failure. We found that speech explanations were preferred to non-verbal and visual cues in terms of similarity to humans. Additionally, projection had a comparable effect in explanation compared to other non-verbal modules. The findings also suggested that in-person and online studies can produce consistent results.

Acknowledgments

I want to give a heartfelt gratitude to my advisor and the chair of my committee, Aaron Steinfeld, for this guidance, advice, and support from the beginning of my journey. His encouragement has been crucial in navigating my research and academic growth.

I am also thankful to Henny Admoni, Nikolas Martelaro, and Roshni Kaushik for their engaging and productive discussions on my research and ideas. Their perspective and expertise have been important in shaping a lot of my ideas and improving the quality of my work.

Furthermore, I extend my appreciation to all members of the Transportation, Bots, & Disability Lab. I am grateful for their support in my qualifying exam rehearsals, in which they provided insightful feedback to improve my presentation skills and research skills.

This research was supported by the Office of Naval Research award N00014-181-2503.

Contents

1 Introduction **1**

2 Related Work **3**

2.1 Robot Failure and Explanation 3

2.1.1 Robot Failure 3

2.1.2 Robot Explanation 3

2.2 Human Perceptions towards Robots 4

2.2.1 Trust 4

2.2.2 Competence 5

2.2.3 Safety and Security 5

2.2.4 Empathy and Engagement 5

2.3 Robot Facial Expression 6

3 Method **7**

3.1 Robot Description 7

3.2 Experiment Setup 7

3.3 Failure and Explanation 9

3.3.1 Failure Types 9

3.3.2 Explanation Conditions 9

3.4 Measures 10

3.5 User Study Design 11

3.5.1 Participants 12

3.5.2 Study Procedure 12

4 Results **15**

4.1 Unexpectedness 15

4.2 Human-Robot Difference 16

4.3 Level of Detail 19

4.4 Conciseness 19

4.5 Post-study Questions: Trust, Competence, Need for Explanation, and Overall Perception 21

4.5.1 Trust 21

4.5.2 Competence 21

4.5.3 Need for Explanation 21

4.5.4 Overall Perception 24

5 Discussion **25**

6 Limitations **27**

7 Conclusion **29**

Bibliography **31**

List of Figures

3.1 Locations of the successful blocks (blue) and the four failures (labeled). The arm position and projected white area with red arc were used for *Head & Projection* during a *Reach* failure. 8

4.1 Unexpectedness scores for four types of failures (left) and four modes of explanations (right). The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. 16

4.2 The distribution of Unexpectedness scores of four explanation conditions in each type of failure. The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. Boxes in blue, orange, green, and red represent *Head*, *Head & Arm*, *Head & Projection*, and *Head & Speech* conditions, respectively. 17

4.3 Human-Robot Difference scores in four types of failures (left) and four modes of explanations (right). The white diamond icons indicate the mean scores. $**$ represents $p - value < 0.01$, and $***$ represents $p - value < 0.001$. 18

4.4 The distribution of Human-Robot Difference scores of four explanation conditions in each type of failure. The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. Boxes in blue, orange, green, and red represent *Head*, *Head & Arm*, *Head & Projection*, and *Head & Speech* conditions, respectively. 19

4.5 Level of Explanation Detail scores in four types of failures (left) and four modes of explanations (right). The white diamond icons indicate the mean scores. 20

4.6 Conciseness scores in four types of failures (left) and four modes of explanations (right). The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. 20

4.7 Distribution of scores in trust-related questions. The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. 22

4.8 Distribution of scores in competence-related questions. The white diamond icons indicate the mean scores. 23

4.9 Distribution of scores in need-for-explanation-related questions. The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. Questions 21, 22, and 23 are discussed separately as they are not based on Likert-type items. 23

List of Tables

3.1 Post-Trial Questions. * indicates questions adapted from [23]. 10
3.2 Post-Study Questionnaire. The * indicates questions adapted from Han et al. [23]. The ** indicates open-ended question. 11

Chapter 1

Introduction

Humans usually provide cues for behaviors in daily life, especially following an unfavorable action, such as failing to do a task. Thus, humans likely expect robots to explain their behaviors in failure situations, verbally or non-verbally. Past work shows that the ability of robots to explain themselves can have a positive effect on the robots' perceived trustworthiness [15, 32] and human-likeness [4].

In this work, we extended a prior study on robot explanation in a cup-handover task [23]. The study examined the handover failure conditions of a cup that was out of reach from the robot arm. The robot explained each failure in handing the cup to participants by looking or shaking its head at the cup and pointing to the cup with its arm. The study found that without head shakes, both the Look and Look & Point conditions were neutral relative to expectedness for participants. Moreover, they found that No Cue (do nothing) increased the level of unexpectedness, and adding head shakes made the robot's behavior more unexpected across all conditions. The study also indicated that the robot should concisely explain its behavior in all circumstances, preferably if the explanations are *in situ*, but only a small percentage of participants thought that humans should explain failures in the same non-verbal way.

In our study, we considered Look as the primary head movement during the explanation. We also had our robot explain its behavior *in situ* when failures happened. In addition, we introduced two new components of explanation, namely Image Projection and Speech, inspired by later work by the same team [7, 22], as other ways of explaining robot failure behaviors.

In our experiment, the robot performed a routine of picking up blocks on the table and placing them onto a tray. We studied four modes of explanations: Head (only look at the object), Head & Arm (look at the object and move the arm), Head & Projection (look at the object and project an image on the workstation), and Head & Speech (look at the object and utter a statement). The two newly added components of Projection and Speech have been proven to be effective in communicating explanations of the robot failures in other contexts [7, 22]. These four conditions were used to explain four types of failures: Out Of Reach (the object is too far away from the robot), Object Size (the object is too large for the robot to grip), Grasp Failure (the robot cannot securely grasp the object), and Perception Failure (the robot hallucinates an object). Our work also served as a partial, in-person replication of the prior online study by including elements from that work [23].

We designed a 26-item questionnaire, partly adapted from Han et al. [23], to measure some key aspects of human-robot interaction: Unexpectedness, Human-Robot Difference, Level of Detail, Conciseness, Trust, Competence, and Need for Explanation in failure situations.

In summary, our contributions in this paper are:

1. An in-person, partial replication of a prior study on robot explanation, showing non-verbal gestures have similar effects on human perception using a different robot, thus confirming the consistency between online and in-person experiments and across robot platforms;
2. Findings showing that projected images for robot performance explanation perform similarly to non-verbal gestures; and
3. Evidence for a prior conjecture about speech being preferred for explaining robot failure and performance.

Chapter 2

Related Work

2.1 Robot Failure and Explanation

2.1.1 Robot Failure

Robotic systems can experience multiple types of failures, either internally from the robot software and hardware, or externally from users and surrounding environments. Honig et al. [25] discussed a taxonomy of human-robot failures in domestic robots that are most frequently seen by customers. Carlson et al. [8] classified in-depth physical failures in the end effector of the robot. Thus, along with the Out Of Reach failure from [23], we studied three other types of failures that are common in a pick-and-place task, namely environment failure (e.g., Out Of Reach, Object Size), control failure (e.g., Grasp Failure), and sensor failure (e.g., Perception Failure).

2.1.2 Robot Explanation

In the motivating prior work that was conducted online, Han et al. [23] studied robot explanation during a cup handover task in which the cup was out of reach from the robot arm. As mentioned, the robot explained its behavior through non-verbal cues, such as arm and head movements, to express the robot's difficulty in reaching the cup placed far away on the table, including Look only, Look & Point, and No Cue, coupled with Headshake or No Headshake. They found that removing headshakes decreased the level of unexpectedness to the explanation in both Look & Point and Look only, and that the robot should always give cues to be perceived as less unexpected. Building on that idea, we eliminated the Headshake portion of the cue, so the only motion for the robot head was to look.

Research on consistency between online and in-person studies has been sparse. Han et al. [23] conducted their experiment online via Amazon's Mechanical Turk. Thus, we conducted an in-person experiment to confirm the consistency of our results with those of their online experiment.

In a later work, Han et al. [22] used verbal and projection indicators, coupled with head and arm motion replay, to communicate past causal information related to tasks. Cao et al. [7] discussed a method of robot proficiency self-assessment, Assumption-Alignment Tracking (AAT), that can make the robot aware of the environment, robot hardware, and assigned tasks. Thus, failure modes can be monitored and assessed to evaluate the robot's capability of performing a task. Likewise, Rosenthal et al. [39] studied the effectiveness of the verbal modality in parallel with visual modality during robot operation. Kelly et al. [27] suggested that speech and nonverbal gestures can be used to complement the meaning of each other, such that speech is used to describe symbolic meanings while gestures are used for holistic information. Sebo et al. [42] suggested using verbal apology to repair human's trust in robot, which was previously broken, in a human-robot competitive game setting. Moreover, verbal explanation has been proven to be effective in failure situations [10, 28]. Thus, we chose to study both visual modalities (projection, gestures) and verbal modalities (speech) in robot failure and explanation.

Image projection is versatile in communicating important information about the contexts of the tasks and behaviors of robots. Previous work by Han et al. [22] demonstrated the effectiveness of projection in revealing task-related information. Projections can indicate boundaries around robots [50] (e.g., maximum robot reach), display information about the robot [51] (e.g., maximum gripper opening), mark locations [43] (e.g., a red X for a failure location), and communicate misperceptions [22] (e.g., hallucinated objects).

Head motion and eye gaze were shown to be important during interactions with robots [6, 18, 38, 45, 46]. Eye gaze can help humans understand robots during tasks. Moreover, facial expressions have been widely used for effective communications in human-human and human-robot collaborations. Having expressive features makes robots be perceived as more intelligent, human-like and trustworthy [26], and even impacts failure situations [34].

Finally, verbal communications in failure situations can be categorized into several types: Apology [10, 30], Explanation [10, 28], and Interjection (Expression of Concern) [44]. Such types of communications can be effective in providing context of and conveying attitude towards the failing behaviors, especially when paired with failure behaviors such as control failures, sensor failures, and environmental / external failures [8, 23, 25].

2.2 Human Perceptions towards Robots

Human's perceptions towards robots have been explored in previous human-robot interaction works. Those perceptions include trust [5, 12, 14, 47, 54], competence [10, 40, 48], safety [1, 2, 52], empathy and engagement [9, 13, 21].

2.2.1 Trust

Trust in robots and autonomous systems has been extensively studied to promote healthy human-robot interaction. Anjomshoae et al. [5] claimed that trust (along with transparency) is the most prominent drive in explanations, and that trust can increase the users' confidence in the systems by understanding how the systems work. When robots explain their actions, humans can correct their mental models and calibrate their level of trust in the systems [12]. Moreover, Yang et al. [54] indicated that trust can be measured through a series of interactions with automation systems. Tolmeijer et al. [47] suggested that offering explanations in failure situations can help mitigate failure and repair trust. Desai et al. [14] stated that trust can change in real time, and that early decreases in robot reliability reduced real-time trust from people compared to later decreases.

Moreover, trust is influenced not only by actions of the robots, but also by their appearances. Li et al. [31] suggested that the more human-like the robot, the more likeable it is perceived by humans, leading to a higher level of perceived trust in the robot. According to Phillips et al. [36], humans can categorize and predict a robot's human-likeness based on its appearance, on a scale of 0 (Not human-like at all) to 100 (Just like a human). The Fetch robot we used in this study [53] has a score of 9.08, which can be considered as not human-like. Thus, we anticipated that the perceived trust in the robot will be low due to its non-humanoid appearance. However, research shown that trust can increase with more interactions with robots [54], and that humans can adjust their perceived trust in robots when the robots explain their behaviors [12]. Thus, if a non-humanoid robot (i.e., the Fetch robot) can formally explain its failures, the perceived trust in the robot might be recovered throughout the entire experiment.

2.2.2 Competence

Humans usually take fellow humans' competence for granted during interaction, whether such interaction is verbal or non-verbal [48]. Scheunemann et al. [40] found that humans prefer to physically interact with social robots that are perceived as warm and competent. Choi et al. [10] measured competence in robots with Likert items based on capability, intelligence, and skillfulness, but did not find significant difference in perceived competence in the case of robots explaining their failures. They also claimed that providing an explanation can increase the perceived warmth from humanoid robots, but not for non-humanoids. In our experiment, we explored the effect of a non-humanoid robot's movements as explanations in robot failure situations, along with the perceived warmth and competence of participants.

2.2.3 Safety and Security

Safety and security are key requirements in designing human-robot interaction [1, 52]. Akalin et al. [1] defined safety as related to perception of possible physical harm. In a later work, Akalin et al. [2] investigated Feeling of Security (humans feeling safe around the robot) and Co-Experience (robot being a good teammate) factors, in which they found that there was a strong correlation between trust and perceived safety. Thus, safety and trust can be studied hand-in-hand with each other.

2.2.4 Empathy and Engagement

Empathy and engagement with robots during human-robot interaction settings were extensively studied in past research. Celiktutan et al. [9] conducted a study involving two participants and a small humanoid robot to investigate the relationship between robot personality and human engagement. They found correlations between humans' personality traits and robot partners' level of extroversion through empathy-related survey questions during interaction. Hall et al. [21] showed that humans' engagement with robots depends on the robots' gestures, especially nodding. In addition, de Kervenoael et al. [13] found that robots' perceived empathy and engagement with humans correlated with humans' intentions to interact with robots in hospitality service settings.

2.3 Robot Facial Expression

Facial expressions have been widely used for effective communications, not only between humans but also in human and robot collaboration. Humans are exceptionally good in recognizing and understanding facial expressions, and humans can recognize facial expressions of robots immediately, especially humanoid robots with explicit faces [17]. Thus, a robot's facial expressions can aid people in understanding and engaging with it. Moreover, Morales et al. [34] indicated that expressive robot facial features have impacts on humans' perceived intelligence, friendliness and human-likeness towards robots, further indicating the importance of testing and designing robots for human-robot interaction. Reyes et al. [38] designed a minimalist robotic face to understand the effect of robot's facial expressions on human's emotional feedback, in which they found that negative expressions such as angry was best recognized in robotic faces by humans in collaborative tasks. Ge et al. [18] developed an expressive social robot that can express emotions such as happy, anger, and sadness, and those emotions can be expressed through only a few modalities such as eye lids, lips, eyes, and eyebrows to test its ability to imitate human's facial expressions. Furthermore, robots having eyebrows are perceived by humans as more mature, intelligent, human-like, and trustworthy, as opposed to not having a mouth [26].

Human communications usually consist of a combination of facial expressions and speech as channels for information exchange [37]. However for human-robot interaction, previous research endeavors have only focused on independent modules modes of explanations: Non-verbal gestures with arm and head [23], speech as verbal cues [22, 28]. In particular, there has been limited research in using facial expressions as a mode of explanation in failure situations, as well as using paired facial expressions on robots and verbal explanations in failure situations. Thus, the pairing of facial expressions and verbal / nonverbal behaviors can be used to explain robot failures, and can be used to recover trust through interactions [12, 54].

Chapter 3

Method

Inspired by the prior work of Han et al. [23], we designed an in-person experiment to collect more reliable responses on human perception of robot failure and explanation.

3.1 Robot Description

While prior work [23] used a Baxter robot [16] for the cup handover task, our study of a pick-and-place task used a Fetch robot [53], which is a mobile manipulator with a 7-DOF arm and a head with built-in cameras. Fetch's arm has the maximum reach of 940.5 mm, which is enough for the pick-and-place task. Furthermore, the smaller frame of a Fetch robot compared to a Baxter robot makes it less imposing. While Fetch has a movable head with "eyes" (cameras), it has no explicit face, which imposes a hardware limitation for eye gaze and facial expressions. Robot Operating Systems (ROS) in Ubuntu 18.04 was used to control the robot.

3.2 Experiment Setup

The arrangement of the table for the pick-and-place task is shown in Figure 3.1. On the table, there were seven blocks of different sizes, shapes, and colors scattered on the table. The robot was assigned a pick-and-place task, in which it had to locate all the blocks on the table, pick them up using its arm, and place them into the tray on the table. Among those blocks, some were designed to be picked up by the robot gripper, while others were there as decoys so that participants could not predict the next block or failure type. For the *Head & Projection* explanation condition, a ceiling-mounted external projector was used to project images onto the table. Although the duration of projection stayed the same, the content of the images was tailored for each failure type. Participants were only informed about the existence of the projection module, without knowing what the projections looked like.

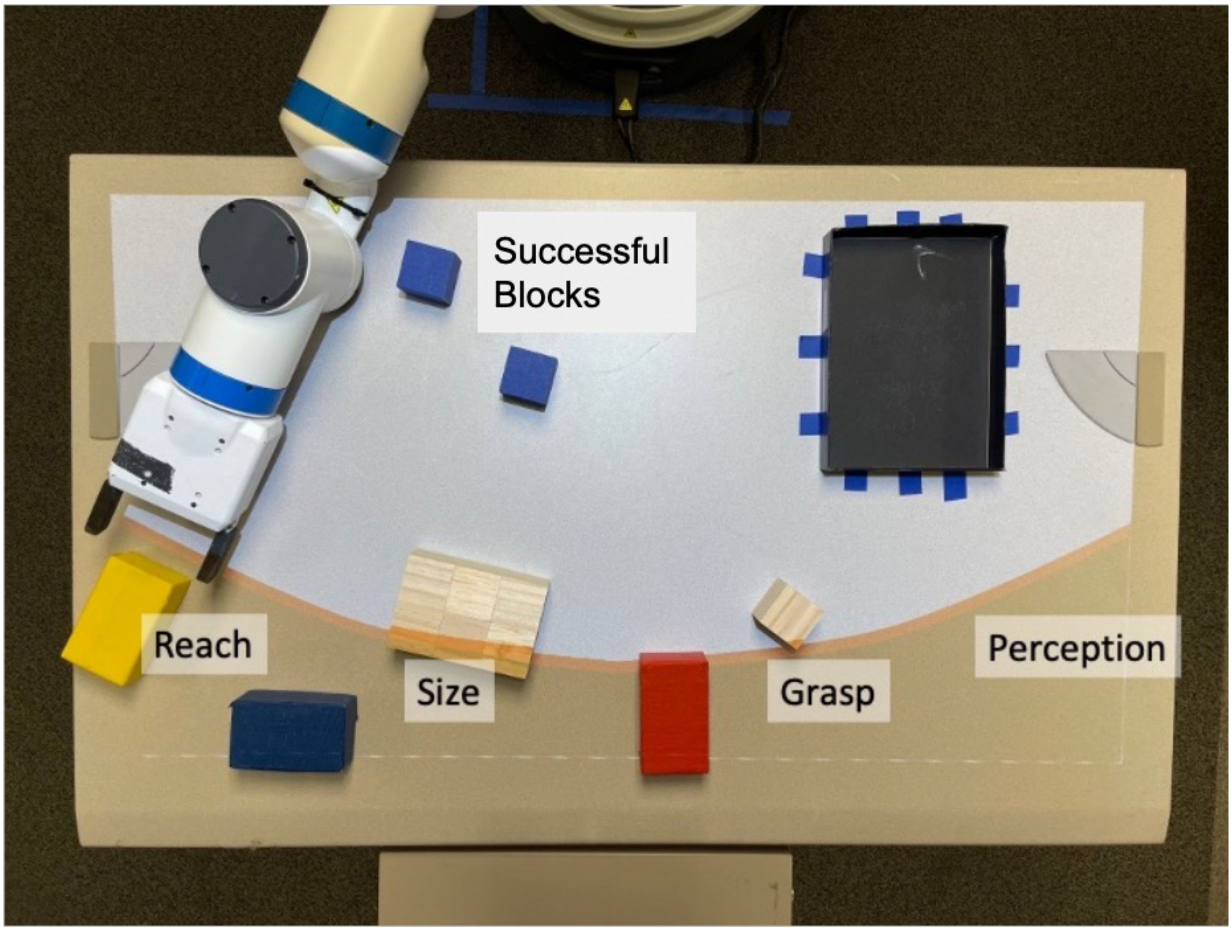

Figure 3.1: Locations of the successful blocks (blue) and the four failures (labeled). The arm position and projected white area with red arc were used for *Head & Projection* during a *Reach* failure.

3.3 Failure and Explanation

Prior work by Han et al. [23] emphasized the need for robots to explain their reaching failures using two main modalities, namely Head (Look or Shake) and Arm (Point). Building upon that, our work also used Fetch's head and arm for explanation. Furthermore, inspired by [7, 22], we incorporated Projection and Speech as two new components in the explanation conditions, and added three new failure types: *Size*, *Grasp*, and *Perception*.

3.3.1 Failure Types

Reach: The block was too far away from the robot arm to reach, even when the arm was fully extended. This was the same type of failure described in [23].

Size: The width of the block was larger than the maximum gripper opening, thus it could not pick up the block.

Grasp: The block was within the reach of the robot arm and was of suitable size for the gripper to grasp. However, the robot miscalculated the inverse kinematics of the arm, leading to an unstable grip. That resulted in the block slipping off the gripper after the gripper closed.

Perception: The block could not be found in the region where the head was pointing, but the robot still hallucinated a block in that area. Thus, the robot arm tried to grasp the hallucinated block, but no block was picked up.

3.3.2 Explanation Conditions

Head: We removed the head shakes from [23]. Instead, the robot only pointed its head at the location of the block.

Head & Arm: The robot pointed its head at the location of the block and moved its arm to form a gesture. In the case of a *Reach* failure, we mimicked the movements of the Baxter robot in [23] towards the block. In other failures, the robot attempted to grasp the block an additional time.

Head & Projection: The robot pointed its head at the location of the block and projected an image onto the table which contained a visual explanation for the robot's failure. The *Reach* failure displayed a red arc denoting the maximum reach of the robot arm (e.g., Figure 3.1). The *Size* failure displayed two red lines across the block showing the maximum gripper opening. The *Grasp* failure displayed a large red X on top of the block. The *Perception* failure displayed a red square where the block was hallucinated.

Head & Speech: The robot pointed its head at the location of the block and uttered a statement explaining its failure using a speaker. For the *Reach* failure, the statement was "My arm cannot reach the block, so I will not be able to pick the block." For the *Size* failure, the statement was "The block is too large, so I will not be able to pick the block." For the *Grasp* failure, the statement was "I was unable to grasp the block, so I will not be able to pick the block." For the *Perception* failure, the statement was "My camera is not working, so I will not be able to pick the block."

3.4 Measures

We prepared a post-trial survey based on questions used in [23] to measure participants' perceptions of robot failure and explanation, as shown in Table 3.1. With the aim of recreating the results from their work and extending our work with new explanation components (Projection and Speech), we merged questions from the prior study with new questions to measure the unexpectedness of the robot's behavior (Unexpectedness), the difference between the ways humans and robots explain themselves (Human-Robot Difference), the level of explanation detail (Level of Detail), and the how concise the explanation should be (Conciseness). To keep the survey questions internally consistent with each other, we asked questions that were very similar or contradictory to each other and used in most questions the 7-point Likert-type scale [41]. Each Likert-type item was coded as -3 (Strongly Disagree), -2 (Disagree), -1 (Moderately Disagree), 0 (Neutral), 1 (Moderately Agree), 2 (Agree), and 3 (Strongly Agree). Among the questions, 6 of them were adapted from [23]: questions 1-3 and 6-8.

Unexpectedness (Cronbach's α = 0.80)
1. I found the robot's behavior confusing.*
2. The robot's behavior matched what I expected. (Reversed)*
3. The robot's behavior surprised me.*
4. The robot's movements were natural. (Reversed)
5. The robot's movements were predictable. (Reversed)
Human-Robot Difference
6. If a person did what the robot did, they should both explain the same behavior in the same way.*
Level of Detail
7. The robot should give a very detailed explanation.*
Conciseness
8. The robot should concisely explain its behavior.*

Table 3.1: Post-Trial Questions. * indicates questions adapted from [23].

In addition to the questions in Table 3.1, we designed a post-study questionnaire (Table 3.2) to gather information about the participants' overall experience and perceptions of the robot. Along with trust, we wanted to measure participants' perceived robot competence and need for explanation when the robot explained its failures. In the line of questioning, we investigated human preference for how and when the robot should explain its behavior and whether robots need to provide explanations in failure situations. We also wanted to explore participants' assessments of the robot's movements and impressions of interacting with the robot. Finally, we wanted to record participants' overall perceptions about the study and the robot. We divided the questions into the subcategories of Trust, Competence, Need for Explanation and Overall Perception. All questions were Likert-type items except for questions 20, 21, and 22, which were multiple-choice questions with specific options, and question 27, which was open-ended.

Trust
9. I am comfortable engaging with a robot that uses movement to signal difficulty.
10. The robot’s movements make me more engaged with it.
11. The robot’s movements affect how much I trust it.
12. I feel empathy for the robot when it fails.
13. The robot is likable.
14. I felt warmth interacting with the robot.
15. The robot can be a good teammate.
16. I felt safe around the robot.
Competence
17. The robot’s movements were clear and lifelike.
18. The robot’s movements help me understand what it can do.
19. The robot is competent.
Need for Explanation
20. I wanted the robot to explain its behavior.*
21. Do you think it is important for the robot to get your attention before starting to explain its behavior?*
22. How should the robot get your attention before starting to explain its behavior?*
23. When would be the best time for the robot to explain its behavior?*
24. A robot signaling failure through its movements is important.
25. I want robots to announce failure out loud.
26. I prefer non-verbal actions from robots when they fail.
Overall Perception
27. Do you have any other comments about the robot’s behavior, its explanations, the robot itself, or the experiment?**

Table 3.2: Post-Study Questionnaire. The * indicates questions adapted from Han et al. [23]. The ** indicates open-ended question.

3.5 User Study Design

We ran each participant across all four modes of explanations. To address practice and other ordering effects, the types of failures and modes of explanations were each ordered using two different four-way Latin Squares [19]. This yielded 16 unique combinations of failure types and explanation conditions and counterbalanced both factors. Due to having four types of failures and four modes of explanations, our data includes 6 iterations over the 4-way pattern, totaling 24 participants.

3.5.1 Participants

In keeping with best practices, we sought gender balance within the 24 participants. Participants included 12 women and 12 men with the mean age to be 34.5. Among the participants, 11 of them were in the range of 18 to 25 years old (46%), 6 of them were in the range of 25 to 35 years old (25%), 2 of them were in the range of 35 to 45 years old (8%), and 5 of them were above 45 years old (21%). Participants' experience with robots ranged from no exposure to years of experience (building robots at school, having vacuum robots, taking robotics courses, participating in robotics studies, etc.).

3.5.2 Study Procedure

The participants were first introduced to the study by a researcher and then asked to sign a consent form, which contained a brief of the study procedure and purpose, risks engaging in the study, and compensation for the study. Before the experiment began, participants provided their demographic information, along with some information about their experience with robots. All questions were conducted using the Qualtrics website on a computer at the study location.

Each participant was given four trials to experience four combinations of robot failures and explanations. Each trial had four consecutive parts: *Success*, *Failure*, *Explanation*, and *Survey*. During the pick-and-place task, participants stood in front of the robot on the other side of the table.

Success: The trial started with the robot in its initial state: its arm was tucked into its torso and its head was held straight. The robot then began scanning the table to search for blocks to pick up. Upon finding two good candidates that were close to the robot (two small blue blocks on the table as seen in Figure 3.1), the robot successfully picked up these two blocks and placed them into the tray. These two manipulations were designed to be successful, indicating that the robot was doing its job properly and mitigating bias. Then, the robot moved on to the Failure part of the trial.

Failure: The robot began the Failure part of the trial by looking at one of the blocks or areas on the table (pertaining to one of the types of failures). After choosing its target, the robot attempted to grasp it by approaching the block area and closing its gripper once. When the gripper was not able to grasp the block, the Failure phase finished and the robot continued to the Explanation phase.

Explanation: Realizing that it could not pick up something, the robot executed one of the explanation conditions. The robot provided an explanation *in situ* when the failure happened. Then, the robot returned to its initial state.

Post-Trial Questions: Next, participants were asked to respond to our post-trial survey on Qualtrics about their observations in the trial. After their responses were recorded, the Survey phase of the trial concluded, marking the end of one trial. Participants returned to the table for the next trial.

After participants finished their four trials, they were asked to respond to a post-study questionnaire about their overall experience participating in the study as well as their general perception of the robot. The entire study took 45 minutes for each participant, and received $10 compensation for their time. This research was approved by Carnegie Mellon University's Institutional Review Board.

Chapter 4

Results

We analyzed Unexpectedness, Human-Robot Difference, Level of Detail, and Conciseness using two-way ANOVAs. For Trust, Competence, Need for Explanation and Overall Perception, we analyzed each question in each item individually.

4.1 Unexpectedness

The Unexpectedness item measures the level of unpredictability of the robot's behavior. The item has questions about whether the robot's behavior was confusing, surprising, natural, and predictable to participants during the trials.

We performed a two-way ANOVA to examine the effect of failure types and explanation conditions on the level of unexpectedness, with the distribution shown in Figure 4.1. From the analysis, we found a statistically significant main effect for failure types ($F(3, 80) = 4.51, p < 0.01$). The main effect of explanation conditions approached statistical significance ($F(3, 80) = 2.25, p = 0.09$), as did the interaction between failure types and explanation conditions ($F(9, 80) = 1.81, p = 0.08$). However, a post hoc pairwise comparison using Tukey's test with Holm-Bonferroni correction [24] ($H_0 : \mu_i = \mu_j$) revealed that there was no pairwise differences between Perception and Reach failures, as well as Perception and Size failures.

There were no statistically significant differences in unexpectedness between the *Head* and *Head & Arm* conditions, which confirmed prior work [23].

We also ran a Least Significant Number analysis on the conditions to see if the marginal significance results were due to our sample size. With the significance level $\alpha = 0.05$, the error standard deviation $\sigma = 1.09$, and the effect size $\delta = 0.3$, the minimum number of participants we needed to run was approximately 108. This is far above a typical in-person experiment, but well below the 366 online participants that Han et al. [23] recruited. This power analysis suggested that even if we had more participants, the bases of our conclusion on unexpectedness would likely remain unchanged.

In order to investigate the unexpectedness effect of the explanation conditions in each type of failure, we performed four additional one-way ANOVA tests for the Unexpectedness measure. We plotted the Unexpectedness scores of all four explanation conditions when paired with each of the failures in Figure 4.2. We found a statistically significant main effect of explanation

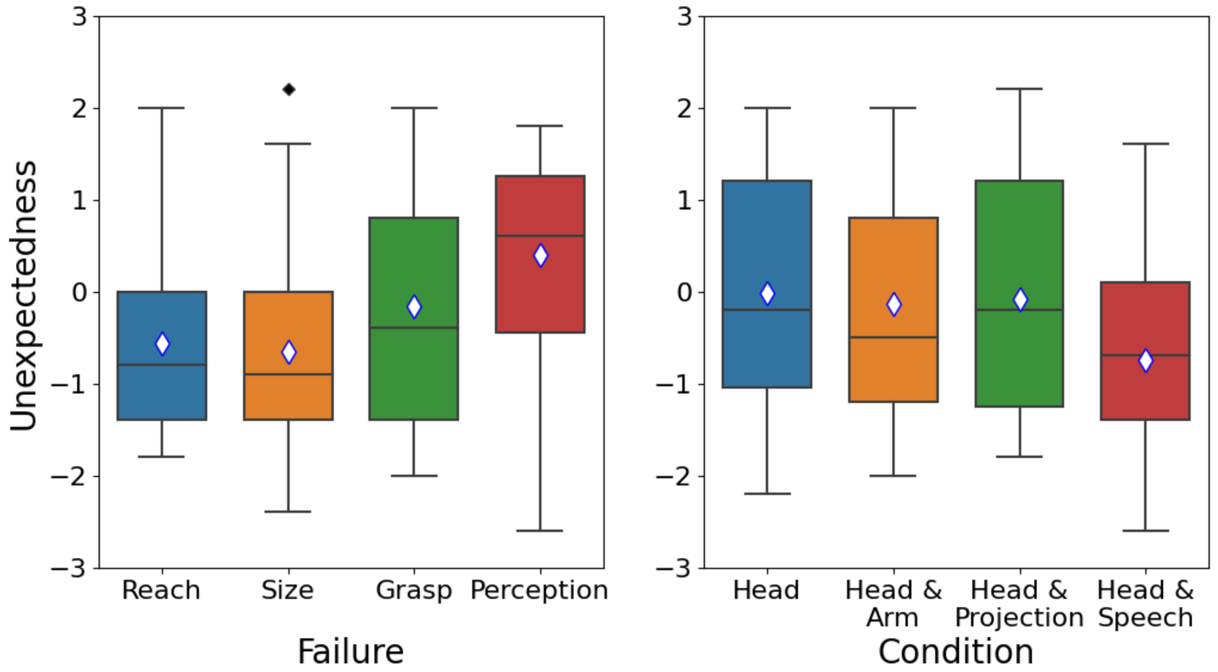

Figure 4.1: Unexpectedness scores for four types of failures (left) and four modes of explanations (right). The white diamond and black diamond icons indicate the mean scores and the outliers, respectively.

conditions when paired with *Perception* failure ($F(3, 20) = 5.75, p < 0.01$), but no significant main effects when paired with *Reach*, *Size*, and *Grasp* failures. Through post hoc pairwise comparisons using Tukey's test with Holm-Bonferroni correction ($H_0 : \mu_i = \mu_j$), we found the difference between *Head & Speech* and *Head* explanation conditions when paired with *Perception* failure approached statistical significance ($meandiff = 1.83, p = 0.07$).

4.2 Human-Robot Difference

The Human-Robot Difference item measures the similarity of the robot's behavior compared to that of a human. This supports discussion on whether robots and humans should explain their failures in the same way.

Similar to the Unexpectedness item, we performed a two-way ANOVA to examine the effect of failure types and explanation conditions on the level of human-likeness, with the distribution, shown in Figure 4.3. From the analysis, we found statistically significant main effects for failure types ($F(3, 80) = 2.82, p < 0.04$) and explanation conditions ($F(3, 80) = 10.34, p < 0.01$). However, we did not find a statistically significant interaction between the two ($F(9, 80) = 0.85, p = 0.57$). We also performed a post hoc pairwise comparison using Tukey's test with Holm-Bonferroni correction ($H_0 : \mu_i = \mu_j$). We found that there were no significant pairwise differences across the failure types. However, across the conditions, the *Head & Speech* explanation condition was found to have a significantly higher mean score than that of the *Head* condition

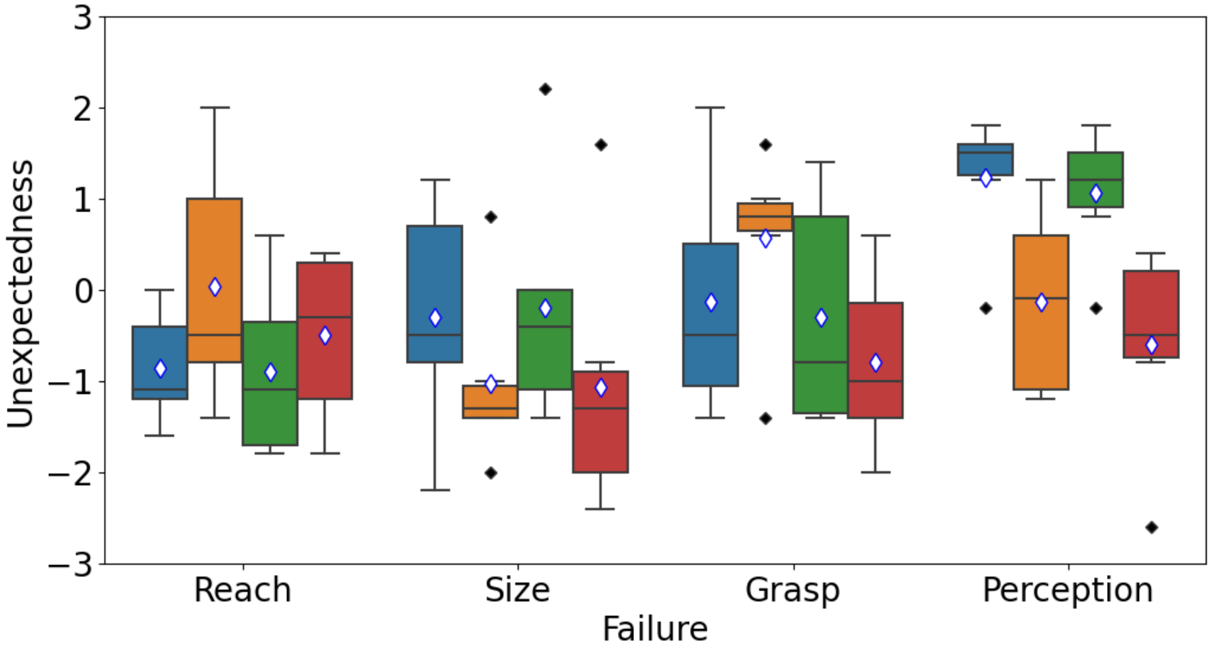

Figure 4.2: The distribution of Unexpectedness scores of four explanation conditions in each type of failure. The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. Boxes in blue, orange, green, and red represent *Head*, *Head & Arm*, *Head & Projection*, and *Head & Speech* conditions, respectively.

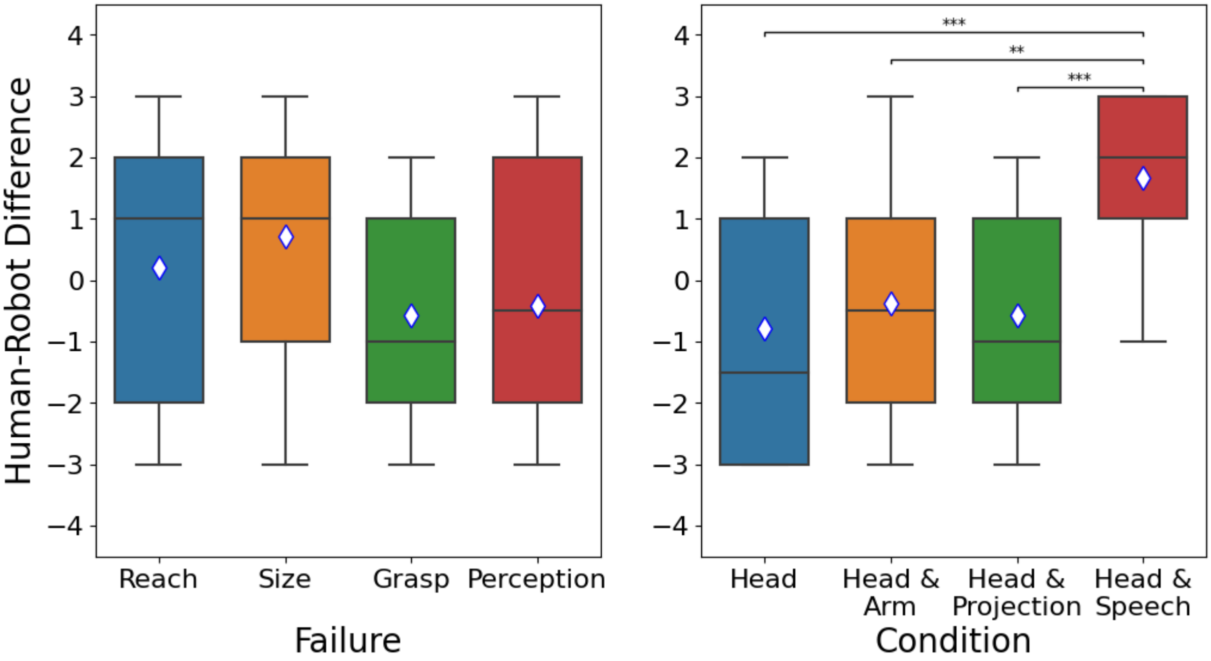

Figure 4.3: Human-Robot Difference scores in four types of failures (left) and four modes of explanations (right). The white diamond icons indicate the mean scores. $**$ represents $p-value < 0.01$, and $***$ represents $p-value < 0.001$.

($meandiff = 2.46, p < 0.001$), the ***Head & Arm*** condition ($meandiff = 2.04, p < 0.01$), and the ***Head & Projection*** condition ($meandiff = 2.25, p < 0.001$). According to participants, the robot explaining its failures with speech was more human-like than with other non-verbal cues. This finding is also consistent with Han et al. [23], as approximately half of their participants preferred the robot to say the same thing as a human does during an explanation.

Similar to the unexpectedness measure, we performed four additional ANOVA tests for the Human-Robot Difference measure. We plotted the Human-Robot Difference scores of all four explanation conditions when paired with each of the failure types, as shown in Figure 4.4. We found a statistically significant main effect for explanation conditions when paired with *Reach* failure ($F(3, 20) = 4.76, p = 0.012$), and with *Perception* failure ($F(3, 20) = 3.67, p = 0.03$). The main effect of explanation conditions approached statistical significance when paired with *Grasp* failure ($F(3, 20) = 2.74, p = 0.07$). We found no statistically significant main effect for explanation conditions when paired with *Size* failure. Through post hoc pairwise comparisons using Tukey's test with Holm-Bonferroni correction ($H_0 : \mu_i = \mu_j$), we found the difference between *Head & Speech* and *Head & Projection* explanation conditions when paired with *Reach* failure approached statistical significance ($meandiff = 3.5, p = 0.055$). However, we did not find any pairwise statistical significance among explanation conditions when paired with *Perception* and *Grasp* failures.

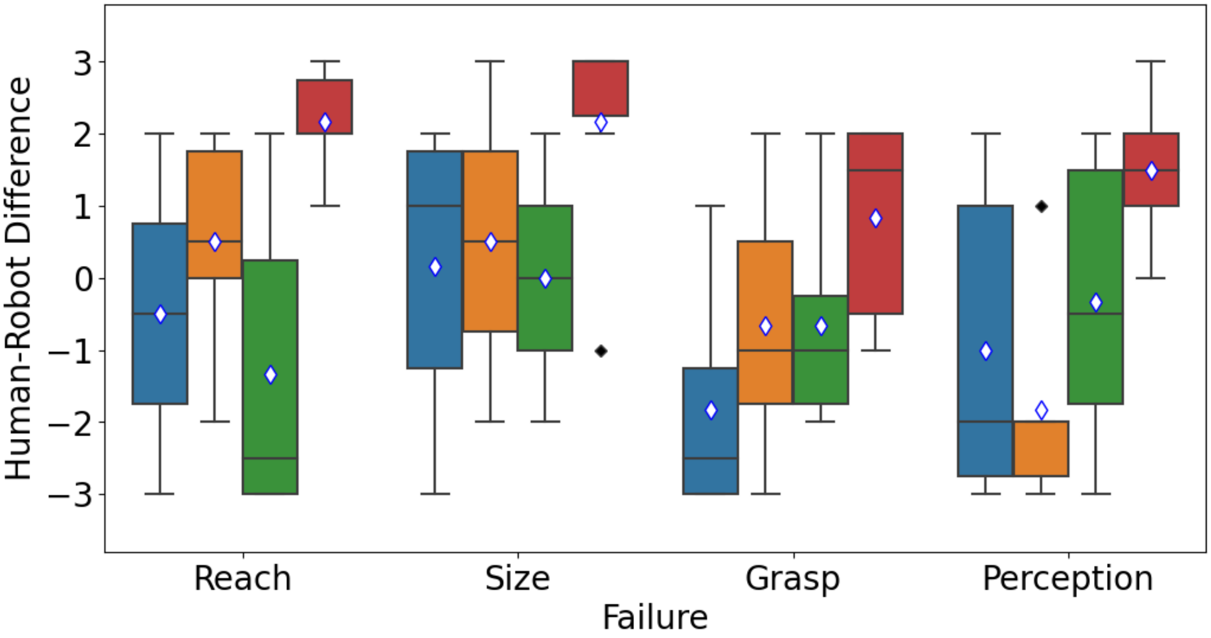

Figure 4.4: The distribution of Human-Robot Difference scores of four explanation conditions in each type of failure. The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. Boxes in blue, orange, green, and red represent *Head*, *Head & Arm*, *Head & Projection*, and *Head & Speech* conditions, respectively.

4.3 Level of Detail

The Level of Detail item measured the degree of completeness of the robot's explanations, in which (Q7) asked whether the robot should explain its behaviors in detail during the trials. The score distribution of the Level of Detail item across four types of failures and four modes of explanations are shown in Figure 4.5. We found no significant main effects or interactions for failure types or explanation conditions. In general, participants agreed that the explanations from the robot should be detailed, with the mean scores of the Level of Detail item between 0 (Neutral) and 1 (Moderately Agree).

4.4 Conciseness

The Conciseness item measured the degree of brevity of the robot's explanations, in which (Q8) asked whether the robot should concisely explain its behaviors during the trials. The score distribution of the Conciseness item across four types of failures and four modes of explanations are shown in Figure 4.6. Again, there were no significant main effects or interactions. Overall, participants reported that the robot should concisely explain its behaviors under all circumstances, with the mean scores of the Conciseness item between 1 (Moderately Agree) and 2 (Agree).

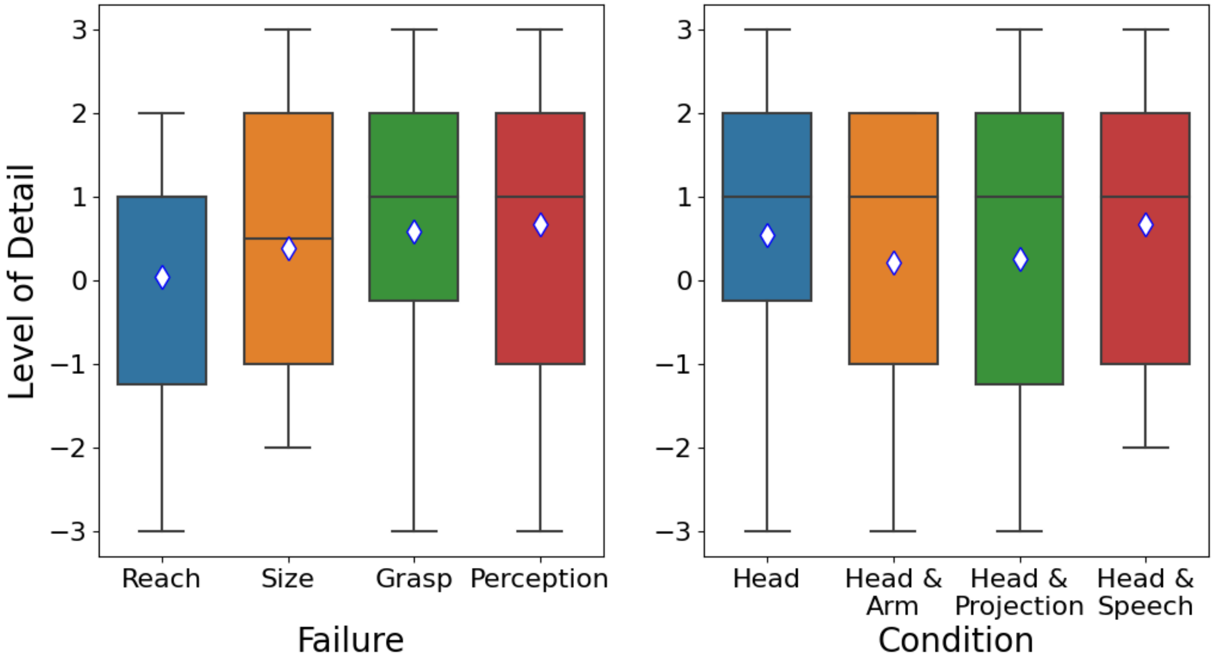

Figure 4.5: Level of Explanation Detail scores in four types of failures (left) and four modes of explanations (right). The white diamond icons indicate the mean scores.

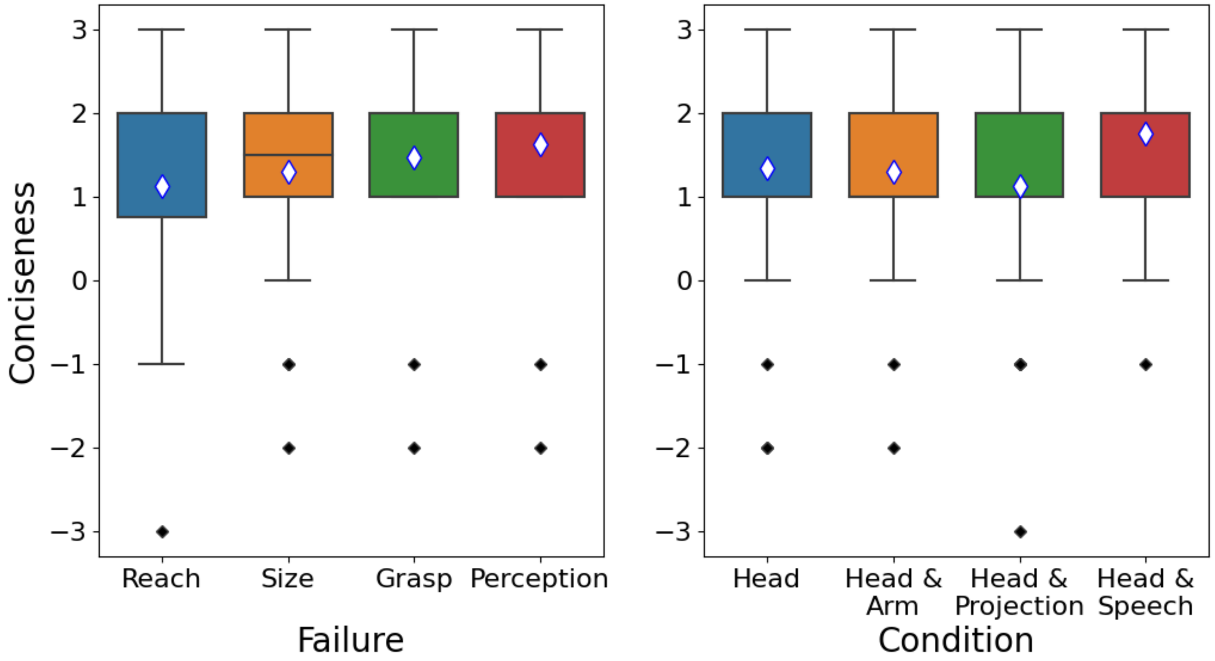

Figure 4.6: Conciseness scores in four types of failures (left) and four modes of explanations (right). The white diamond and black diamond icons indicate the mean scores and the outliers, respectively.

4.5 Post-study Questions: Trust, Competence, Need for Explanation, and Overall Perception

We present summary statistics for each question of the Post-Study Questionnaire individually to provide detailed insights on participants' general trust in the robot, the perceived competence of the robot, the need for explanation from the robot, and the overall perception of participants towards the robot. This data is illustrated in Figure 4.7, Figure 4.8, and Figure 4.9.

4.5.1 Trust

Participants, on average, found themselves to be engaging with the robot (Q9 and Q10) and safe around the robot (Q16), with the mean scores above 1 (Agree in Likert-type scale). Participants also moderately agreed that the robot was trustworthy (Q11), likable (Q13), and a potentially good teammate (Q15), with the mean scores between 0 (Neutral) and 1 (Moderately Agree). However, participants were neutral in empathy for the robot when it failed (Q12), with a mean score of 0. Participants did not feel warmth interacting with the robot (Q14) with the mean score between -1 (Moderately Disagree) and 0 (Neutral). Overall, we found that the robot was trustworthy to the participants during interaction and the pick and place task.

4.5.2 Competence

Participants generally found the robot to be competent (Q17-19, Figure 4.8), with the mean scores of all questions between 0 (Neutral) and 2 (Agree). They found that the robot's movements were clear, lifelike, and important in helping them understand the ability of the robot.

4.5.3 Need for Explanation

As mentioned, Questions 21-23 are not based on Likert-type items, so we reported the responses separately from the others. Participants preferred the robot to get their attention before starting to explain its behavior (Q21), with 79% of them agreeing. To get their attention (Q22), 33% of participants preferred the robot to look at them, 75% of participants preferred the robot to raise its volume or play some sounds to alert the participants, and others preferred the robot to perform some actions like waiving the arm. This result is consistent with our finding of participants' preference in the Head & Speech explanation condition in our previous section. In terms of explanation timing (Q23), 79% of participants preferred the robot to explain its behavior whenever something unexpected happens, 17% of participants preferred the robot to explain its behavior at the end, and the rest preferred the timing to be before something unexpected happens. This finding is consistent with [23] as they claimed approximately half of their participants preferred *in situ* explanations, and only 18% preferred explanation before something unexpected happens.

As seen in Figure 4.9, we also found that participants strongly preferred the robot to explain its behavior (Q20), with a mean score above 2 (Agree). Moreover, participants acknowledged the value of signaling failure through gestures (Q24), but preferred out loud verbal announcements

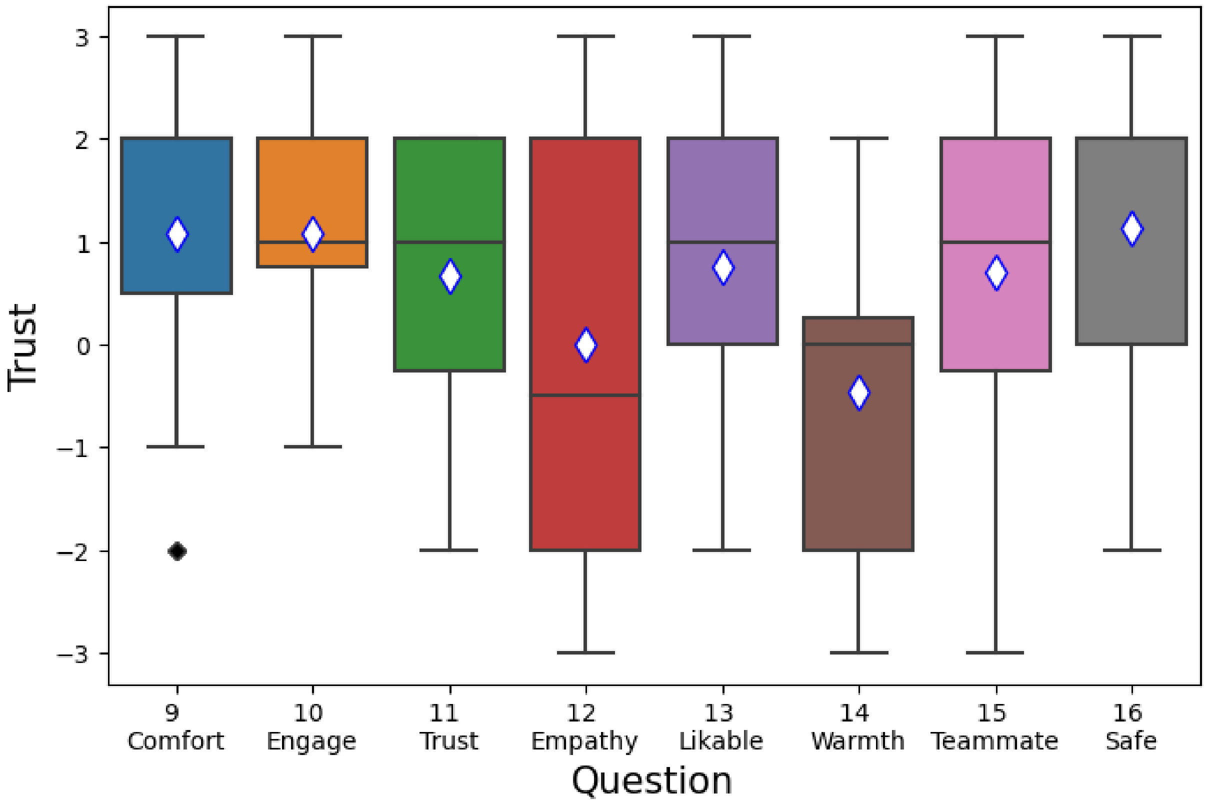

Figure 4.7: Distribution of scores in trust-related questions. The white diamond and black diamond icons indicate the mean scores and the outliers, respectively.

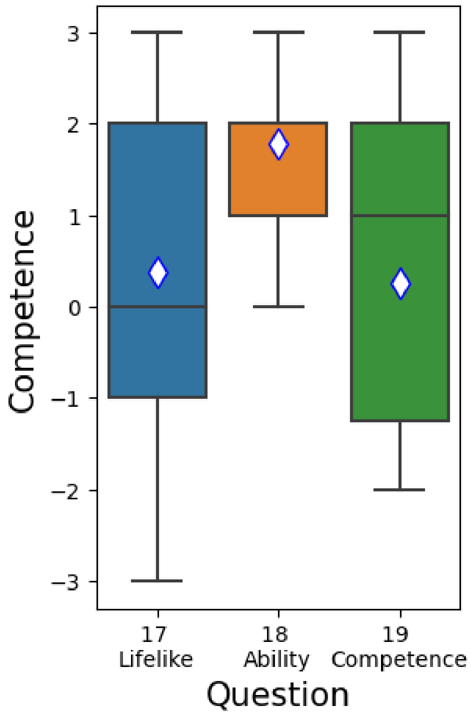

Figure 4.8: Distribution of scores in competence-related questions. The white diamond icons indicate the mean scores.

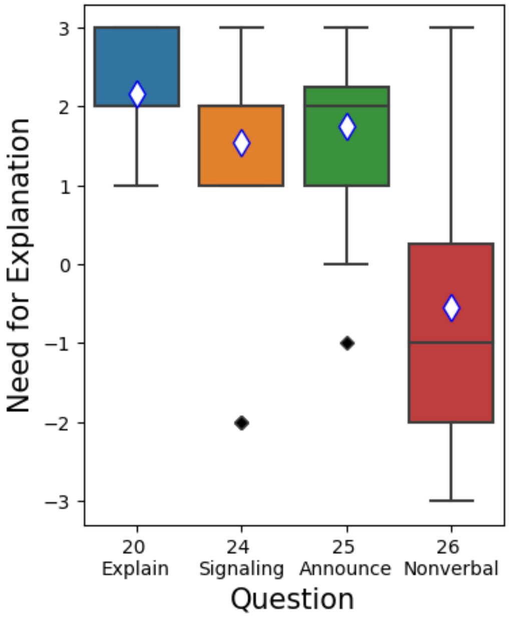

Figure 4.9: Distribution of scores in need-for-explanation-related questions. The white diamond and black diamond icons indicate the mean scores and the outliers, respectively. Questions 21, 22, and 23 are discussed separately as they are not based on Likert-type items.

of failure from the robot (Q25), with the mean scores being between 1 (Moderately Agree) and 2 (Agree). Finally, participants did not prefer non-verbal explanations from the robot (Q26), with a mean score being between -1 (Moderately Disagree) and 0 (Neutral). This has face validity since there were statistical differences between Head & Speech condition and other non-verbal and visual cues in the Human-Robot Difference item.

4.5.4 Overall Perception

Participants were asked an open-ended question at the end of the Post-Study Questionnaire about their overall perception of the robot (Q27). Some participants preferred the robot's voice to be louder and more human-like. They also liked the fact that the robot focused visually on its target, as its head always looked at the block it was planning to pick up and the tray. Some mentioned adding a signal before the explanation to alert humans about its intention to explain its behavior. A few participants commented that they did not understand the projection mode of explanation as they only saw an image, not an explanation or a solution to the failure. They also thought that the robot's movements were natural for a robot, but its ability to rotate its joints 360 degrees might be perceived as not natural for humans.

Chapter 5

Discussion

Research on the consistency of results between online and in-person studies in human-robot interaction has been sparse. The findings among study replications across different robot platforms have also been inconsistent [49]. Thus, we designed and carried out the in-person experiment under the assumption that our results could be different from those of the online experiment conducted by Han et al. [23]. We found that replicated conditions had consistent findings between their online study and our in-person study. Our results revealed that there was minimal difference in effects on participants of both *Head* and *Head & Arm* conditions, with both resulting in Neutral scores on the level of unexpectedness. As seen in [23], participants also preferred the robot to get their attention before explaining its behavior, preferably with an *in situ* explanation. Moreover, participants also acknowledged the importance of robots signaling their failures. Therefore, our study successfully replicated and reinforced findings from the prior study.

Along with our successful replication of prior work results, we found that the introduction of the Projection component did not lead to additional benefits in participants' perceptions of the robot. This finding is related, but not identical, to Han et al. [22], which claimed that standalone projection markers can worsen participants' causal inference performance, as seen by the fact that in their experiment, only half of the participants correctly inferred the missing information about the object picking task when only projection marker was used. In contrast, our projection condition was comparable to the other non-verbal conditions.

Findings from [23] suggested that speech is preferred to gesture for robot explanations, which our findings confirmed. Moreover, participants rated that the *Head & Speech* explanation condition is the most human-like among all explanation conditions to explain robot failures.

Findings from our post-study questionnaire provided further evidence that speech was a preferred component of failure explanation, as most participants wanted the robot to announce its failure out loud and raise its volume to alert them before giving explanations. Those findings, along with the positive perceptions of the robot's trustworthiness and competence, agree with prior work suggesting human preference for verbal over nonverbal communication. For example, Nikolaidis et al. [35] reported that short sentences as verbal commands were more effective than non-verbal actions and promoted trust between a human and a robot in a collaborative table-carrying task, and Grigore et al. [20] revealed that humans perceived robots as more friendly and socially present with speech-based communication compared to action-based communication. Furthermore, Maggioni et al. [33] indicated that robots capable of verbal interaction are perceived as “more human” and can eliminate the difference between human- human interaction and human-robot interaction, all with short and easy-to-understand utterances.

Trustworthiness, competence, empathy, and warmth go hand in hand with each other as one or more attributes can positively influence others in human-robot interactions [11]. However, from our study, participants found the robot to be trustworthy and competent, but they were neutral about empathy when the robot failed and did not feel warmth when interacting with it. This neutral feeling of empathy could be due to our pre-experiment briefing that the robot may fail during the experiment. Moreover, the lack of feeling of warmth could partly be due to participants only observing the robot instead of directly interacting with it during the task, as prior work suggested that warmth is enhanced through physical interactions [40].

Chapter 6

Limitations

Among humans, eye gaze can be a powerful action that promotes inter-human interaction. Eye gaze and facial expressions have been proven to be effective in communicating intentions, strengthening human-robot collaboration, and improving engagement [26, 46]. Moreover, robot facial expressions have been studied and used to aid understanding and engagement of humans [18, 34, 38]. Due to the limitations of our hardware, we were not able to incorporate eye gaze and facial expressions into our modes of explanations. Thus, we recommend that eye gaze and facial expressions be incorporated into future studies on human perceptions of robot failures and explanations.

We conducted our experiment in-person in a university lab, which inherently limits the sample size. Han et al. [23] recruited 366 online participants compared to our 24 in-person participants. Our sample size was near 30 (Central Limit Theorem, [29]) and our power analyses suggested that increasing the sample size would be unlikely to change key findings. Having said this, there is still a chance that some of our significance tests would be different with a larger sample.

Next, there is value in multi-modal verbal/non-verbal communication in human-robot interaction [3, 27]. Therefore, there should be future studies combining verbal and non-verbal modes of failure explanation. This would help explain the impact of multi-modal communication on perceptions of the level of unexpectedness, human-likeness, and the level of detail in explanation.

Finally, due to our experiment design, we were only able to measure absolute trust and competence as opposed to those measures after participants observed different failures or explanations. The absolute trust and competence measures are dependent on a number of unreplicable factors such as task sequence, experiment design, or the specific robot itself. Measuring trust and competence as the participants were exposed to robot behaviors would be helpful in modeling such measurements and designing future interactions.

Chapter 7

Conclusion

We extended prior work on non-verbal motion cues for robot explanation from an online study to an in-person study. Our findings suggested that head motions and paired head and arm motions produced comparable effects for failure explanation from robots. Our results confirmed the prior findings and demonstrated consistency between online and in-person studies. We also gathered new data on two additional methods of explanation, namely Projection and Speech, and found that speech was the most preferred mode of failure explanation in terms of human-likeness. However, the use of Projection as a component of explanation performed similarly to the status quo of head motions and paired head and arm motions.

Bibliography

[1] Neziha Akalin, Annica Kristoffersson, and Amy Loutfi. Evaluating the sense of safety and security in human–robot interaction with older people. *Social robots: Technological, societal and ethical aspects of human-robot interaction*, pages 237–264, 2019. 2.2, 2.2.3

[2] Neziha Akalin, Annica Kristoffersson, and Amy Loutfi. Do you feel safe with your robot? factors influencing perceived safety in human-robot interaction based on subjective and objective measures. *International journal of human-computer studies*, 158:102744, 2022. 2.2, 2.2.3

[3] Amir Aly and Adriana Tapus. Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction. *Autonomous Robots*, 40:193–209, 2016. 6

[4] Jakob Ambsdorf, Alina Munir, Yiyao Wei, Klaas Degkwitz, Harm Matthias Harms, Susanne Stannek, Kyra Ahrens, Dennis Becker, Erik Strahl, Tom Weber, et al. Explain yourself! effects of explanations in human-robot interaction. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 393–400. IEEE, 2022. 1

[5] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019. 2.2, 2.2.1

[6] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*, pages 708–713. IEEE, 2005. 2.1.2

[7] Xuan Cao, Alvika Gautam, Tim Whiting, Skyler Smith, Michael A Goodrich, and Jacob W Crandall. Robot proficiency self-assessment using assumption-alignment tracking. *IEEE Transactions on Robotics*, 2023. 1, 2.1.2, 3.3

[8] Jennifer Carlson, Robin R Murphy, and Andrew Nelson. Follow-up analysis of mobile robot failures. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 5, pages 4987–4994. IEEE, 2004. 2.1.1, 2.1.2

[9] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions*

on Affective Computing, 10(4):484–497, 2017. 2.2, 2.2.4

[10] Sungwoo Choi, Anna S Mattila, and Lisa E Bolton. To err is human (-oid): how do consumers react to robot service failure and recovery? *Journal of Service Research*, 24(3): 354–371, 2021. 2.1.2, 2.2, 2.2.2

[11] Lara Christoforakos, Alessio Gallucci, Tinatini Surmava-Große, Daniel Ullrich, and Sarah Diefenbach. Can robots earn our trust the same way humans do? a systematic exploration of competence, warmth, and anthropomorphism as determinants of trust development in hri. *Frontiers in Robotics and AI*, 8:640444, 2021. 5

[12] Maartje MA De Graaf and Bertram F Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, 2017. 2.2, 2.2.1, 2.3

[13] Ronan de Kervenoael, Rajibul Hasan, Alexandre Schwob, and Edwin Goh. Leveraging human-robot interaction in hospitality services: Incorporating the role of perceived value, empathy, and information sharing into visitors' intentions to use social robots. *Tourism Management*, 78:104042, 2020. 2.2, 2.2.4

[14] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 251–258. IEEE, 2013. 2.2, 2.2.1

[15] Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, Yixin Zhu, Ying Nian Wu, Hongjing Lu, and Song-Chun Zhu. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, 4(37):eaay4663, 2019. 1

[16] Cliff Fitzgerald. Developing baxter. In *2013 IEEE conference on technologies for practical robot applications (TePRA)*, pages 1–6. IEEE, 2013. 3.1

[17] Chris Frith. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3453–3458, 2009. 2.3

[18] Shuzhi Sam Ge, Chen Wang, and Chang Chieh Hang. Facial expression imitation in human robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 213–218. IEEE, 2008. 2.1.2, 2.3, 6

[19] David A Grant. The latin square principle in the design and analysis of psychological experiments. *Psychological bulletin*, 45(5):427, 1948. 3.5

[20] Elena Corina Grigore, Andre Pereira, Ian Zhou, David Wang, and Brian Scassellati. Talk to me: Verbal communication improves perceptions of friendship and social presence in human-robot interaction. In *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, Proceedings 16*, pages 51–63. Springer, 2016. 5

[21] Joanna Hall, Terry Tritton, Angela Rowe, Anthony Pipe, Chris Melhuish, and Ute Leonards. Perception of own and robot engagement in human–robot interactions and their dependence on robotics knowledge. *Robotics and Autonomous Systems*, 62(3):392–399, 2014. 2.2, 2.2.4

[22] Zhao Han and Holly Yanco. Communicating missing causal information to explain a robot's

past behavior. *ACM Transactions on Human-Robot Interaction*, 12(1):1–45, 2023. 1, 2.1.2, 2.3, 3.3, 5

[23] Zhao Han, Elizabeth Phillips, and Holly A Yanco. The need for verbal robot explanations and how people would like a robot to explain itself. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(4):1–42, 2021. (document), 1, 2.1.1, 2.1.2, 2.3, 3, 3.1, 3.3, 3.3.1, 3.3.2, 3.4, 3.1, 3.2, 4.1, 4.2, 4.5.3, 5, 6

[24] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979. 4.1

[25] Shanee Honig, Alon Bartal, Yisrael Parmet, and Tal Oron-Gilad. Using online customer reviews to classify, predict, and learn about domestic robot failures. *International Journal of Social Robotics*, pages 1–26, 2022. 2.1.1, 2.1.2

[26] Alisa Kalegina, Grace Schroeder, Aidan Allchin, Keara Berlin, and Maya Cakmak. Characterizing the design space of rendered robot faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 96–104, 2018. 2.1.2, 2.3, 6

[27] Spencer D Kelly, Dale J Barr, R Breckinridge Church, and Katheryn Lynch. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language*, 40(4):577–592, 1999. 2.1.2, 6

[28] Parag Khanna, Elmira Yadollahi, Mårten Björkman, Iolanda Leite, and Christian Smith. Effects of explanation strategies to resolve failures in human-robot collaboration. *arXiv preprint arXiv:2309.10127*, 2023. 2.1.2, 2.3

[29] Sang Gyu Kwak and Jong Hae Kim. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144–156, 2017. 6

[30] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 203–210. IEEE, 2010. 2.1.2

[31] Dingjun Li, PL Patrick Rau, and Ye Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2:175–186, 2010. 2.2.1

[32] Joseph B Lyons, Izz aldin Hamdan, and Thy Q Vo. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior*, 138:107473, 2023. 1

[33] Mario A Maggioni and Domenico Rossignoli. If it looks like a human and speaks like a human... communication and cooperation in strategic human–robot interactions. *Journal of Behavioral and Experimental Economics*, 104:102011, 2023. 5

[34] Cecilia G Morales, Elizabeth J Carter, Xiang Zhi Tan, and Aaron Steinfeld. Interaction needs and opportunities for failing robots. In *Proceedings of the 2019 on designing interactive systems conference*, pages 659–670, 2019. 2.1.2, 2.3, 6

[35] Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha Srinivasa. Planning with verbal communication for human-robot collaboration. *ACM Transactions on Human-Robot Interaction (THRI)*, 7(3):1–21, 2018. 5

[36] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F Malle. What is human-

like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pages 105–113, 2018. 2.2.1

[37] Christina Regenbogen, Daniel A Schneider, Raquel E Gur, Frank Schneider, Ute Habel, and Thilo Kellermann. Multimodal human communication—targeting facial expressions, speech content and prosody. *Neuroimage*, 60(4):2346–2356, 2012. 2.3

[38] Mauricio E Reyes, Ivan V Meza, and Luis A Pineda. Robotics facial expression of anger in collaborative human–robot interaction. *International Journal of Advanced Robotic Systems*, 16(1):1729881418817972, 2019. 2.1.2, 2.3, 6

[39] Stephanie Rosenthal, Sai P Selvaraj, and Manuela M Veloso. Verbalization: Narration of autonomous robot experience. In *IJCAI*, volume 16, pages 862–868, 2016. 2.1.2

[40] Marcus M Scheunemann, Raymond H Cuijpers, and Christoph Salge. Warmth and competence to predict human preference of robot behavior in physical human-robot interaction. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1340–1347. IEEE, 2020. 2.2, 2.2.2, 5

[41] Mariah L Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C Gombolay. Four years in review: Statistical practices of likert scales in human-robot interaction studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 43–52, 2020. 3.4

[42] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. "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. IEEE, 2019. 2.1.2

[43] Jinglin Shen and Nicholas Gans. Robot-to-human feedback and automatic object grasping using an rgb-d camera–projector system. *Robotica*, 36(2):241–260, 2018. 2.1.2

[44] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian 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*, pages 178–186, 2018. 2.1.2

[45] Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 69–76, 2011. 2.1.2

[46] Peter Tisnikar, Lennart Wachowiak, Gerard Canal, Andrew Coles, Matteo Leonetti, and Oya Celiktutan. Towards autonomous collaborative robots that adapt and explain. In *IEEE ICRA 2022 Workshop on Prediction and Anticipation Reasoning in Human Robot Interaction*, 2022. 2.1.2, 6

[47] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, pages 3–12, 2020. 2.2, 2.2.1

[48] Sylvaine Tuncer, Christian Licoppe, Paul Luff, and Christian Heath. Recipient design in human–robot interaction: the emergent assessment of a robot's competence. *AI & SOCIETY*, pages 1–16, 2023. 2.2, 2.2.2

[49] Daniel Ullman, Salomi Aladia, and Bertram F Malle. Challenges and opportunities for replication science in hri: A case study in human-robot trust. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, pages 110–118, 2021. 5

[50] Christian Vogel, Maik Poggendorf, Christoph Walter, and Norbert Elkmann. Towards safe physical human-robot collaboration: A projection-based safety system. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3355–3360. IEEE, 2011. 2.1.2

[51] Christian Vogel, Christoph Walter, and Norbert Elkmann. Exploring the possibilities of supporting robot-assisted work places using a projection-based sensor system. In *2012 IEEE International Symposium on Robotic and Sensors Environments Proceedings*, pages 67–72. IEEE, 2012. 2.1.2

[52] Astrid Weiss, Regina Bernhaupt, Manfred Tscheligi, and Eiichi Yoshida. Addressing user experience and societal impact in a user study with a humanoid robot. In *AISB2009: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction (Edinburgh, 8-9 April 2009), SSAISB*, pages 150–157, 2009. 2.2, 2.2.3

[53] Melonee Wise, Michael Ferguson, Derek King, Eric Diehr, and David Dymesich. Fetch and freight: Standard platforms for service robot applications. In *Workshop on autonomous mobile service robots*, pages 1–6, 2016. 2.2.1, 3.1

[54] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 408–416, 2017. 2.2, 2.2.1, 2.3