Human Perception of Robot Failure and Explanation During a Pick-and-Place Task

Huy Quyen Ngo[1], **Elizabeth Carter**[1], **Aaron Steinfeld**[1]

[1]The Robotics Institute, Carnegie Mellon University
huyquyen@andrew.cmu.edu, ejcarter@andrew.cmu.edu, steinfeld@cmu.edu

Abstract

In recent years, researchers have extensively used non-verbal gestures, such as head and arm movements, to express a 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, within-subjects experiment with 24 participants, we studied four modes of explanations across four types of failures. Some of these were chosen to mimic combinations from prior work in order to both extend and replicate past findings by the community. We found that speech explanations were preferred to non-verbal and visual cues in terms of similarity to humans. Additionally, projected images had a comparable effect on explanation as other non-verbal modules. We also found consistent results with a prior online study.

Introduction

Humans frequently provide explanations 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 (Edmonds et al. 2019) and human-likeness (Ambsdorf et al. 2022).

In this work, we extended a prior study on robot explanation in a cup-handover task (Han, Phillips, and Yanco 2021). The study examined failure conditions of a cup that was out of reach from a Baxter robot. 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 worked well as they 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 only considered *Look* as the primary head movement during the explanation. We also had our robot explain its failures *in situ*. In addition, we introduced two new explanation components, namely Image Projection and Speech, inspired by later work by the same team (Han and Yanco 2023). Mixing prior and new interactions supports both extension and replication of past work.

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 (Han and Yanco 2023; Cao et al. 2023). These 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). Moving the study to in-person also allows us to see if results from the replicated combinations are consistent with the prior online study (Han, Phillips, and Yanco 2021).

We designed a 15-item questionnaire, partly adapted from Han, Phillips, and Yanco (2021), to measure some key aspects of human-robot interaction: Unexpectedness, Human-Robot Difference, Level of Detail, Conciseness, 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 having 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 have similar effect on human perception compared to non-verbal gestures; and
3. Evidence for a prior conjecture about speech being preferred for explaining robot failure and performance.

Related Work

Robot Failure and Explanation

In the motivating prior work, Han, Phillips, and Yanco (2021) studied robot explanation during a cup handover task in which the cup was out of reach from the robot arm. To explain its behavior, the robot used 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. Moreover, Han, Phillips, and Yanco (2021) conducted their experiment online. 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 and Yanco (2023) used verbal and projection indicators, coupled with head and arm motion replay, to communicate past causal information related to tasks. Cao et al. (2023) 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, Selvaraj, and Veloso (2016) studied the effectiveness of the verbal modality in parallel with visual modality during robot operation. Moreover, verbal explanation has been proven to be effective in failure situations (Choi, Mattila, and Bolton 2021; Khanna et al. 2023). Thus, we studied 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 and Yanco (2023) demonstrated the effectiveness of projection in revealing task-related information. Projections can indicate boundaries around robots (Vogel et al. 2011) (e.g., maximum robot reach), display information about the robot (Vogel, Walter, and Elkmann 2012) (e.g., maximum gripper opening), mark locations (Shen and Gans 2018) (e.g., a red X for a failure location), and communicate misperceptions (Han and Yanco 2023) (e.g., hallucinated objects).

Robotic systems can experience multiple types of failures, either from the robot software and hardware, or from surrounding environments. Honig et al. (2022) discussed a taxonomy of human-robot failures in domestic robots that are most frequently seen by customers. Carlson, Murphy, and Nelson (2004) classified in-depth physical failures in the end effector of the robot. Thus, along with the Out Of Reach failure from (Han, Phillips, and Yanco 2021), 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).

Human Perceptions Towards Robots

Trust in robots and autonomy has been extensively studied to promote effective human-robot interaction. Anjomshoae et al. (2019) 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 (De Graaf and Malle 2017). Moreover, Yang et al. (2017) indicated that trust can be measured through a series of interactions with automation systems. Tolmeijer et al. (2020) suggested that offering explanations can help mitigate failure and repair trust.

Humans usually take fellow humans' competence for granted during interaction, whether such interaction is verbal or non-verbal (Tuncer et al. 2023). For humans and robots to collaborate, humans need to learn the competence of robots through prolonged interaction. Scheunemann, Cuijpers, and Salge (2020) found that humans prefer to physically interact with robots that are perceived as warm and competent. Choi, Mattila, and Bolton (2021) measured competence in robots 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 failures.

Method

Inspired by the prior work of Han, Phillips, and Yanco (2021), we designed an in-person, within-subjects experiment to collect more reliable responses on human perception of robot failure and explanation.

Robot Description

While prior work (Han, Phillips, and Yanco 2021) used a Baxter robot (Fitzgerald 2013), our study used a Fetch robot (Wise et al. 2016), 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 System (ROS) was used to control the robot.

Experiment Setup

The arrangement of the table for the pick-and-place task is shown in Figure 1. There were seven blocks of different sizes, shapes, and colors scattered on the table and the robot's task was to pick up the blocks and place them into the tray. Among those blocks, some were designed to be picked up by the robot gripper, while others were there as decoys. For the *Head & Projection* explanation condition, a ceiling-mounted external projector was used to project images onto the table. Participants were only informed about the existence of the projection module, without knowing what the projections looked like.

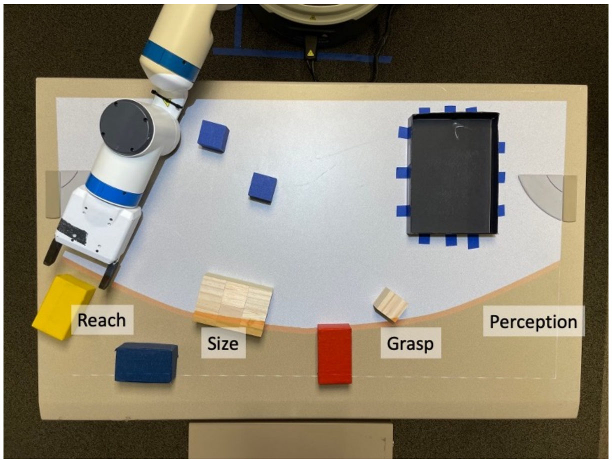

Figure 1: Locations of the successful blocks (blue) and the four failures (labeled). The arm position and projected white area with red arc were used for *Head & Projection* during a *Reach* failure.

Failure and Explanation

Prior work by Han, Phillips, and Yanco (2021) 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 Han and Yanco (2023), we incorporated Projection and Speech as two new components in the explanations, and added three new failure types: *Size*, *Grasp*, and *Perception*.

Explanation Conditions

Head: We removed the head shakes from Han, Phillips, and Yanco (2021). 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 Han, Phillips, and Yanco (2021) 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 that 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 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."

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 failure described in Han, Phillips, and Yanco (2021).

Size: The width of the block was larger than the maximum gripper opening, so 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.

Measures

We prepared a post-trial survey based on questions used in Han, Phillips, and Yanco (2021) to measure participants' perceptions of robot failure and explanation, as shown in Table 1. With the aim of replicating 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 primarily used the 7-point Likert-type item (Schrum et al. 2020). Each Likert-type item is coded as -3 (Strongly Disagree), -2 (Disagree), -1 (Moderately Disagree), 0 (Neutral), 1 (Moderately Agree), 2 (Agree), and 3 (Strongly Agree).

In addition to the questions in Table 1, we designed a post-study questionnaire (Table 2) to gather information about the participants' 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. Questions were Likert-type items except for questions 10, 11, and 12, which were multiple-choice questions with specific options. The choices for Question 10 are "Yes" and "No". The choices for Question 11 are "It should look at me", "It should raise its volume", and "Other (Please elaborate)". The choices of Question 12 are "At the end", "Whenever something unexpected happens", "Before something unexpected happens", and "Other (Please elaborate)".

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 1: Post-Trial Questions. * indicates questions adapted from Han, Phillips, and Yanco (2021).

Need for Explanation
9. I wanted the robot to explain its behavior.*
10. Do you think it is important for the robot to get your attention before starting to explain its behavior?*
11. How should the robot get your attention before starting to explain its behavior?*
12. When would be the best time for the robot to explain its behavior?*
13. A robot signaling failure through its movements is important.
14. I want robots to announce failure out loud.
15. I prefer non-verbal actions from robots when they fail.

Table 2: Post-Study Questionnaire. * items were adapted from Han, Phillips, and Yanco (2021).

User Study Design

We ran each participant across four modes of explanations. To address practice and ordering effects, the types of failures and modes of explanations were each ordered using a four-way Latin Square (Grant 1948). 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.

Participants

In keeping with best practices, we sought gender balance. The 24 participants included 12 women and 12 men. Participants were of varying age range from 19 to 82 years, with the mean age of 34.5. Participants' experience with robots ranged from no exposure to years of experience (building robots at school, having robot vacuums, etc.). Participants were recruited through an online platform for human behavioral studies, flyers, and word of mouth.

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 details and information about their experience with robots. All responses were recorded 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 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) farther from the robot. 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: Upon 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.

Survey: 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 their four trials, participants were asked to respond to a post-study questionnaire about their need for explanation from robots. The study took 45 minutes, and they received \$10 compensation for their time. This research was approved by our university's Institutional Review Board.

Results

We analyzed Unexpectedness, Human-Robot Difference, Level of Detail, and Conciseness using two-way ANOVAs. For Need for Explanation, we summarized our findings.

Unexpectedness

The Unexpectedness item measured the level of unpredictability of the robot's behavior. The item had 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 effects of failure types and explanation conditions on the level of unexpectedness, with the distribution shown in Figure 2. From the analysis, we found a statistically significant main effect for failure types ($F(3, 80) = 4.51, p < 0.01$). Post

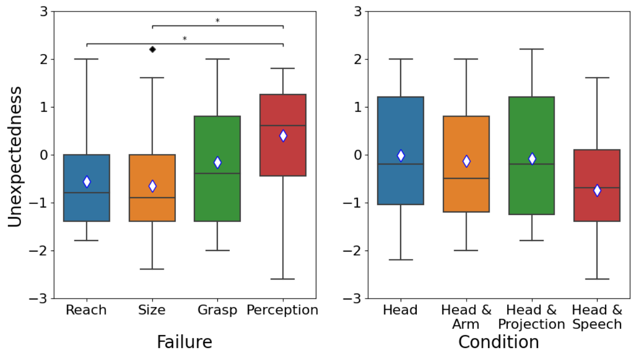

Figure 2: Unexpectedness scores for four types of failures (left) and modes of explanations (right). The white and black diamonds indicate mean scores and outliers, respectively. $*$ represents $p - value < 0.05$.

hoc pairwise comparisons using Tukey's HSD revealed that there were significant pairwise differences across the failure types. The *Perception* failure was found to have a significantly higher mean score (e.g., more unexpected) than that of the *Reach* failure ($meandiff = 0.95, p - adjusted = 0.028$) and *Size* failure ($meandiff = 1.04, p - adjusted = 0.013$). According to participants, *Perception* failure was significantly more unexpected than *Reach* and *Size* failures.

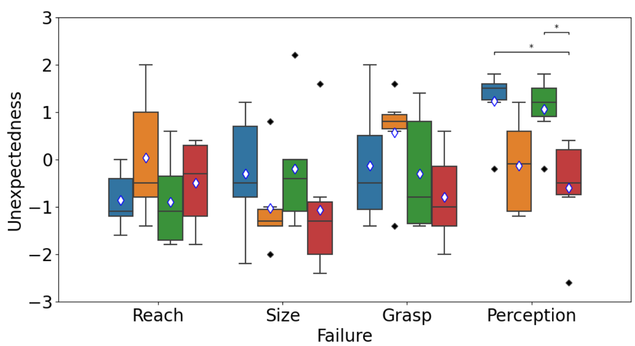

Figure 3: The distribution of Unexpectedness scores of four explanation conditions in each type of failure. The white and black diamonds indicate mean scores and outliers, respectively. Boxes in blue, orange, green, and red represent *Head*, *Head & Arm*, *Head & Projection*, and *Head & Speech* conditions, respectively. $*$ represents $p - value < 0.05$.

To investigate the unexpectedness of the explanation conditions, we performed four additional one-way ANOVAs for the Unexpectedness measure. We plotted the Unexpectedness scores of four explanation conditions when paired with each of the failures in Figure 3. We found a statistically significant main effect of explanation conditions when paired with *Perception* failure ($F(3, 20) = 5.75, p < 0.01$). Post hoc pairwise comparisons using Tukey's HSD revealed the significant difference between *Head & Speech* and *Head* ($meandiff = 1.83, p - adjusted = 0.012$) explanation conditions and between *Head & Speech* and *Head & Projection* explanation conditions ($meandiff = 1.67, p - adjusted = 0.024$) when paired with *Perception* failure .

Human-Robot Difference

The Human-Robot Difference item asks whether robots and humans should explain their failures in the same way.

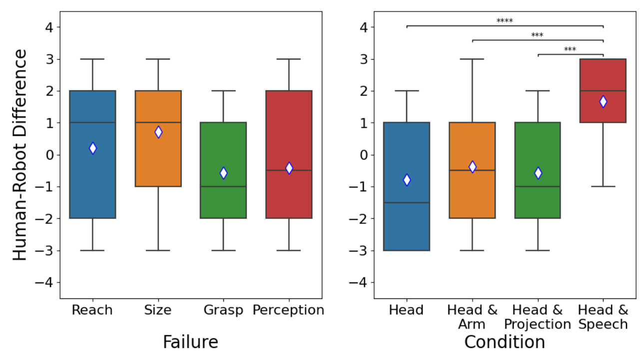

Figure 4: Human-Robot Difference scores in four types of failures (left) and four modes of explanations (right). The white diamonds indicate mean scores. $***$ represents $p - value < 0.001$, and $****$ represents $p - value < 0.0001$.

We performed a two-way ANOVA to examine the effects of failure types and explanation conditions on the level of human-likeness, with the distribution shown in Figure 4. From the analysis, we found statistically significant main effects for failure types ($F(3, 80) = 2.82, p = 0.044$) and explanation conditions ($F(3, 80) = 10.34, p < 0.0001$). However, we did not find a statistically significant interaction between the two ($F(9, 80) = 0.85, p = 0.57$). Post hoc pairwise comparisons using Tukey's HSD revealed 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 ($meandiff = 2.46, p < 0.0001$), the *Head & Arm* condition ($meandiff = 2.04, p < 0.001$), and the *Head & Projection* condition ($meandiff = 2.25, p < 0.001$). According to the results, participants found the robot explaining failures with speech to be more human-like than other nonverbal cues.

We performed four additional ANOVA tests for the Human-Robot Difference measure. We plotted the Human-Robot Difference scores of all four conditions when paired with each of the failure types, as shown in Figure 5. 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 ($F(3, 20) = 3.67, p = 0.03$). Post hoc pairwise comparisons using Tukey's HSD revealed a significant difference between *Head & Speech* and *Head & Projection* ($meandiff = 3.5, p - adjusted < 0.01$) explanation conditions when paired with *Reach* failure. Moreover, we also found a significant difference between *Head & Speech* and *Head & Arm* explanation conditions ($meandiff = 3.33, p - adjusted = 0.022$) when paired with *Perception* failure.

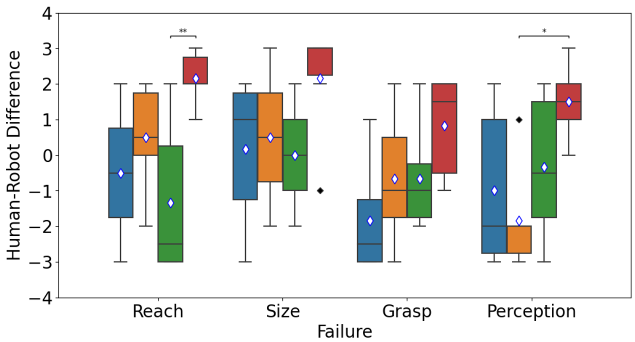

Figure 5: The distribution of Human-Robot Difference scores of four explanation conditions in each type of failure. The white and black diamonds indicate mean scores and outliers, respectively. Boxes in blue, orange, green, and red represent *Head*, *Head & Arm*, *Head & Projection*, and *Head & Speech* conditions, respectively. $*$ represents $p - value < 0.05$, and $**$ represents $p - value < 0.01$.

Level of Detail and Conciseness

The Level of Detail and Conciseness items measured the degree of completeness (Q7) and the degree of brevity (Q8) of the robot's explanations. We found no significant main effects or interactions for failure types or explanation conditions in both Level of Detail and Conciseness. Participants agreed the explanations from the robot should be detailed, with the mean scores of Q7 being between 0 (Neutral) and 1 (Moderately Agree), but the robot should concisely explain its behaviors under all circumstances, with the mean scores of Q8 being between 1 (Moderately Agree) and 2 (Agree).

Post-study Questions: Need for Explanation

Participants preferred the robot to get their attention before starting to explain its behaviors, with 79% of them agreeing (Q10). To get the participants' attention, 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 an actions like waving its arm (Q11). 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 (Q12). We also found that participants strongly preferred the robot to explain its behavior (Q9), with a mean score above 2 (Agree). Participants acknowledged the value of signaling failure through gestures (Q13), but preferred verbal announcements of failure from the robot (Q14), with the mean scores being between 1 (Moderately Agree) and 2 (Agree). Finally, participants did not prefer non-verbal explanations from the robot (Q15), with a mean score being between -1 (Moderately Disagree) and 0 (Neutral).

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 (Ullman, Aladia, and Malle 2021). Thus, we designed our in-person experiment under the assumption that our results could be different from those of the online experiment conducted by Han, Phillips, and Yanco (2021). However, replicated conditions had consistent findings between their online Baxter study and our in-person Fetch 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. In addition, 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 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 and Yanco (2023), which claimed that standalone projection markers can worsen participants' causal inference, as only half of their participants correctly inferred the missing information about the object picking task when only projection marker was used. In contrast, projection was comparable to the other non-verbal conditions.

Findings from Han, Phillips, and Yanco (2021) suggested that speech is preferred to gesture for robot explanations, which our findings partially confirmed (Q13-15). Moreover, findings from our post-study questionnaire provided further evidence that speech was a preferred component of 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 agree with prior work suggesting human preference for verbal over nonverbal communication. For example, Nikolaidis et al. (2018) reported that short verbal commands were more effective than non-verbal actions to promote human-robot trust in collaborative tasks.

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, Projection as an explanation component performed similarly to the status quo of head motions and paired head and arm motions.

Acknowledgments

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

References

Ambsdorf, J.; Munir, A.; Wei, Y.; Degkwitz, K.; Harms, H. M.; Stannek, S.; Ahrens, K.; Becker, D.; Strahl, E.; Weber, T.; et al. 2022. Explain yourself! effects of explanations in human-robot interaction. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 393–400. IEEE.

Anjomshoae, S.; Najjar, A.; Calvaresi, D.; and Främling, K. 2019. 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*, 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

Cao, X.; Gautam, A.; Whiting, T.; Smith, S.; Goodrich, M. A.; and Crandall, J. W. 2023. Robot Proficiency Self-Assessment Using Assumption-Alignment Tracking. *IEEE Transactions on Robotics*.

Carlson, J.; Murphy, R. R.; and Nelson, A. 2004. Follow-up analysis of mobile robot failures. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 5, 4987–4994. IEEE.

Choi, S.; Mattila, A. S.; and Bolton, L. E. 2021. To err is human (-oid): how do consumers react to robot service failure and recovery? *Journal of Service Research*, 24(3): 354–371.

De Graaf, M. M.; and Malle, B. F. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

Edmonds, M.; Gao, F.; Liu, H.; Xie, X.; Qi, S.; Rothrock, B.; Zhu, Y.; Wu, Y. N.; Lu, H.; and Zhu, S.-C. 2019. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, 4(37): eaay4663.

Fitzgerald, C. 2013. Developing baxter. In *2013 IEEE conference on technologies for practical robot applications (TePRA)*, 1–6. IEEE.

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

Han, Z.; Phillips, E.; and Yanco, H. A. 2021. 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.

Han, Z.; and Yanco, H. 2023. Communicating missing causal information to explain a robot's past behavior. *ACM Transactions on Human-Robot Interaction*, 12(1): 1–45.

Honig, S.; Bartal, A.; Parmet, Y.; and Oron-Gilad, T. 2022. Using online customer reviews to classify, predict, and learn about domestic robot failures. *International Journal of Social Robotics*, 1–26.

Khanna, P.; Yadollahi, E.; Björkman, M.; Leite, I.; and Smith, C. 2023. Effects of Explanation Strategies to Resolve Failures in Human-Robot Collaboration. *arXiv preprint arXiv:2309.10127*.

Nikolaidis, S.; Kwon, M.; Forlizzi, J.; and Srinivasa, S. 2018. Planning with verbal communication for human-robot collaboration. *ACM Transactions on Human-Robot Interaction (THRI)*, 7(3): 1–21.

Rosenthal, S.; Selvaraj, S. P.; and Veloso, M. M. 2016. Verbalization: Narration of Autonomous Robot Experience. In *IJCAI*, volume 16, 862–868.

Scheunemann, M. M.; Cuijpers, R. H.; and Salge, C. 2020. 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)*, 1340–1347. IEEE.

Schrum, M. L.; Johnson, M.; Ghuy, M.; and Gombolay, M. C. 2020. 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*, 43–52.

Shen, J.; and Gans, N. 2018. Robot-to-human feedback and automatic object grasping using an RGB-D camera–projector system. *Robotica*, 36(2): 241–260.

Tolmeijer, S.; Weiss, A.; Hanheide, M.; Lindner, F.; Powers, T. M.; Dixon, C.; and Tielman, M. L. 2020. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, 3–12.

Tuncer, S.; Licoppe, C.; Luff, P.; and Heath, C. 2023. Recipient design in human–robot interaction: the emergent assessment of a robot's competence. *AI & SOCIETY*, 1–16.

Ullman, D.; Aladia, S.; and Malle, B. F. 2021. Challenges and opportunities for replication science in hri: A case study in human-robot trust. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, 110–118.

Vogel, C.; Poggendorf, M.; Walter, C.; and Elkmann, N. 2011. Towards safe physical human-robot collaboration: A projection-based safety system. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3355–3360. IEEE.

Vogel, C.; Walter, C.; and Elkmann, N. 2012. 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*, 67–72. IEEE.

Wise, M.; Ferguson, M.; King, D.; Diehr, E.; and Dymesich, D. 2016. Fetch and freight: Standard platforms for service robot applications. In *Workshop on autonomous mobile service robots*, 1–6.

Yang, X. J.; Unhelkar, V. V.; Li, K.; and Shah, J. A. 2017. 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*, 408–416.