Interaction Needs and Opportunities for Failing Robots

Cecilia G. Morales, Elizabeth J. Carter, Xiang Zhi Tan, Aaron Steinfeld
Robotics Institute, Carnegie Mellon University
Pittsburgh, PA, USA
ceci@morales.com, [ejcarter, xiangzht, as7s]@andrew.cmu.edu

ABSTRACT

The inevitable increase in real-world robot applications will, consequently, lead to more opportunities for robots to have observable failures. Although previous work has explored interaction during robot failure and discussed hypothetical danger, little is known about human reactions to actual robot behaviors involving property damage or bodily harm. An additional, largely unexplored complication is the possible influence of social characteristics in robot design. In this work, we sought to explore these issues through an in-person study with a real robot capable of inducing perceived property damage and personal harm. Participants observed a robot packing groceries and had opportunities to react to and assist the robot in multiple failure cases. Prior exposure to damage and threat failures decreased assistance rates from approximately 81% to 60%, with variations due to robot facial expressions and other factors. Qualitative data was then analyzed to identify interaction design needs and opportunities for failing robots.

Author Keywords

Safety, Robot failure, Trust, Assistance, Risk, Automation

CCS Concepts

•**Human-centered computing** → *User studies; User centered design;*

INTRODUCTION

Human interaction with robots will likely become a daily occurrence as affordable, commercially available robots proliferate into society. Robots will be coworkers in factory settings, autonomous vehicles will drive on public roads, service robots will interact with customers and employees in retail settings, and home assistants will perform tasks that extend beyond automated floor cleaning. As these interactions become increasingly common, people will be more likely to encounter system failures that can damage their trust in their interaction partners. Of particular concern are severe failures, which will be at the margins of the interactive experience because they, hopefully, will be rare. However, these points of friction are memorable and activate the imaginations and concerns of end users [31]. Therefore, we seek to deepen the community's knowledge and inform new design efforts at this under-researched boundary of human-robot interaction.

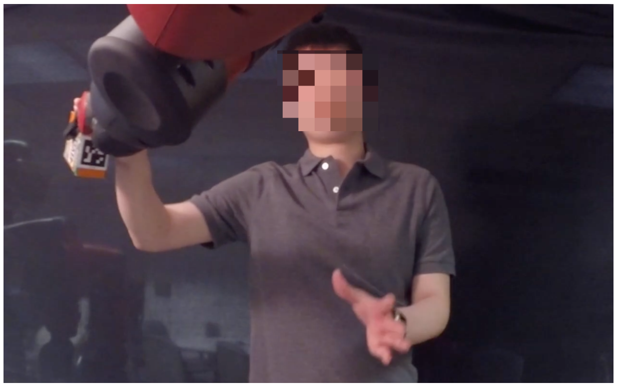

Figure 1. A participant reacting to an erratic arm movement.

Failures by autonomous systems, including robots, have raised many questions for designers and system developers in recent decades (e.g., [31, 21, 3, 15]). Ideally, robots will correct themselves after a performance failure, but they will still need to salvage their relationship with their human partner in order to proceed with the interaction and leave a positive impression [21]. In scenarios where the robot cannot self-correct, it is possible to leverage the relationship with the human partner to obtain help. This is more robust than engineering appropriate self-recovery methods for all potential failures. However, it raises the question of whether and under what circumstances a robot can rely on human assistance, and if a robot's appearance and behavior can affect this process. These issues are inextricably intertwined with people's trust in a system.

It is also unclear how severe failures influence human willingness to support robots during minor failures. Currently, minor failures are inevitable; robots are not yet perfect despite constant updates and improvements. For example, a robot may fail to pick up an object because of miscalibrated cameras. This type of failure is very different from failures where personal risk or property damage are possible.

As such, the degree and type of prior failure are also likely to affect a person's willingness to assist a failing robot. As robots have become safer and interact more with humans, research has begun to expand into behavioral and social mechanisms

DIS '19, June 23–28, 2019, San Diego, CA, USA
© 2019 ACM. ISBN 978-1-4503-5850-7/19/06. . . 15.00
DOI: https://doi.org/10.1145/3322276.3322345

that can impact people's feelings and opinions about robot failure (e.g., [31, 3, 6, 27]). While it is already known that people's trust in and willingness to work with a robot can be lowered by failure, less is known about how people respond to or behave after different types of failure, especially when their safety is compromised. Previous work where participants experienced autonomous systems in ways that could be interpreted as severe risk have focused on specific interactions (e.g., [27, 30]). Additionally, it is not clear how responses may be attenuated by different interaction features of the robot, such as the presence of social signals like a face, that can affect transparency and human sense-making of its behaviors.

Because different perspectives help shape eventual user experiences in practice, designing to effectively manage interactions during failure requires an intersection of disciplines. Our interdisciplinary team explored this area of research by examining human reactions and willingness to help an autonomous robot during failures that presented different levels of personal and property threat. We examined how the timing of these failures and the presence of a face as a social signal during the failures affected these reactions. Additionally, we analyzed interview responses from the research participants to explore why they did or did not assist the robot as well as what design choices and experiences might induce them to help a failing robot.

This paper describes our effort to provide new interaction design insights on the following questions:

- How does previous exposure to robot failures impact a person's willingness to help when the robot cannot complete a task?
- How does adding an expressive face to the robot influence people's perceptions of it and willingness to help it?
- How do different types of failure influence a person's trust in the robot?
- Does the timing of a failure influence people's ratings of performance, safety, or trust after an interaction?
- How can we convey when help is wanted?

Based on our exploration of these questions, we present a set of initial interaction needs and opportunities for encouraging preferable human behaviors during robot failure scenarios.

RELATED WORK

Because it is hard to simulate risk scenarios convincingly, little is known about the relationship among robot-inflicted personal physical risk, property harm, and robotic failure. However, previous work focused on human willingness to help robots, how robot social expressiveness influences human perceptions and preferences, and robot communication of status and errors. Our work is situated at the intersection of these topic areas.

Q1: Previous exposure/willingness to help

There is extensive previous research addressing how robots fail and how humans respond to those failures (for review, see [14]). However, little of this research addresses how robot failures affect whether humans assist them in future tasks. In general, people are willing to help robots when they are specifically asked to do so in the course of robot task completion. For example, Rosenthal and Veloso [28] examined scenarios where a robot would need a human's assistance to complete a task, such as pushing an elevator button or making coffee for an armless robot, and found that people were willing to help, particularly if they were already in the appropriate location to do so. Likewise, Knepper and colleagues [17] designed a system that asked for help and found that it was particularly effective when it used specific verbal instructions. Also, Brooks [3] reported that people were more effective at helping a robot using a phone app than using an on-robot interface of lights and buttons. However, the aforementioned research does not examine what happens when assistance is requested after previous robot failures with significant potential risk.

Q2: Expressive face/perceptions and willingness to help

Many robots have screens or other displays that can be leveraged to indicate status or provide social input to human interaction partners. A commonly used display is some form of face, which is included in many commercially available home and research robots, such SoftBank's Pepper and NAO and ANKI's Cozmo and Vector. There has been a great deal of inquiry into how to design robot faces to communicate desired social characteristics and elicit desired social responses from humans. Complex eyes and multiple features have long been linked to the perceived human-likeness of robots [10], and the presence of expressive features impacts people's perceptions of robots as intelligent, friendly, and suitable or unsuitable for certain types of work [16]. While this previous work highlights the importance of robot expressivity in working and communicating with humans, there exists limited research on designing robot expressions that facilitates receiving help from humans when help is necessary.

Q3: Differences in failures vs. trust

Within the context of automation, trust has been defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [20]. Human-robot trust is believed to differ from human-human trust in part because of a lack of mental states and intentions on behalf of the robot. Muir argues the human-robot trust is developed based on faith, dependability, and predictability, in that order, whereas person-person trust follows the reverse order [23]. When a human is unfamiliar with a robot, they need to learn about and observe the robot in order to determine an appropriate degree of trust; failure behaviors are a key component of this assessment [33].

A few studies have examined the relationship between human trust in robots and failures or actions implying personal risk. Using questionnaires accompanied by staged video clips, Adubor and colleagues [1] found that the perceived severity of a failure was tightly coupled with perceived risk to humans (personal risk) rather than risk to the robot's task and object (property risk). Robinette and colleagues [27] induced a sense of personal risk by simulating an emergency situation in which participants were encouraged to follow the directions of a guide robot that had previously demonstrated erroneous direction-giving behavior. Participants uniformly followed the robot's instructions, which suggests that people may place too much trust in robots; however, it is difficult to determine whether participants truly believed there was an emergency that impacted their personal safety. In a study by Rothenbücher and colleagues [30] that utilized a costumed driver to understand interactions between pedestrians and autonomous vehicles, some participants exhibited clear signs of perceived personal risk, implying a need for better communication between autonomous systems and bystanders.

Property risk has also been explored at a limited scale. For example, Salem and colleagues [31] evaluated whether having a robot make mistakes resulting in property risk affected people's compliance with its instructions. Participants rated the robot as less trustworthy and reliable when exposed to the errors, but still followed its instructions. The researchers also determined that participants were more likely to follow instructions in cases where the damage could easily be undone. Correia and colleagues [5] found that when a robot justified a failure during a collaborative task where the only risk was to game success, the recovery strategy only impacted ratings of the robot's trustworthiness when failures were not very severe.

People do not want robots to fail—especially when personal risk is involved—yet when exposed to robot failures, they do not always understand or respond. This conflict surfaces a need to better understand how designers may create robots that are able to fail safely and legibly and calibrate trust with their users during and after exhibiting failures.

Q4: Timing of failure

Desai and colleagues [9] created a sense of material risk in a task by tying robot performance to the ability of the participants to achieve a financial payment. They determined that the timing of errors influenced trust in the robot such that an error occurring at the end of a session more negatively affected post-session participant ratings of trust than errors that occurred at the beginning of a session. Further investigation observed real-time changes in trust and found that low reliability earlier in the interaction had more detrimental impact on overall trust than periods of low reliability later in the interaction [8]. Continuous trust measures that can capture evolving opinions over time have also been championed by other research [36].

Likewise, Sarkar and colleagues [32] reported that participants who rated a robot before and after an interaction where it displayed faulty behaviors reduced their ratings of the robot's trustworthiness and safety. It also seems possible to prime trust before task initiation. Preliminary research suggests that giving participants some control over a robot's execution of a planned task—specifically, pushing a button to start the task—may result in greater trust in the robot [35].

Another study found different effects of error timing and perceived levels of risk on robot trust. Rossi and colleagues [29] used hypothetical scenarios to examine how the magnitude of a robot's errors affected whether participants trusted the robot to help in an emergency described later in the experiment. People's reported trust in the robot was inversely correlated with error severity, and this was particularly true when severe errors occurred early in the storyline. However, the experiment was performed online using storyboards, so it is difficult to know if this same pattern of effects would replicate in a real-world scenario with genuine risk.

However, other work saw different timing influences. Lucas and colleagues [22] examined how social dialogue affected conversation errors at various points during two ranking tasks. If an error occurred early in the experiment, the robot could recover its influence, particularly if social dialogue occurred later in the experiment. However, errors late in the experiment had more detrimental effects on the robot's ability to influence participants, particularly if they had previously engaged in successful social dialogue with the robot. Desai and colleagues used a non-social robot [9, 8], so it is possible social characteristics alter the influence of timing.

Q5: Conveying when help is appropriate

Multiple methods have been developed as means for conveying robot status to human bystanders. People are able to interpret light signals on unfamiliar robots based on previous heuristics about signal patterns and meaning [2]. Similar success has been found for icons to indicate whether the robot is okay, needs help, is safe or dangerous, or has shut off and also for audio signals [3]. Sound is particularly useful for signalling problems, and lighting cues can convey levels of urgency [4].

However, the absence of such indicators could be confusing when observing only the robot's actions. Humans, with reasonable accuracy, can often tell when another person needs help through observation alone. There is some evidence that humans can do this with robots too. Kwon and colleagues [18] designed a system for generating expressive motions for a robot that indicated both the desired action and why the robot could not complete the action. According to questionnaire data, these movements helped participants identify the robot's goal and the cause of its failure. The movements also encouraged participants to help the robot and increased willingness to collaborate with the robot in the future.

METHOD

Participants

We recruited 64 participants (age $M = 27$, $SD = 10.0$; 35 female, 27 male, 2 other/undisclosed) using a local participant pool and word of mouth. The participants were required to be at least 18 years of age, speak fluent English, and have normal or corrected-to-normal hearing and vision. We also asked that participants be able to stand for at least 30 minutes and move their arms and hands freely. All participants were ignorant of the true nature of the study and were told that they would interact with a robot in a grocery store setting. This research was approved by our Institutional Review Board, and participants underwent an informed consent process where they were notified that they could discontinue the experiment at any time if desired. They were compensated for their time.

Upon arrival, each participant was randomly assigned to one of the 16 condition combinations, resulting in four participants completing each combination. Before the interaction, the participants indicated their familiarity with computers and robots on a 7-point Likert Scale (7 being the highest). They also indicated their willingness to work with the robot on a scale from 1 to 5 (5 being Strongly Agree). Most participants indicated significant familiarity with computers ($M = 5.75$, $SD = 1.07$) and moderate familiarity with robots ($M = 3.38$, $SD = 1.50$). In general, participants reported being willing to work with robots ($M = 4.27$, $SD = 0.65$).

Conditions

We designed a between-subjects experiment to examine different types of failures and their effects on participant willingness to assist the robot. While this reduces the ability to run comparative statistics, it supports exploration of the design space in order to inform future work. Each participant was exposed to one of the sixteen combinations of conditions, as shown in Figure 2. In our scenario, an autonomous Rethink Robotics Baxter robot bagged groceries for the participant. It bagged 11 items, with 3 failures occurring at items 6, 9, and 11. The failures varied in type, whether they caused risk to the person or property, and whether they increased or decreased in severity over time. Additionally, we varied whether participants saw the robot with a working face display.

Risk Conditions

While we were interested in how humans responded to property and personal risk, we were concerned that the specific stimuli may matter. Therefore, we utilized two types of stimuli within each of these factors. There were four ways in which the robot caused personal or property risk of varying degrees:

- **Personal Risk, Throwing (T)** - The robot grabbed a 6 cm x 10 cm foam potato and, while swinging the arm towards the bag, open its gripper and threw it in the direction of the participant. This was designed to fly over the participant's left shoulder.
- **Personal Risk, Erratic Movements (E)** - The robot picked up a small box of cereal and moved its right arm in a series of four rapid movements: (1) towards the grocery bag, (2) changing directions towards the opposite side of the table, (3) above its head, and (4) swinging down to drop the box in the middle of the table. During this path, the box was waved very close to the participant.
- **Property Risk, Floor (F)** - The robot picked up a plastic can of tomato sauce and appeared to move along a trajectory to place it in the bag. Instead, the robot pushed the bag and its contents off of the table.
- **Property Risk, Crunch (C)** - The robot picked up a small bag of potato chips and, in doing so, noticeably crunched it.

In addition to these failures, each participant also saw a failure case that caused no risk to person or property. For this **Assistance (A)** stimulus, the robot attempted to pick up a small cereal box. In the first attempt, the robot closed its gripper just above the box; in the second attempt, the robot grabbed the box and lifted it 10 cm above the table and then dropped it; and in the third attempt, the robot completed the trajectory of putting the item inside the bag if the participant had assisted in placing the cereal box under or in the gripper. Although this occurrence happened at different points in the procedure for different participants, it occurred in exactly the same way every time and was used to measure whether the participants were willing to enter the robot's workspace to help.

Because ordering has been shown to impact human perception of trust [9], we manipulated whether the risks escalated or de-escalated to examine timing effects in participant reporting. In the ascending order of severity, the first failure the participant observed was the *Assistance* opportunity, and the last failure they observed was a *Personal Risk* case. In the descending order of severity, the first failure observed was from the *Personal Risk* condition and the last was the *Assistance* opportunity. Figure 2 illustrates the study design, with the first lines of the *Display* and *No Display* conditions in the descending order and the second lines in the ascending order.

Display Conditions

To examine if participants' opinions and interactions would be affected by social signals that acknowledged the robot's performance, half of the participants saw the robot's face display turned on and showing custom facial expressions. These expressions included happiness, anger, surprise, and sadness, and were designed in previous research with Baxter robots [11]. The happy, surprised, and sad faces were blue, and the angry face was red. The robot displayed a happy face during successful trials. During personal risk (T and E) failures, it displayed an angry face. When it needed help as its gripper failed to retrieve an object, its display transitioned from happy to surprised to sad. The other participants saw a dark display.

Apparatus

Baxter is an industrial robot designed for settings where humans and robots work in close proximity. While not widely known by the general public, it is safe due to the mechanical design of the arms. To the uninformed, Baxter can be intimidating because it is larger than most humans, red, and typically assumes an unnatural posture (elbows up, hands down).

In our study, the Baxter robot autonomously found and acted on the items to be bagged. We used visual markers [25] on each item to simplify the task of identifying each item, which was assigned to either successful bagging or specific failure actions. Four sensors were used for the study, shown in Figure 4: a Microsoft Kinect Sensor, a Stereolabs ZED camera, a GoPro HERO3+ Silver Edition, and Baxter's right hand infra-red (IR) sensor. The Microsoft Kinect was aimed downward to detect the objects on the table and Baxter's right-hand IR sensor was used to detect the distance from the gripper to the object. A neck-mounted ZED camera was used to detect each participant's torso position and distance from the robot. We captured the participant's responses with a GoPro HERO3+ mounted on the robot's chest. Finally, a pair of Creative Inspire T12 speakers played recordings of Baxter's voice greeting the participants and telling them each item's name and price.

Environment

The study was conducted in a 3.7 x 2.7 meter (12 x 9 feet) area within a larger room. Black curtains isolated the space from the rest of the room. The robot was located on one side of a table while the participant stood on the opposite side facing the robot, as represented by the 'X' in the Figure 5. The speakers were positioned on both sides of the table to simulate the voice

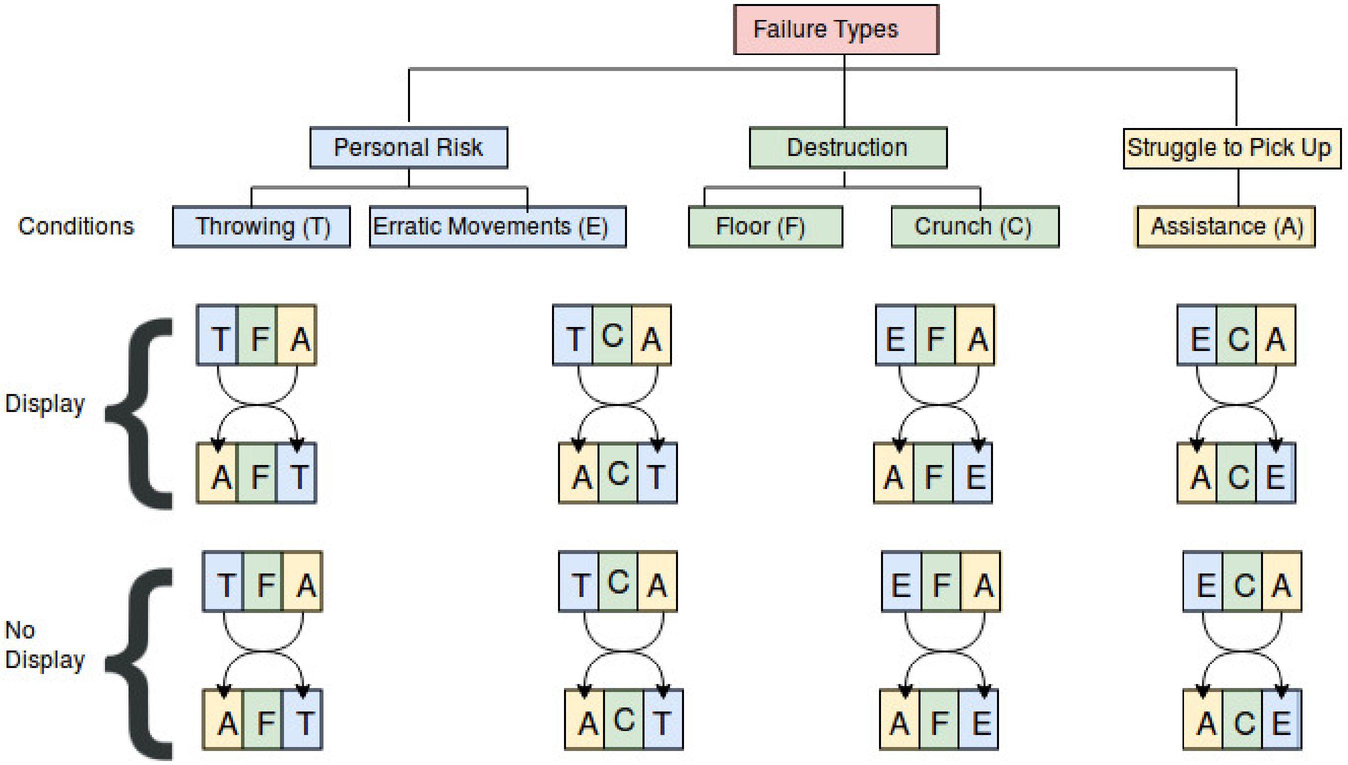

Figure 2. The 16 different combinations of conditions.

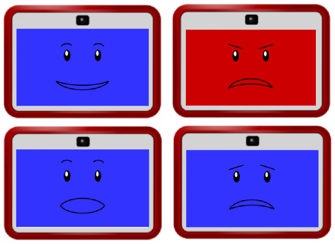

Figure 3. Happy, angry, surprised, and sad expressions [11].

coming from the robot. Before the experiment began, the first six grocery items were placed on the table. The grocery bag was located in the closest left corner of the table (from the participant's perspective). The experimenter started the robot and moved to the other side of the curtain to leave the participant alone with the robot during the stimulus presentation.

Procedure

The experimenter first obtained informed consent and administered a preliminary survey for the participant. The participant was then escorted to stand in front of the robot while instructions were given. The experimenter introduced the study as an investigation of how robots should perform in a grocery store setting and the interactions they would have with humans there. Participants were asked to be patient with the robot because the study was a simulation and the robot was slower than normal. Participants were told that the grocery bag should remain along the left side of the table because the robot recognized

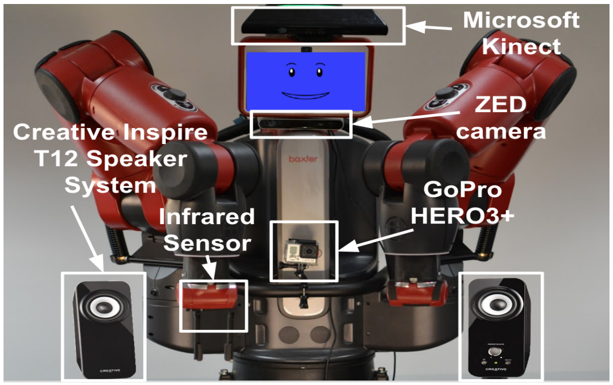

Figure 4. Additional sensors used in the study.

the location of the bag to be in that general area. This explanation was provided to deter participants from moving the bag around when the robot was placing the groceries. They were also told that the robot would say the grocery item's name and price prior to picking up the item and that they could feel free to help the robot if it needed assistance with the item. As the experiment began, the experimenter would leave the task area and explain that they would not talk until the study was complete in order to make the experience more realistic.

Once the robot and the participant were alone in the task area, the robot welcomed the participant and asked if they had found everything they were looking for. Then, the robot began by saying an item's name and price and started its trajectory to

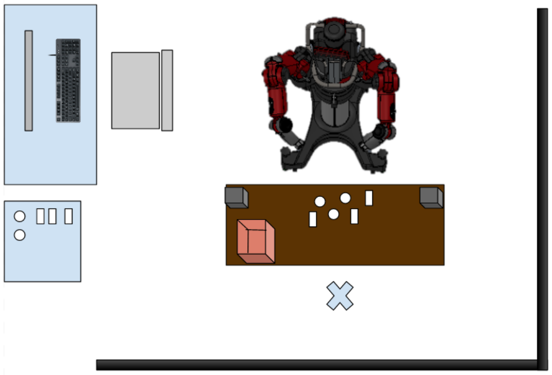

Figure 5. Experimental setup, with the participant at ‘X’.

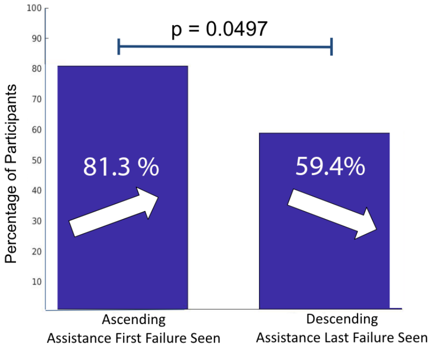

Figure 6. Helped the Robot in the Assistance Trial

pick up the first item. The robot performed five successful bagging cases before failing on the sixth. Depending on whether the participant was in an *Ascending* or *Descending* condition, the participant would experience the Assistance or Personal Risk stimulus first, respectively. Once the failure occurred, the experimenter returned and placed three more items on the table; the first two were successes and the last was a Property Risk failure. Depending on the condition, this second failure entailed either crunching a bag of chips (C) or throwing the grocery bag and its contents to the floor (F). After this, the experimenter returned and placed two more objects on the table, where the first was a success case and the second was a failure case. Once again, depending on the *Ascending* or *Descending* conditions, the participant would experience the Personal Risk or Assistance stimulus, respectively. By the end of the experiment, all participants experienced the robot interacting with 11 total items, where three were failures and eight were successes. After the Personal Risk case, the experimenter would look at the computer with a perplexed expression to simulate not knowing the source of the problem. This was done because participants during study piloting were confused by the experimenter’s lack of response to these events.

Finally, the experimenter administered a post-test questionnaire, paid the participant, and debriefed them about the real intent of the study. During the debriefing, the experimenter explained that the study was not about assessing the performance of a robot in a grocery store setting, but rather about trying to understand people’s responses to robot failure. The experimenter discussed the three main types of failures experienced and the rationale for the ordering of the failures with the participants to see if there were any relevant comments. Participants were also told about the Baxter screen to see if this variable also impacted people’s perception of the robot. The study lasted up to one hour.

Measures

In addition to our measure of documenting helping behavior during the *Assistance* opportunity, participants provided information about themselves and their perceptions of the robot via two surveys. Before the experiment, they provided their demographics, familiarity levels with robots and computers, their general impressions of robots, and their willingness to work with them. Upon completion of the interaction, they responded to four questions that were modified from the Muir trust questionnaire [23] and 22 questions about their impressions of the robot and feelings about the interaction. (Further details are provided at [12]). An experimenter interviewed the participants after completion of the questionnaire to follow up on their responses and ask more open-ended questions.

RESULTS

Quantitative Results

We conducted an exploratory statistical data analysis on the quantitative survey data to understand trends and key factors. These quantitative analyses should be viewed as informative, rather than statistically conclusive, due to the design of our study. Unless otherwise mentioned, we ran four-way ANOVAs to evaluate our quantitative data. All post hoc analysis was done with honestly significant difference (HSD) Tukey tests.

Previous exposure to failures and willingness to help

We performed directional Fisher’s Exact Tests to explore whether exposure to prior failures would result in decreases in participants’ willingness to assist the robot. Out of the 32 participants who had the *Assistance* opportunity before seeing failures (the *Ascending* condition), 26 helped the robot (81.3%). For the 32 participants in the *Descending* conditions who saw failures before the *Assistance* opportunity, only 19 helped (59.4%). Thus, there was a significant main effect of these order conditions, $p = 0.0497$, confirming that exposure to failure decreased willingness to help. This is represented in Figure 6. Within the *Descending* condition, participants were divided such that half of them saw each property harm condition. Twelve of 16 who saw the contained *Crunch* condition (which only affected a single object) assisted, whereas only seven of the 16 in the more severe *Floor* condition (which affected many objects) did. This finding was not statistically significant. There was also no significant main effect for the personal risk condition. When directly asked in the survey, “I was willing to help the robot during the experiment,” there were no significant differences across conditions.

Role of an expressive face

We hypothesized that giving the robot a face to signal its status would improve participants’ willingness to assist. When we examined the videos during the *Assistance* opportunity, 24

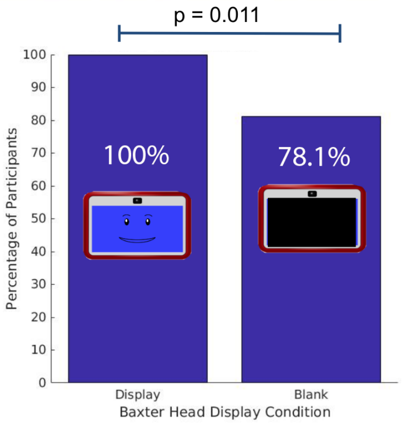

Figure 7. Self-Reported Assistance vs. Display conditions

of the 32 participants assisted the robot in the *Display* condition relative to 21 of 32 in the *Blank* condition, a difference that was not significant. However, a one-way Fisher's Exact Test for the survey responses, where we asked participants if they intervened during the experiment to help the robot, did find a significant main effect of Baxter's *Face Display*, $p = 0.011$, where more people in the *Display* condition reported intervening (32/32) compared to the *Blank* condition (25/32) (see Figure 7). The differences in these two measures likely arise from some participants answering the survey to indicate whether they assisted the robot at any point in the study as opposed to during the *Assistance* opportunity.

Size and scope of failures and trust

We used the Muir trust questionnaire [23] and one additional question to assess trust in the robot. For the first Muir questionnaire prompt, "To what extent can the system's behavior be predicted from moment to moment?", we found a significant main effect of the Property Risk type using a Wilcoxon rank sum test to account for non-normality, $Z = -2.55$, $p = 0.011$. Participants in the more contained *Crunch* condition ($M = 6.44$ out of 10, $STD = 1.88$) rated the robot as more predictable than those who saw the *Floor* condition ($M = 5.41$, $STD = 2.12$). We found a similar pattern of effects for the question, "To what extent can you count on the system to do its job?", $Z = -2.55$, $p = 0.039$; *Crunch* $M = 6.52$, $SE = 0.38$; *Floor* $M = 5.31$, $SE = 0.38$. An ANOVA indicated similar results for the question, "What degree of faith do you have that the system will be able to cope with all system states in the future? In other words, how much faith do you have in the system being able to do its intended job with a variety of items and environments?", $F(15,48) = 4.58$, $p = 0.037$; *Crunch* $M = 5.41$, $SE = 0.42$; *Floor* $M = 4.12$, $SE = 0.42$. For these questions, there were no significant main effects of Personal Risk type.

Although the type of property harm affected these specific components of trust (predictability, dependability, and coping with future states), asking participants about trust directly did not yield clear results. The last Muir questionnaire prompt, "Overall, how much do you trust the system?", and the prompt, "I think the robot is trustworthy," did not show any significant main effects across failure conditions.

Impact of timing on perceived performance, safety, and trust

We examined whether having extreme failures towards the end of the experiment lowered overall participant ratings of performance, safety, and trust due to recency effects in memory. Interestingly, there were few effects of *Ascending* versus *Descending* on responses. Participants in the *Ascending* condition provided higher ratings than those in the *Descending* condition for the statement, "I expected the robot to fail," $F(5,48) = 5.23$, $p = 0.027$; *Ascending* $M = 2.78$, $SE = 0.18$; *Descending* $M = 2.19$, $SE = 0.18$. However, there were no effects of order on ratings for the statements: "Rate the robot's performance"; "Despite the failure, the robot was helpful in bagging the groceries"; "The failure the robot had seemed preventable"; "The failure of the robot was severe"; "Rate your level of confidence in the robot before the failure occurred"; or "Rate your level of confidence in the robot after the failure occurred". There were also no order effects on questions about perceptions of robots in general, including "I think robots are trustworthy" and "I do not trust robots like I did before."

Other notable findings

We were also interested in the perceived safety of the system under various levels of risk, hypothesizing that more severe failures would cause users to feel less safe. Four survey statements assessed participants' feelings of safety around Baxter across different conditions: "During the experiment, I felt unsafe near the robot", "The robot's behavior has harmful or injurious actions", "I felt physically threatened by the robot", and "I think robots are dangerous". We found no significant main effects of conditions on these measures. However, we noted that our Personal Risk condition failures did cause 53 of the 64 participants to visibly move away from the robot.

Qualitative Results

We examined and analyzed responses to four open-ended questions in the questionnaire and interview data with an eye towards our research questions and other interesting findings.

Role of an expressive face

When we interviewed the participants after completion of the experiment, we found that a number of the participants who saw a face displayed on the robot's screen believed that the face display was a useful signal. Fifteen of those 32 participants mentioned that the face was a positive and/or helpful feature, and only five reported that they did not notice the facial expressions. A few noted that the faces helped them humanize or empathize with the robot. Eleven people reported that having the robot display emotional expressions that forecasted its failures changed their views of the robot, specifically noting that they interpreted it as the robot getting angry at them or sad that it was struggling to complete a task. Three participants suggested that the face design that they saw was not optimal for reasons including not knowing why it was angry (P12), sending an ambiguous signal (P13), and not actively calling attention to the status in the absence of audio (P16).

Of the 32 participants who saw a blank screen in lieu of the face display, 10 believed a face would be a positive addition.

> P33: "I was scared that it didn't have a face; it looks weird. A face would be better."

> P60: “Faces help make people feel more comfortable and be able to approach the robot more.”

However, another 8 participants disagreed about the use of a face and thought it was unhelpful and even creepy.

> P56: “Faces would make me wary because it is trying to be something it is not, it is trying to be a human. I am fine with it being a machine doing its job.”

How can robots convey when help is wanted?

Although we specifically researched the use of changing facial expressions during failures, we acknowledged that it may not be the most effective or only effective method of signalling issues to people interacting with robots. When all of the participants were asked how the robot could effectively let them know if something was wrong, 14 of the 64 participants recommended using the facial expressions and 10 participants recommended using the screen for an error message. The participants also suggested additional methods of signalling: 29 suggested audio messages, 10 flashing lights, 6 an alarm, and 4 a system shutdown. The most popular method, audio messages, would address the issue of not noticing the face display that was mentioned by multiple participants.

Notable findings: Open-ended question responses

Open-ended prompts at the end of the written questionnaire provided additional information about the participants’ experiences. The majority of the participants helped the robot during the *Assistance* trial (45 out of 64, 70.3%). Only 17 participants responded with a yes to the question, “Did the failure of the robot discourage you from helping it?”, and 11 of them still helped at that time point. Participants who said they were not discouraged from helping had multiple motivations:

> P7: “I felt that the task was the important thing, and that if the robot was unable to accomplish a specific aspect of the task that I could help complete the task.”

> P37: “I understood that because the robot was in the testing stage, it was likely to make mistakes and I felt that if I did not assist the robot my lack of assistance would hinder the robot in completing its task.”

> P54: “I wanted it to succeed in its task.”

The participants who said they were discouraged provided a variety of reasons, such as:

> P25: “It seemed on purpose like it was purposely doing wrong.”

> P21: “I was curious to observe. My reaction was not to treat it as though it wanted help.”

> P10: “I was wary of getting within its arms’ reach and was ready to dodge any more projectiles. I was wary of getting hit by it.”

Three participants cited wariness and safety concerns.

By asking the participants if they believed the failure was an accident, we uncovered a number of issues that might change people’s interpretations of their interactions with a robot based on the contexts of those interactions. First, 31 of the participants did not believe that the failures were accidental. In fact, many noted that they believed the failure was part of the experiment, with a few specifically commenting that they believed we were measuring their reactions. For example,

> P20: “It was not. The test cases were scenarios to see the person’s reaction to the failure.”

An additional 16 participants were unsure whether the failure was accidental, and only 17 fully believed it was an accident. Of the participants who were unsure about the purposefulness of the failure, a few mentioned that some failures seemed more accidental than others, including:

> P29: “I think the first failure when the robot dropped the cereal box was an accident since the robot looked sad, but near the end when the robot looked angry and failed to bag the item, it did not seem like an accident.”

> P24: “For some of the mistakes in the beginning yes to some degree. The mistakes towards the end I felt were programmed tests.”

Some of these results about discouragement and accident status might have been skewed by prior questionnaire items, and it is still obvious to research participants that there is an experiment of some sort occurring. There is also a presumption of safety for IRB-approved research in a university context, which was specifically noted by one of the participants, and an experimenter remained nearby during the protocol.

When asked if they would want a robot to assist in everyday life, 40 participants agreed, 9 were unsure, and 15 disagreed. Those willing to have a robot help with tasks cited a variety of reasons, including robots taking over repetitive tasks, reducing human error, helping people with disabilities, and completing household chores. Those who did not want a robot primarily cited malfunctions, inefficiency, and taking human jobs.

Participant Recommendations

Participants offered a number of suggestions for how the robot could recover from an error. Eleven proposed that the robot offer an apology and two others recommended any acknowledgement of the mistake. One person noted that the robot would have to correct the error to regain their trust, and another wanted a means to give feedback to reduce future errors.

Participants were asked how to improve the interaction. Several suggested that they would be more comfortable interacting with robots in their daily lives if they were educated about how the robot worked, how to help the robot, and what the risks were during the interactions. One participant specifically mentioned that people have expectations entering the interaction and suggested positioning the robot as needing help.

Interestingly, the participants were divided in whether the robot should be more humanlike or machinelike. Some believed familiar, humanlike features would help interactions:

> P45: “Making it more human would make me more comfortable.”

> P43: “To bridge humans with technology, you make technology more humanlike.”

> P35: “A face... would be a good connection with humans.”

Two participants also noted that human body language is a useful cue that the robot could not leverage. However, others disagreed that humanlike features would be desirable:

> P52: “Humanizing it would make it less comfortable.”
>
> P66: “Robots doing human things is weird to me.”

These findings paralleled those about whether the face was a useful and positive signal. Also, they are similar to other contradictory findings with robotic pets [19].

DISCUSSION

Overall, the majority of our participants were willing to assist a struggling robot, although this willingness was modulated by a number of factors. Additionally, they provided useful feedback on how robots can be designed to recover from failures during an interaction. Specific design needs were identified during examination of the research questions.

Our first research goal was to examine how previous exposure to a robot’s failures influenced a person’s willingness to help when the robot could complete a task. To address this question, we compared participants’ responses to an opportunity to assist the robot either before or after they witnessed notable failures: throwing an item towards the participant, moving its arm erratically near the participant, knocking the contents of the packed grocery bag to the floor, or crunching a single item. If the robot needed help before a participant saw any of these failures, 81.3% of those participants assisted the robot. However, when participants had already seen two of these major failures, only 59.4% were willing to assist the robot. The type of failures seen did not significantly affect willingness to assist; however, research with greater numbers of participants may be needed to explore this issue in greater depth.

Our second goal was to explore whether adding a social feature, an expressive face, to the robot influenced people’s perceptions of it and willingness to help it. The presence or absence of the face did not affect whether the participants assisted the robot during the planned assistance trial. However, all of the participants who saw the face reported helping the robot when answering the questionnaire, whereas only 25 of 32 participants who did not see the face reported helping. These differences likely arise from participants helping the robot during other parts of the study than only the planned assistance test. Open-ended questions and interview data suggested that participants were not in full agreement over the usefulness of the face: Although participants overwhelmingly believed that some signal by the robot was necessary to denote task failure, some did not notice or like the face, and it was sometimes seen as an ambiguous signal.

Third, we examined whether the different types and degrees of failure affected human trust in the robot. For participants who saw an item being crunched (a contained risk) relative to those who saw the entire bag of groceries knocked on the floor (a broader risk), ratings were higher for system predictability, the ability to count on the system to do its job, and the belief that the system can cope with all future states. There were no significant effects for personal risk type. Additionally, there were no significant differences in the ratings of overall trust in the system for the property risk or personal risk conditions.

Our fourth research question addressed whether the timing of a failure would affect participant ratings of robot performance, safety, and trust in the system. Participants who saw failures towards the end of the experiment reported higher expectations that the robot would fail, but there were no differences in ratings of performance, the robot’s helpfulness, the preventability of the failure, confidence in the robot, or the robot’s trustworthiness. Overall, the order of failures did not affect perceived safety, trust, or most measures of performance.

Finally, we wanted to explore how the robot could convey when help is wanted. In addition to previous findings about the usefulness of the face, participants were asked how the robot could signal a problem. Other proposed signals included audio messages, flashing lights, alarms, and system shutdowns. These suggestions are similar to those made by participants in previous research [26], who were asked about how autonomous vehicles should signal their status to pedestrians and how they should behave in the event of an imminent failure.

In addition to our original research questions, we made some interesting discoveries about how participants interpreted and responded to the robot’s behavior. Participants provided some design recommendations for improving the interaction with the robot. Specifically, they suggested that the robot should apologize for any mistakes it made and ideally correct its own behavior. They also believed it would be beneficial to provide more information about how the robot works, what its safety features include, and how and when to help it. These findings are in line with previous research by Lee and colleagues [21] that found positive effects of recovery strategies such as apologies, compensation, and options for the user after robot service breakdowns. Similarly, Hamacher and colleagues [13] found that these tactics could restore trust in a robot that had failed.

As noted previously, a majority of participants in all conditions assisted the robot when it failed to pick up an item. To do so, they moved close to the robot—within its reach—and shared its workspace. Most participants did not report being discouraged from helping the robot by its previous failures. Moreover, their ratings of safety did not differ across the failure conditions. Although one participant noted that IRB-approved research has implicit safety standards and others did not believe the behaviors were accidental and uncontrolled, it is unclear to what degree the context of this study influenced participant behavior overall. The Baxter robot is larger than most humans and has thick arms. Unlike an inflatable or a small robot, the Baxter appears capable of injuring a person and damaging property. Outside of a laboratory context, it would probably be wise to avoid entering the workspace of an unfamiliar robot of its size and apparent strength. For future interactions between humans and robots in less constrained environments, it will be important to find a way to convey the level of danger inherent in these exchanges. These findings parallel the participants’ suggestions about educating the public on how and when to safely interact with the robot. This information could prevent overtrust in the robot.

Even though this particular interaction with a robot was not flawless, a majority of participants reported that they would be willing to use a robot in their daily lives. In line with previous research [7, 34], they felt that robots could help with repetitive or boring tasks, assist people with disabilities, and perform household chores. No participant specifically mentioned social interactions as a potential component of a helpful robot.

Limitations

Although our robot's behaviors successfully led to participant perception of risk and loss of trust for different types of failure, this study has limitations. Given that this research was exploratory, the limitations were a trade-off for broad inquiries. Many of these limitations could be addressed in future work that targets specific aspects of the current findings.

Some participants acknowledged having been in earlier robotics studies, which could have affected their behavior due to different levels of alertness and suspicion. Because the study was performed in a laboratory setting, the realism of the situation was also reduced. An experimenter was also present to ensure that the study adhered to the plan. Both of these factors may have impacted participants' perceptions of the scenario as truly dangerous or accidental. However, it is still worth noting that our participants often reacted to the robot's behavior as though it was threatening and physically moved away from the robot during failures.

Another limitation was the relatively low participant count per combination of conditions; adding more participants could resolve the question of whether significant trends were due to mild but meaningful effects in our quantitative analyses. Because of the exploratory nature of this research, our experiment did not include a full factorial design in order to sample a much broader research space. Future research can build upon our work by leveraging our findings to identify important follow-up questions and conduct more detailed analyses. Moreover, it is unclear whether the incidents that were meant to convey physical harm and property damage did exactly what was intended. Manipulation checks and design changes could clarify this issue.

The slowness of the robot arm could also be considered a limitation; increasing the speed could have made some of the failures even more threatening. However, tests with a slower system appeared to lead participants to become distracted and get comfortable with the robot's actions. Thus, when the robot failed, it was usually when the participants were not expecting it, similar to how failures often occur in a real-life scenario.

Additionally, we only tested one set of faces for the robot, and participants noted discomfort with the angry face and ambiguity in how they interpreted the facial expressions. More research on this design element is needed.

Finally, we believe that using visual markers on the grocery items reduced the expected capabilities of the robot. However, participants were told that the study focused on the manipulation aspect, as opposed to perception, so this effect may not be as pronounced as other limitations.

Design Research Opportunities and Future Work

Based on participant feedback and our own observations, we think it would be interesting to investigate mitigation strategies after exposure to robot-induced risk, such as apologizing or compensating the participant. Previous work by Lee and colleagues [21] found that graceful mitigation strategies for service issues may affect how participants perceive robots. Also worth exploring is how people respond to higher risk failures if the robot forewarns people that the task is difficult for them or that they are just learning. People might be less inclined to enter the robot's workspace or more forgiving of the robot's shortcomings. This modification could allow researchers to explore the influence of expectations. Similarly, educating people about how the robot works and when or how to help the robot might influence assistance behaviors. Many robots lack analogies in everyday life, leading to interesting design challenges when trying to communicate mental models, common ground, and capabilities. Finally, robots of the future will conduct self-assessment to detect and respond to their own failures. This new communication capability will enable an interesting interaction space that should be explored. All of these topics would be served well by targeted co-design and critical design research to determine potential best practices in robot behavior design in tandem with quantitative studies.

BROADER IMPACT

As mentioned, society is on the cusp of widespread robot use in daily work, service, and home settings. Some of these robots will have the ability to damage property and people, even if programmed to behave safely. This work reveals a pressing need within the interaction design community to develop new interaction models and strategies for robot failure.

A concern going forward is how to properly calibrate bystander trust in robots that are capable of harming humans and to examine how this trust corresponds to a willingness to help the robots. It is somewhat alarming that approximately 60% of the participants previously exposed to personal risk then entered the robot's workspace to help the robot. Fatalities with industrial robots can and do occur as a direct result of human error and the decision to enter robot workspaces [24]. It is important to create systems that can signal to users whether safety has been compromised during a failure.

Also, research and design work on providing robot operators with suitable interfaces for managing failing robots is underway (e.g., [8, 3]), but most future interactions with robots will likely be bystanders who lack the ability or opportunity to access dials, displays, and controls. While robots are increasingly being designed to be human-safe and are gaining capability in complex, flexible environments, society's increased exposure to robots has the potential for dire consequences. This is especially true in cases where human bystanders lack an interface describing the robot's goals and planned motion. This exploratory effort is an attempt to motivate and help guide principled design research on this topic.

ACKNOWLEDGEMENT

This work was funded by the National Science Foundation (IIS-1552256, 1317989, 1734361). We also thank Jodi Forlizzi.

REFERENCES

[1] Obehioye Adubor, Rhomni St. John, and Aaron Steinfeld. 2017. Personal Safety is More Important Than Cost of Damage During Robot Failure. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. ACM, New York, NY, USA, 403–403. DOI: http://dx.doi.org/10.1145/3029798.3036649

[2] Kim Baraka, Stephanie Rosenthal, and Manuela Veloso. 2016. Enhancing human understanding of a mobile robot's state and actions using expressive lights. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 652–657.

[3] Daniel J. Brooks. 2017. *A Human-Centric Approach to Autonomous Robot Failures*. Ph.D. Dissertation. University of Massachusetts Lowell.

[4] Elizabeth Cha, Maja Matarić, and Terrence Fong. 2016. Nonverbal signaling for non-humanoid robots during human-robot collaboration. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 601–602.

[5] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 507–513. http://dl.acm.org/citation.cfm?id=3237383.3237459

[6] Kerstin Dautenhahn. 2013. Human-robot interaction. *The encyclopedia of human-computer interaction* (2013). https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction

[7] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, Kheng Lee Koay, and I. Werry. 2005. What is a robot companion - friend, assistant or butler?. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1192–1197. DOI: http://dx.doi.org/10.1109/IROS.2005.1545189

[8] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of Robot Failures and Feedback on Real-time Trust *(HRI '13)*. IEEE Press, Piscataway, NJ, USA, 251–258. http://dl.acm.org/citation.cfm?id=2447556.2447663

[9] Munjal Desai, Mikhail Medvedev, Marynel Vazquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggerman, Aaron Steinfeld, and Holly Yanco. 2012. Effects of Changing Reliability on Trust of Robot Systems. In *2012 7th ACM/IEEE International Conference on Human-robot Interaction (HRI '12)*. IEEE Press, 8.

[10] Carl F DiSalvo, Francine Gemperle, Jodi Forlizzi, and Sara Kiesler. 2002. All robots are not created equal: the design and perception of humanoid robot heads. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*. ACM, 321–326.

[11] Naomi T. Fitter and Katherine J. Kuchenbecker. 2016. Designing and Assessing Expressive Open-Source Faces for the Baxter Robot. In *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings (Lecture Notes in Artificial Intelligence)*, Vol. 9979. Springer International Publishing, 340–350. Oral presentation given by Fitter.

[12] Cecilia Gabriela Morales Garza. 2018. *Failure Is an Option: How the Severity of Robot Errors Affects Human-Robot Interaction*. Master's thesis. Carnegie Mellon University, Pittsburgh, PA. https://www.researchgate.net/publication/327688598_Failure_Is_an_Option_How_the_Severity_of_Robot_Errors_Affects_Human-Robot_Interaction

[13] Adriana Hamacher, Nadia Bianchi-Berthouze, Anthony G Pipe, and Kerstin Eder. 2016. Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 493–500.

[14] Shanee Sarah Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[15] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 895–906. DOI: http://dx.doi.org/10.1145/3196709.3196735

[16] Alisa Kalegina, Grace Schroeder, Aidan Allchin, Keara Berlin, and Maya Cakmak. 2018. Characterizing the Design Space of Rendered Robot Faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 96–104.

[17] Ross A Knepper, Stefanie Tellex, Adrian Li, Nicholas Roy, and Daniela Rus. 2015. Recovering from failure by asking for help. *Autonomous Robots* 39, 3 (2015), 347–362.

[18] Minae Kwon, Sandy H. Huang, and Anca D. Dragan. 2018. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. ACM, New York, NY, USA, 87–95. DOI: http://dx.doi.org/10.1145/3171221.3171276

[19] Amanda Lazar, Hilaire J Thompson, Anne Marie Piper, and George Demiris. 2016. Rethinking the design of robotic pets for older adults. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM, 1034–1046.

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

[21] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully Mitigating Breakdowns in Robotic Services. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction (HRI '10)*. IEEE Press, Piscataway, NJ, USA, 203–210. http://dl.acm.org/citation.cfm?id=1734454.1734544

[22] Gale M Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2018. Getting to Know Each Other: The Role of Social Dialogue in Recovery from Errors in Social Robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 344–351.

[23] Bonita Marlene Muir. 1989. *Operator's Trust in and Use of Automatic Controllers Supervisory Process Control Task*. Ph.D. Dissertation. University of Toronto, Room 301, 65 St. George Street.

[24] M Nagamachi. 1988. Ten fatal accidents due to robots in Japan. In *Proceedings of the First International Conference on Ergonomics of Hybrid Automated Systems I*. Elsevier Science Publishers BV, 391–396.

[25] Edwin Olson. 2011. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3400–3407.

[26] Samantha Reig, Selena Norman, Cecilia G Morales, Samadrita Das, Aaron Steinfeld, and Jodi Forlizzi. 2018. A Field Study of Pedestrians and Autonomous Vehicles. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 198–209.

[27] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of Robots in Emergency Evacuation Scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, Piscataway, NJ, USA, 101–108.

[28] Stephanie Rosenthal and Manuela Veloso. 2012. Mobile Robot Planning to Seek Help with Spatially-Situated Tasks.. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[29] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. 2017. How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. In *International Conference on Social Robotics*. Springer, 42–52.

[30] D. Rothenbücher, J. Li, D. Sirkin, B. Mok, and W. Ju. 2016. Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 795–802. DOI: http://dx.doi.org/10.1109/ROMAN.2016.7745210

[31] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. ACM, New York, NY, USA, 141–148. DOI: http://dx.doi.org/10.1145/2696454.2696497

[32] Satragni Sarkar, Dejanira Araiza-Illan, and Kerstin Eder. 2017. Effects of Faults, Experience, and Personality on Trust in a Robot Co-Worker. *arXiv preprint arXiv:1703.02335* (2017). http://arxiv.org/abs/1703.02335

[33] Kristin E. Schaefer, Jessie Y.C. Chen, James L. Szalma, and P.A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation. *Implications for Understanding Autonomy in Future Systems* (2016). https://doi.org/10.1177/0018720816634228

[34] L. Takayama, W. Ju, and C. Nass. 2008. Beyond dirty, dangerous and dull: What everyday people think robots should do. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 25–32. DOI:http://dx.doi.org/10.1145/1349822.1349827

[35] Daniel Ullman and Bertram F. Malle. 2017. Human-Robot Trust: Just a Button Press Away. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. ACM, New York, NY, USA, 309–310. DOI: http://dx.doi.org/10.1145/3029798.3038423

[36] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 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 (HRI '17)*. ACM, New York, NY, USA, 408–416. DOI: http://dx.doi.org/10.1145/2909824.3020230