Flailing, Hailing, Prevailing: Perceptions of Multi-Robot Failure Recovery Strategies

Samantha Reig
sreig@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Elizabeth J. Carter
ejcarter@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Terrence Fong
terry.fong@nasa.gov
NASA Ames Research Center
Moffett Field, CA, USA

Jodi Forlizzi
forlizzi@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Aaron Steinfeld
steinfeld@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

We explored different ways in which a multi-robot system might recover after one robot experiences a failure. We compared four recovery conditions: Update (a robot fixes its error and continues the task), Re-embody (a robot transfers its intelligence to a different body), Call (the failed robot summons a second robot to take its place), and Sense (a second robot detects the failure and proactively takes the place of the first robot). We found that trust in the system and perceived competence of the system were higher when a single robot recovered from a failure on its own (by updating or re-embodying) than when a second robot took over the task. We also found evidence that two robots that used the same socially interactive intelligence were perceived more similarly than two robots with different intelligences. Finally, our study revealed a relationship between how people perceive the agency of a robot and how they perceive the performance of the system.

CCS CONCEPTS

• **Human-centered computing** → **Laboratory experiments**.

KEYWORDS

failure, recovery, trust, re-embodiment, multi-robot interaction

ACM Reference Format:
Samantha Reig, Elizabeth J. Carter, Terrence Fong, Jodi Forlizzi, and Aaron Steinfeld. 2021. Flailing, Hailing, Prevailing: Perceptions of Multi-Robot Failure Recovery Strategies. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3434073.3444659

1 INTRODUCTION

Robots that work with humans will be prevalent before they are perfect. Hopefully, severe failures will be few and far between, but some breakdowns are inevitable. The use of social cues to communicate states, needs, and processes to humans is especially important during cases of failure, which can have lasting effects on perceived competence and trustworthiness [13, 14, 42]. Sometimes, these cues will be intended to solicit help from human collaborators (e.g., [17, 29, 42, 63]) or bystanders (e.g., [64]). In these cases, they will be critically important to both robot function and human-robot relationships. In other cases, robots may recover autonomously without seeking human intervention, but they will still need to communicate to humans to repair trust and relationships [34].

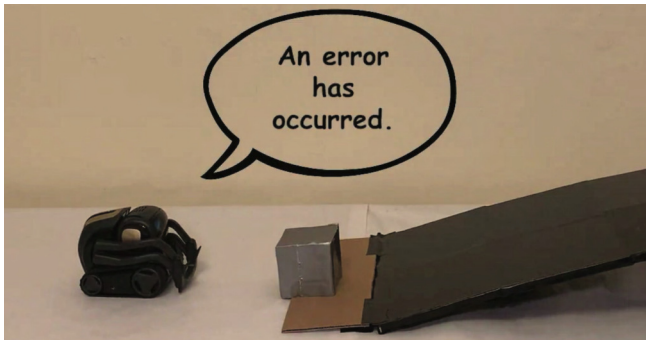

Figure 1: A robot drops a package at the bottom of a ramp.

When multiple robots work together, there may be cases in which a single robot experiences a failure from which it cannot recover sufficiently quickly (e.g., signal loss) or at all (e.g., severe hardware damage). One possibility is that the failure ends the task. However, it is also possible that the failed robot could find a way to resume the task (e.g., by downloading an update that improves its vision) or even hand the task off to another robot to complete. In these situations, will a violation of trust in the *robot system as a whole* be best repaired with a single robot that demonstrates resilience, or with a second robot that does not have the stain of a prior failure on its record? Could the software intelligence of the first robot migrate to ("re-embody" [36]) another physical embodiment to achieve the best of both worlds? It remains an open question how a multi-robot system's response to failure affects humans' perceptions of the robots, individually and collectively.

In this paper, we describe a protocol where participants watched multiple failures and recovery types as a robot system completed a task. We investigated the effects of these recovery types on trust in

HRI '21, March 8–11, 2021, Boulder, CO, USA.

ACM ISBN 978-1-4503-8289-2/21/03.
https://doi.org/10.1145/3434073.3444659

and social attributions to the system. We performed two studies that examined four possible recovery interactions. Our findings suggest that people may trust multi-robot systems more and perceive them as more competent if a single robot demonstrates resilience and self-repairs after a failure than if a second robot replaces the robot that failed. Additionally, if a robot *re-embodies* its artificial intelligence (AI) to another robot after failure, the system may be perceived more positively than a system with two independent robots. Our work advances the understanding of failure, trust, and agency in human-multirobot interaction and has implications for design.

2 RELATED WORK

2.1 Interactions among robots

Human-multirobot interaction is a growing area within human-robot interaction (HRI). Prior work suggests that the way robots interact with one another in the presence of humans can impact both task outcomes and social perceptions.

2.1.1 Goal-oriented outcomes. Our work explores a scenario where a team of two robots has a shared goal. Work on interdependence for human-robot teams (e.g., [25]) and shared mental models (e.g., [21, 22, 57]) has examined how robots with common goals can work effectively with human teammates. In one study [21], robots that used a shared mental model to share task- and team-related information with each other led to faster task completion time and higher neglect tolerance (a behavioral measure of trust) from participants. However, self-report measures of trust did not show an effect of the shared mental model. Other previous research examined the efficiency of cooperative repair on robot teams under various circumstances, but did not examine human responses [8]. Our scenario examines how people might react when robots dynamically reassign subtasks in response to failures.

2.1.2 Social outcomes. Numerous HRI papers have discussed how robot behavior can influence human behavior in groups. For example, robots that agree with each other can encourage humans to conform to their opinion [54], and vulnerable behavior by a robot can lead to "ripple effects" in which other team members have more trust-related interactions [59]. In group interactions, robots are perceived more positively when they behave prosocially [12, 48] (e.g., in ways that benefit the group, even at personal cost). When humans merely observe an interaction between robots, the way the robots treat each other impacts how humans perceive them. In some cases, robots that talk to each other overtly (rather than covertly) are perceived as more social, engaging, and easy to understand [61, 65]. However, observable communication among robots and agents that are not clearly distinguished, individual, social entities may also be seen as unnecessary [19] or unnerving [36]. This perceived entitativity of a group of robots also impacts social perceptions. When people perceived multiple robots as a single group entity rather than as individuals, they viewed the robots as a less-approachable, impermeable in-group [20]. Because re-embodying intelligence for robots that use social cues is a new paradigm for HRI, it is not clear how findings from prior work on implicit persuasion, overt and covert communication, and entitativity will apply. This question motivates our research.

2.2 Migrating software intelligence

One way to design the human-facing elements of coordinated behavior for multiple physical robots is to have the robots be controlled by a single AI. The concept of software intelligence migrating across physical platforms was originally proposed by Duffy and colleagues [15]. Their Agent Chameleon framework proposed an architecture in which software agents could move between virtual and physical environments as well as mutate (e.g., by gaining or losing a physical feature) within an individual environment. That framework also emphasized the importance of equipping agents with basic survival instincts: if an agent perceived that an external force (such as a dying battery) would soon cause it to be unable to function, it could either migrate or "save" its internal state to storage. The LIREC (Living with Robots and intEractive Characters) project [1] positioned migrating intelligence as a companion technology that could provide continual social support while offsetting the power costs of carrying a physically embodied agent from place to place [33]. Related work has studied appropriate and legible cues for identity migration [31, 37] and how relationships between humans and artifacts are mediated by migrating agents [45].

This idea has been extended by research testing the psychological boundaries of robots [27], determining children's concepts of migrating intelligences [60], and probing potential configurations of embodiments and intelligences [31]. Research in situated laboratory contexts has exhibited prototypes of migrating intelligence in robots inhabiting mock smart homes [30] and compared the effects of identity (i.e., behavior and personality) migration and information (i.e., data) migration [62]. Both [30] and [62] suggested that a persistent "identity" of an AI over time and embodiment is a crucial benefit of migrating intelligence. Re-embodiment is also a promising design for robots in service settings, where the same intelligence (or "personality" or "identity") can appear across the multiple touchpoints of a service to provide a seamless, comfortable, and personalized experience [36, 50].

2.3 Robot failures and trust

Characterizing, measuring, and creating trust in human-robot interaction is a key area of HRI research (e.g., [23, 55, 56]). Prior work shows that the timing, risk type, and severity of a failure can impact the degree of detrimental effects on trust. Desai and colleagues [13] found that early periods of low reliability damage trust more than late periods of low reliability. Adubor and colleagues [3] compared personal risk to property risk and found that people prioritize personal safety over financial cost. Humans may also "overtrust" robots. People continue to trust robots when given subpar information [38] and when asked to complete strange requests [9, 53]. Also, they may trust a robot that they have already seen malfunction [42], even in crisis situations [52]. Robots with social capabilities may also be able to use trust- and rapport-building to encourage people to conform to their suggestions and disclose information [5]. Some HRI work (e.g., [41, 47, 49]) has explored the pratfall effect—in which someone who is perceived as highly competent is more likeable if they make a minor mistake [4]—with mixed results.

The service design and HRI literatures have insight to offer pertaining to recovery from failure. In general, customers are more satisfied with services when they avoid failure than when they

experience failure and then recover [39]. Good communication is critically important during service breakdowns [7, 26]. For robots, effective strategies for recovery include giving notifications about the error [10] and acknowledging the difficulty of a task [13, 35, 43]. To our knowledge, HRI research has not yet explored how trust is impacted by robot re-embodiment after failure.

3 STUDY 1

To examine possible effects of recovery strategies with many participants, we designed an online study that showed videos of a package delivery scenario where robots carried boxes from point A to point B. This is similar to a paradigm from [28] in which participants cooperated with a delivery robot in an assembly task.

3.1 Study design

This study had a between-subjects design with four conditions (one video per condition). Each video involved a small robot attempting to carry a small, solid, grey cube (a "package") from a starting point to an ending point. We used two Vector [2] robots from Anki/Digital Dream Labs. These are small robots that have expressive, pixelated eyes and a bulldozer-like form. Each robot has a lift that is capable of picking up and placing down small objects. The robots used spoken natural language to explain what was happening. We also included speech bubbles to help participants understand the dialogue. We chose to use speech bubbles rather than captions because they could be placed next to the correct robot and thus be part of the scene.

All videos began the same way. First, the robot picked up the package and said, "Beginning package delivery." Then, it drove the package across a flat surface toward a ramp. At the bottom of the ramp, the robot swiveled back and forth, reversed, and put the package on the ground. It said, "Package dropped." After attempting to recover the package (by moving toward it and raising and lowering the lift), it reversed again, and declared: "Cannot recover package. Delivery failed. An error has occurred." Then, one of four recovery conditions was executed to complete delivery of the package.

- ***Update:*** *One intelligence, one robot. After a robot experienced a failure, it fixed the problem and then completed the task.* After acknowledging the error, the robot said, "Let me update my software," and drove back to the starting point. It then turned away from the camera and then back toward it, and it said, "The problem is fixed. I will not experience the same error again."
- ***Call:*** *Two intelligences, two robots. After a robot experienced a failure, it called a second robot that replaced the first one and completed the task.* After the first robot acknowledged the error, it said, "Let me call another robot," and drove back to the starting point. A second robot entered the frame, and said, "I will not experience the same error as the previous robot."
- ***Sense:*** *Two intelligences, two robots. After a robot experienced a failure, a second robot noticed the problem and replaced the first robot to complete the task.* After the first robot acknowledged the error, it drove back to the starting point. A second robot entered the frame and said, "I will take over from here. I will not experience the same error as the previous robot."
- ***Re-embody:*** *One intelligence, two robots. After a robot experienced a failure, it re-embodied (moved its intelligence to) a different physical robot to complete the task.* After acknowledging the error, the robot said, "Let me move my brain over to a better robot body," and drove back to the starting point. Its eyes and face went dark. A second robot entered the frame and said, "The problem is fixed. In this robot body, I will not experience the same error again."

At this point, the recovery robot (the same robot in the Update and Re-embody conditions; a second robot in the Call and Sense conditions) drove to the package, picked it up, and said, "Beginning package delivery." Then, it drove the package to the top of the ramp, placed it down, backed away from it, and said, "Delivery complete." All four conditions followed the exact same narrative up until the failure, and they resumed similar narratives after the recovery. Figure 1 shows an example of the failure event. Full videos are included in the Supplementary Materials.

A pilot study with 154 participants on Amazon Mechanical Turk confirmed that (1) the package drop was perceived as a failure; (2) a successful robot was perceived as more trustworthy ($F(1, 152) = 37.76, p < .0001$, Cohen's $d = .97$) and competent ($F(1, 152) = 17.78, p < .0001, d = .68$) than a failing robot; and (3) they accurately understood the speech.

3.2 Hypotheses

We predicted that the recovery method used after a failure would impact participants' trust. Prior work suggested that robots can recover from negative associations brought about by mistakes during sustained interactions using socially appropriate behaviors [35]. Prior work also suggested that re-embodiment is perceived as a desirable and efficient design [36, 50] and that identity migration positively impacts social perceptions [62]. Thus, we predicted:

- **H1** Participants will have higher **trust** in a robot system following a **Re-embody recovery** than following an Update recovery.
- **H2** Participants will perceive a robot system that uses a **Re-embody recovery** as most **competent**.

Research on groups and teams of robots (e.g., [21, 22, 40, 57]) informs our hypotheses regarding two-robot recoveries.

- **H3** Participants will have higher **trust** in a team of robots when the second robot **senses the first's failure** than when the first robot calls the second.
- **H4** Participants will perceive higher **competence** in a team of robots when the second robot **senses the first's failure** than when the first robot calls the second.

Our final hypothesis follows from the suggestions by previous work [46] that favorable social perceptions of robots increase willingness to work with them in the future.

- **H5** Participants will report a greater **desire to use** the system in the future when they perceive it to be more **warm and likeable**.

3.3 Measures

Our assessments included a mix of questions from prior work and questions written for this study. The response format of the closed-ended questions was 5-point (*attitudes toward robots in general*), 7-point (*trust*), and 9-point (*competence*, *warmth*, *likeability*) scales.

3.3.1 Validation questions. To confirm that participants perceived the failures and recoveries as intended, we asked open-ended questions about their interpretations of the robot's behavior during the task. We also included two attention checks that all passed.

3.3.2 Trust in the robot system. We evaluated trust through self-report measures. Participants were asked to answer several questions modified from the Jian scale [24] and a few additional questions that we created specifically for this study.

3.3.3 Social attributions to the robot system. We used a subset of the 18-item Robotic Social Attributes Scale (RoSAS) [11] to measure perceptions of *competence* and *warmth*. We analyzed both of these two factors and their individual items to examine more specific traits. To measure likeability, we used three Likert-type items inspired by words from the GODSPEED likeability subscale [6].

3.3.4 Attitudes toward robots. We included five Likert-type items to obtain judgments of overall trust in robots, perceived helpfulness of robots, interest in robots, and perceived personal importance and societal importance of robots. Four of these were modified from a scale proposed (but not validated) in prior work [51]. One, pertaining to overall trust, was new as of this work.

3.4 Procedure

Because some pilot responses from Amazon Mechanical Turk users suggested that people had glossed over some questions, we conducted the study on Prolific.co, which is a survey research platform with users who are used to longer-form studies. We described the task as gathering impressions of a prototype of a robotic package delivery system. Potential participants were redirected to Qualtrics for the study. After providing informed consent, participants were semi-randomly presented with one of the four videos (Update, Call, Sense, or Re-embody).[1] Below the video, participants were asked if the system experienced a failure and how it recovered from that failure. They then answered the questions about trust (presented in a random order), social attributes (in a random order), and attitudes toward robots. Then, they answered demographic questions, including about their age, gender, languages, employment, experience with computers and robots, and an open-ended question meant to capture additional demographic information. Finally, participants had the option to provide feedback about the study.

3.5 Participants

A total of 403 people participated in this study. There were 100 participants in the Update condition, 100 in Re-embody, 101 in Call, and 102 in Sense. To be eligible for the study, Prolific users had to be 18 years of age or older, be located in the U.S. or Canada, be proficient in English, and have a previous submission approval rate of at least 95%. Participants ranged in age from 18 to 78 years ($M = 31.25, SD = 10.89$). 162 were female, 234 male, 5 were other genders, and 1 did not specify a gender. They had a variety of professional backgrounds, including engineering, medicine, psychology, art, and sales. They generally had some experience using computers and little experience using AI personal assistants and robots (on a 7-point scale with 7 being more use, computers: $M = 6.70, SD = 0.70$; AI assistants: $M = 2.88, SD = 1.98$; robots: $M = 1.98, SD = 1.45$). 251 owned a pet, 257 owned an AI assistant, and 57 owned a robot. Participants took an average of 14 minutes to complete the study (min: 5, max: 45, median: 12) and were paid 2.50 USD each. Our study was approved by an Institutional Review Board.

[1]The video only allowed for pause and play; participants could watch the video more than once, but could not fast forward, rewind, or change the playback speed. Participants were told that they would only be able to watch the video straight through and that they could not proceed to the next questions until an amount of time equal to the video duration elapsed.

4 STUDY 1 RESULTS

Explanations of the failure and recovery accurately reflected the differences between the robot behavior in the different conditions, suggesting that the conditions were interpreted as intended. We analyzed the data using a linear model fit with REML.

The *trust* questions were correlated at Cronbach's $\alpha = .89$. The RoSAS *competence* items had $\alpha = .88$, and the *warmth* items had $\alpha = .90$. We treated these as factors. We analyzed *likeability* as an individual item because *meanness* and *friendliness* only weakly correlated with it. The *attitudes toward robots* questions correlated strongly ($\alpha = .85$) and were treated as a factor.

We included the *attitudes toward robots* questions to understand whether preexisting associations or biases had an effect on our dependent variables. In an exploratory analysis, we found that the factor had a significant effect on *trust*, *warmth*, perceived *competence*, and *likability*, $p < .0001$ for all variables. We placed these items at the end of our study rather than at the beginning in order to prevent priming the participants to rate the videos according to the immediate availability of their preexisting attitudes rather than our manipulation. We were concerned that the attitude questions could have been affected by our manipulation, thus invalidating attitude as an independent variable. We ran a nonparametric Wilcoxon rank sum/Kruskal-Wallis test to check for this. We did not find any significant effects of condition on attitudes (in fact, all means were $M = 3.7$). After confirming that it was not affected by condition, we included attitude in our model as a covariate. We used Tukey's Honest Significant Difference (HSD) test for post-hoc comparisons.

4.1 Trust in the robot system

We found a main effect of Recovery method on *trust*, $F(3, 395) = 3.16, p = .025$. Post-hoc pairwise tests revealed that trust was higher in the Update condition ($M = 4.03, SE = 0.10$) than in the Sense condition ($M = 3.67, SE = .10$). Because there are different dimensions of trust, we also looked at the individual items from the scale. We found a main effect of Recovery condition on perceptions that the system was *reliable*, $F(3, 395) = 2.71, p = .0345$. Post-hoc tests showed that the Re-embody recovery ($M = 4.19, SE = .14$) was rated higher than the Sense recovery ($M = 3.67, SE = .13$). We also found a main effect of Recovery condition on desire to use the system in the future, $F(3, 395) = 2.99, p = .031$, which was higher for Update ($M = 4.39, SE = .15$) than Sense ($M = 3.83, SE = 1.69$). We did not find trust differences between Update and Re-embody, so **H1 was not supported**. We also did not find any trust differences between the Call and Sense conditions, so **H3 was not supported**.

4.2 Perceived competence of the robot system

For perceived *competence*, we found a main effect of Recovery method, $F(3, 395) = 3.25, p = .022$. In particular, Update ($M = 5.81, SE = .14$) was perceived as more competent than Sense ($M = 5.22, SE = .14$). We also found an interaction effect of Recovery method and attitudes toward robots, $F(3, 395) = 3.31, p = .020$. Higher scores on the attitudes index combined with a Re-embody

recovery led to higher perceptions of competence, $p = .046$. This did not directly support H2, but it did suggest that re-embodiment was perceived as a more competent design by participants who had positive attitudes toward robots. We analyzed the individual items for the competence scale as well, and we found a main effect of Recovery condition on perceptions of the system as *knowledgeable*, $F(3, 395) = 3.56, p = .015$. Specifically, Re-embody ($M = 5.81, SE = .20$) was perceived as more knowledgeable than Sense ($M = 4.97, SE = .20$). Re-embody was higher than Sense, but not Call, and only on one item of the *competence* construct; this meant that **H2 was partially supported**. We did not find differences for competence between Call and Sense, so **H4 was not supported**.

4.3 Social attributions to the robot system

We did not find any effects of our manipulation on *warmth* or *likeability*. However, we found an interaction effect of Recovery method and attitudes toward robots on likeability, $F(3, 395) = 3.94, p = .009$. Higher attitudes scores combined with a Re-embody recovery led to higher likeability, $p = .023$. Desire to use the robot system in the future was moderately correlated with perceived *warmth*, $r = .37$ and with *likeability*, $r = .45$, both $p < .0001$, **supporting H5.**

5 STUDY 1 DISCUSSION

In Study 1, we predicted that a Re-embody recovery would result in the highest perceived trust and competence, and that Sense would be perceived as more trustworthy and competent than Call. Three of our hypotheses were not supported, and one received only partial support. In general, Re-embody was not an improvement over Update, and Sense was not an improvement over Call. Instead, the common thread across our findings was that Update was perceived most favorably, and particularly more favorably than Sense.

To explore possible explanations, we looked at the qualitative data, which consisted of reflections on the recovery, explanations of the trust and social attribute ratings, and general feedback. We noticed that participants anthropomorphized the robots (e.g., "He wants to update his software so he won't experience the same error again,"–P391) and viewed them as cute (e.g., "The voice was very cute and so were its little eyes,"–P51). However, they were not willing to associate robots with words meant to measure perceived *warmth* because "robots do not have emotions" (many participants). In particular, when participants saw two robots, they especially anthropomorphized the first robot and thought it "made you feel bad for the little guy when he failed" (P210). This endearing failure caused them to see the first robot more positively when it recovered. For example, P121 said, "It didn't get grumpy while experiencing an error but instead acted promptly and made an immediate effort to find a solution." P270 said, "I honestly thought the first robot looked very distressed [...] The little fella looked cute as hell and I was touched." In contrast, participants viewed the second robot negatively when it took over. P288 said, "I felt sad for the first robot." P258 said, "The second robot was 'mean' by dismissing the first robot, and I was weirdly almost rooting for it to fail."

We reason that participants anthropomorphized the first robot and then favored Update because it was the condition in which the first robot showed the most agency: it failed, was able to repair the error on its own, and then continued the task successfully. Conversely, in the Sense condition, the first robot had the least agency: it simply stopped and waited for another robot to come and take over. Besides forming an attachment to the first robot, participants also felt that the need for a second robot made the system as a whole less reliable. For example, P233 said, "Ideally, there should be no need to depend on a second robot," and P235 said, "The first robot should have made another attempt."

We also noticed a pattern where participants commented that they based their ratings of trust entirely on the fact that the first robot failed to deliver the package on the first try. For example, P7 said, "It looks like it's in early testing, and it doesn't seem too reliable as the first one failed the simple task." The timing of a trust violation influences changes in trust [13, 14]. In our study, there was no "burn-in period" for building up trust before the error occurred. It is possible that the effects of our manipulation were dwarfed by the effect of seeing only a single, failed first attempt at delivery.

Results may have also been impacted by participants taking the perspective of the package *recipient*, rather than that of someone who *worked with* the robots. Many participants mentioned that they would not be willing to trust the system enough to use it until it showed major technical improvement (e.g., "I'm not confident that it could be trusted in more complex, real-world settings,"–P317; "I would likely not use [it] in case of future errors that could not be automatically resolved,"–P365). Several participants mentioned concerns that the robot(s) would not be able to handle stairs (e.g., P60, P87, P140) or bad weather (e.g., P53, P209), or that packages would be subject to theft (e.g., P61, P91, P351). From the vantage point of an end-user who would only ever *see* such a system if it succeeded, people were hesitant to view it as trustworthy and competent if it could not successfully perform its task even once.

This study provided evidence that participants did not make social attributions to the robots despite anthropomorphizing them, that people generally preferred a one-robot recovery over a two-robot recovery, and that participants formed impressions of the robot(s) from the perspective of an end-user or customer rather than a collaborator. With these new insights, we conducted another study to better understand these findings.

6 STUDY 2 METHOD

We adapted the method from Study 1. We used the same videos, recruitment platform (Prolific), and survey template (in Qualtrics).

6.1 Methodological adjustments

In this section, we describe the changes from Study 1. Methods not described here (e.g., recruitment, consent) remained the same.

6.1.1 Scenario framing. We revised the introductory blurb for the study to invoke a collaboration with the robots rather than receiving a service. It read: *"In this study, you will learn about and watch videos of a prototype for a robotic package delivery system. Imagine that you work with the robots that are part of this system. You are responsible for managing them as they coordinate to deliver packages. Because of various obstacles in the environment, they sometimes fail, but they have protocols in place to resume the task after a failure."*

6.1.2 Within-subjects design. To further examine differences in perceptions and attributions between "one-intelligence" (Update

and Re-embody) and "two-intelligence" (Call and Sense) conditions, we used a within-subjects design. Each participant viewed all four conditions in a random order.[2] This also enabled us to ask participants to rank the four designs in order of preference.

6.1.3 Timing of the failure. We added a Baseline video in which a single robot successfully delivered the package on the first try. Thus, success was shown as a possibility and the first failure was not experienced as early. We expected this addition, along with the within-subjects design, to recalibrate participants' ratings of the system's trustworthiness and competence after recoveries.

6.1.4 Measures. The Study 1 findings about non-social treatment of the system as a whole, anthropomorphism of the first robot, and attributions of failure informed our measures for Study 2.

Trust questions. We used the Muir trust scale [43] rather than the Jian trust scale [24]. The wording of the questions in the Muir trust scale is less evocative of relational aspects of trust, which makes more sense for a study in which participants are not interacting with robots or viewing them socially. Prior work on failures in HRI has shown that both scales elicit similar ratings of trust [13].

Attribution of failure. We added a question about whether participants attributed the robot's failure to get up the ramp to a hardware problem, a software problem, both, or another problem. We asked this question for each condition.

Agency of the first robot. In Study 1, the RoSAS *warmth* subscale was subject to a floor effect: participants did not attribute the descriptions of words like "emotional" and "organic" to the robots they saw in the video. However, they did anthropomorphize the first robot in their qualitative descriptions, and this seemed to influence their perceptions of the two-robot conditions. Therefore, we replaced the RoSAS *warmth* subscale with measures of *agency* and *anthropomorphism*. We used analogical statements from Ezer's robot anthropomorphism instrument [16], items from Kozak et al.'s Mind Attribution Scale for perceptions of agency [32], and one new item ("The robot is capable of complex thought"). These instruments have been used in prior HRI work on robots in groups [18].

6.2 Hypotheses

We approached Study 2 with a novel set of hypotheses. Because the Study 1 results implied that perceptions of the whole system were primarily shaped by perceptions of the *first robot*, we predicted:

- **H6a** Participants will perceive a robot that experiences a failure to have more **agency** when it **recovers on its own** than when it requires help from another robot.
- **H6b** Participants will perceive a robot that experiences a failure to be more **competent** when it **recovers on its own** than when it requires help from another robot.
- **H6c** Participants will have higher **trust** in a robot system in which one robot recovers on its own than in a robot system that uses a two-robot recovery.
- **H7a** Participants will have a greater **desire to work with** a system in which they perceive a failing robot to have more **agency**.
- **H7b** Participants will **prefer** a robot system that recovers using the **same hardware and the same software**.

We also tested the suggestion from Study 1 that participants formed an attachment to and "rooted for" the first robot's AI:

- **H8** A failure that is recovered with a **re-embodiment** will be perceived as a **hardware problem** (rather than a software problem) more often than will a failure that is recovered by the same robot without a re-embodiment or by a second robot.

6.3 Participants

We recruited 130 participants for this study, none of whom participated in Study 1. Participants ranged in age from 18 to 64 ($M = 29.81, SD = 9.67$). 51 identified as female, 57 as male, 1 as nonbinary, and 1 as agender. As in the first study, many different personal and professional backgrounds were represented (e.g., engineering, law, science, retail), experience with computers was high ($M = 6.75, SD = 0.65$), and experience with AI personal assistants and robots was relatively low (AI personal assistants: $M = 2.52, SD = 1.81$; robots: $M = 1.76, SD = 1.08$). 59 owned a pet, 73 owned an AI personal assistant, and 15 owned a robot. Participants took an average of 38.2 minutes to complete the study (excluding one outlier) and were paid 5.00 USD each.

We excluded data from 20 participants who (a) failed the attention checks, (b) perceived the Baseline video to have a failure, (c) did not perceive one of the failures to be a failure (this would have interfered with the way their impressions changed across conditions), or (d) used a mobile device (we could not prevent scrubbing the video for mobile viewing). This left us with a total of 110 participants.

7 STUDY 2 RESULTS

The residuals were non-normally distributed, so we used Friedman tests and Wilcoxon signed-rank tests with a Bonferroni correction for post-hoc comparisons unless otherwise noted. Where possible, we report effect sizes with Kendall's W for Friedman tests and with r for post-hoc tests. We report sample medians as M.

The Muir *trust* scale had a Cronbach's $\alpha = .94$. The RoSAS *competence* items had $\alpha = .89$. We created a factor out of the analogical statements for *anthropomorphism*, which had $\alpha = .77$. Four of the five *agency* items had $\alpha = .77$. One of them, "The robot is capable of doing things on purpose", was only weakly correlated with the other items, so we excluded it from the agency factor.

7.1 Trust in the robot system

We found a main effect of Recovery method on *trust*, $\chi^2(4) = 98.8, p < .0001, W = .22$. Trust was significantly higher in Update ($M = 5.38$) than in Re-embody ($M = 4.81$), Call ($M = 4.75$), and Sense ($M = 4.38$), all $p < .0001, r > .48$. Trust was significantly higher in Re-embody than in Sense, $p < .0001, r = .45$, but there was no significant difference between Re-embody and Call. Also, trust for Call was significantly higher than for Sense, $p = .002, r = .35$. Finally, trust was lower in Call and Sense than in the Baseline ($M = 5.38$), $p < .0001$ ($r = .47$ and .65, respectively), and lower in Re-embody than in the Baseline, $p = .0007, r = .37$. Trust in the Update condition was not significantly different from Baseline. These results **support H6c**.

[2] Because the order was randomly chosen each time by our survey software, the 24 $\binom{4}{4}$ ordering conditions were not balanced. However, the number of times each Recovery condition occurred in each position was sufficiently distributed.

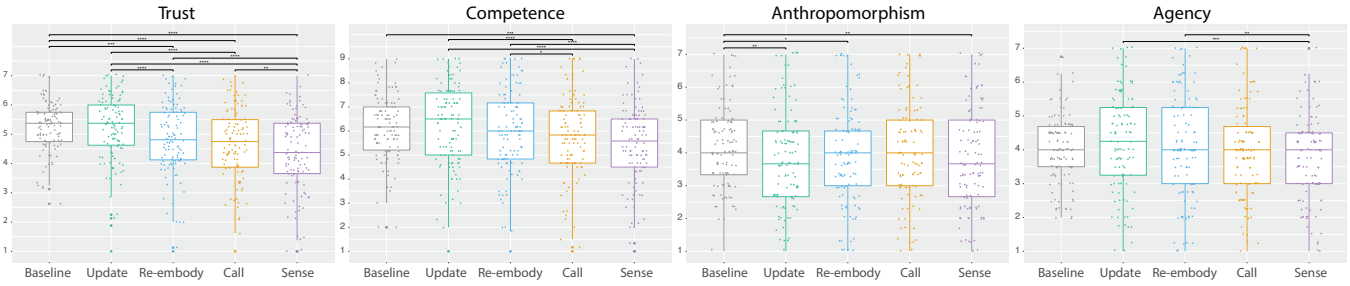

Figure 2: Box plots showing trust, competence, agency, and anthropomorphism for the Baseline video and each of the four Recovery conditions (Update, Re-embody, Call, Sense). Brackets marked * are significant at the .05 level, ** shows significance at the .01 level, * shows significance at the .001 level, and **** shows significance at the .0001 level.**

7.2 Perceived competence of the robot system

There was a small but significant main effect of Recovery method on perceived *competence*, $\chi^2(4) = 44.3, p < .0001, W = .10$. Specifically, perceived competence was significantly higher for Re-embody ($M = 6.00$) than for Call ($M = 5.83$), $p = .022, r = .29$, and for Sense ($M = 5.58$), $p < .0001, r = .46$, **supporting H6b**. Update ($M = 6.50$) had the highest rating and was also perceived as more competent than both Call and Sense, $p < .0001$, ($r = .44$ and $.53$, respectively), **supporting H6b**. There was no significant difference between Update and Re-embody, nor between Call and Sense.

7.3 Social attributions to the robot system

There was a small effect of Recovery on perceptions of the first robot's *agency*, $\chi^2(4) = 17.4, p = .0016, W = .04$. The robot was perceived to have more agency in Re-embody ($M = 4.00$) than in Sense ($M = 4.00$), $p = .002, r = .35$, and more agency in Update ($M = 4.20$) than in Sense, $p = .0002, r = .41$. There was also a small effect of Recovery on the *anthropomorphism* of the first robot, $\chi^2(4) = 22.90, p = .0001, W = .05$. The robot in Baseline ($M = 4.00$) was perceived as more anthropomorphic than the first robot in Re-embody ($M = 4.00$), Sense ($M = 3.67$), and Update ($M = 3.67$) ($r = .29, .36, .35$) but there were no significant differences between Baseline and Call ($M = 4.00$) or among the failure conditions. As such, **H6a was partially supported**. Desire to work with the system in the future moderately correlated with increased ratings of the first robot's anthropomorphism, Pearson's $r = .46, p < .0001$, and its agency, $r = .37, p < .0001$, **supporting H7a**.

7.4 Attributions of failure

We used Cochran's Q test to examine effects of Recovery condition on attributions of the failure, treating each possible attribution as a binary variable (1 if it was the participant's answer, 0 if it was not). There was a significant effect of Recovery on ratings of the failure as a hardware problem, $\chi^2(3) = 129.0, \eta^2 = .39$, as a software problem, $\chi^2(3) = 178.0, \eta^2 = .54$, as both, $\chi^2(3) = 60.9, \eta^2 = .18$, and as other, $\chi^2(3) = 24.8, \eta^2 = .08$, all $p < .0001$. We used pairwise McNemar tests for post-hoc comparisons. The failure was attributed to a hardware problem significantly more in the Re-embody condition ($n = 66$) than in the Update condition ($n = 1$), $p < .0001$. We also found that the failure was attributed to a hardware problem significantly more in Re-embody than in Call ($n = 23$), $p < .0001$ and Sense ($n = 13$), $p < .001$. These results **supported H8**.

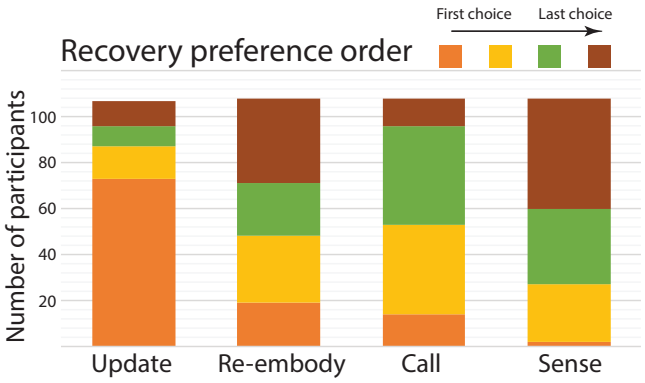

Figure 3: Most Study 2 participants rated Update as their first choice. Re-embody and Sense were commonly ranked last.

7.5 Preference

A majority of participants ($n = 73$) ranked Update as their most-preferred recovery (Figure 3), followed by Re-embody ($n = 19$), Call ($n = 14$), and Sense ($n = 2$). Most participants ($n = 48$) ranked Sense as their last choice. Interestingly, Re-embody was also frequently the least-preferred recovery ($n = 37$). **H7b was supported.**

7.6 Other findings

We conducted additional exploratory analyses to look for effects of Recovery condition on the individual analogical statement items from [16]. We used the Skillings-Mack test to look for effects on on perceptions that the first robot was *like a pet* and *like a teammate* because some values were missing. We used Friedman's test for *like an assistant*. There was a main effect of Recovery on perceptions of the first robot as a pet, $\chi^2 = 9.65, p = .047$. Post-hoc Wilcoxon signed-rank tests revealed that the first robot was perceived as more like a pet in the Baseline than in Call, $p = .041, r = .30$, Re-embody, $p = .006, r = .32$, and Update, $p = .015, r = .29$, but not in Sense. There was also a small effect of Recovery on perceptions of the first robot as an assistant, $\chi^2 = 43.7, p < .0001, W = .10$. Ratings were higher for the Baseline than for all four failure conditions, all

$p < .001$, $.39 < r < .50$. There was no effect of Recovery condition on perceptions that the first robot was like a teammate.

8 GENERAL DISCUSSION

The recovery strategies we tested compared a single-robot-single-AI recovery (Update condition), a multi-robot-single-AI recovery (Re-embody), and two forms of multi-robot-multi-AI recoveries (Call and Sense). We approached these two studies expecting to see a pattern in which the recoveries with more-efficient designs would be perceived more favorably. Instead, we found that people "rooted for" a robot that had failed: they perceived the system to be more trustworthy and competent in the single-AI Update and Re-embody conditions than in the Call and Sense conditions.

It is interesting that attachment to a single robot and perceptions of agency played a role in shaping trust and perceived competence despite relatively low ratings of warmth (Study 1) and anthropomorphism (Study 2). This suggests that people viewed the robots through a social lens despite claiming to consider them functionally. The field of HRI has long known that humans can form and benefit from bonds with machines despite *knowing* that they are machines that do not themselves have feeling. Nass' famous Computers Are Social Actors theory emphasized that social treatment of machines impacts human-machine relationships and occurs *independently* of true mind attribution and even anthropomorphism [44]. It follows that when robots experience damage or fail, their human partners will emotionally invest in their recovery. In fact, this has been reported in stories of soldiers whose life-saving robots have been damaged [58]. Our results demonstrate a type of preference or attachment for the first robot to attempt recovery even in a non-interactive scenario. This raises an interesting question about how to rebuild trust in robots after failure and the relationships among failure recovery and form, agency, and anthropomorphism.

Taken together, our findings suggest that the software update recovery was perceived most positively overall. However, the re-embodiment condition—in which the same interactive AI continued the task by moving into a different physical robot—was a fairly close second on many outcomes (see Figure 2). This has implications for specialized, goal-oriented, and high-risk environments: Robots that work closely with humans in task-oriented settings might be designed to take on a social "software identity" that can persist across embodiments to maintain trust after unexpected errors and failures. Relatedly, in Study 1, individual differences influenced how positively participants responded to the re-embodiment recovery. It is likely that the impact of re-embodiment recoveries on trust repair and human-robot relationships varies according to other individual differences as well. Socially interactive robots can be designed to behave differently when recovering after a failure depending on task domain, team dynamics, and personal traits of the current user(s). This is an opportunity area for future research.

8.1 Limitations and future work

Our study was conducted on one recruitment platform with a relatively small sample from the U.S. and Canada. The perspectives in our results may be limited by the sample's demographics, and our findings may not generalize to other populations. All of our findings were based on self-report measures, which do not always correspond to behavioral metrics meant to assess similar variables (e.g., objective and subjective trust measures do not always correlate).

Additionally, it is possible that aspects of our video stimuli not related to the manipulation impacted the results. Making videos that varied only by the minimum amount of dialogue and robot movement necessary to differentiate the recovery strategies was an intentional choice to minimize possible confounds. However, it is possible that the videos were too alike, especially in the Call and Sense conditions, for participants to find them noticeably different. The use of the word "software" in the Update condition and "brain" in the Re-embody condition may have impacted perceptions of anthropomorphism, and results more generally. We intended for the Sense condition to be interpreted as one robot proactively helping another after detecting its failure, but participants may have instead interpreted this as the first robot implicitly summoning the second. A stronger signal of a proactive response by the second robot might have drawn a starker contrast between Call and Sense, which were perceived overall similarly in both of our studies.

We also used robots that were small and toy-like, and which many participants called "cute". Although the robots had a functional form, their expressive eyes, high-pitched voices, and use of natural language likely raised expectations about anthropomorphism. The study results might have been markedly different had we used a different robot. Even with the Vector robots, we might have seen different patterns if the robots' eyes had been hidden, or if the state had been conveyed through different signals (e.g., as simple messages on a scrolling text log).

Finally, our study is limited in that it sought insight into human-multirobot interaction but did not involve in-person human interaction with real robots. We found that supplementing our closed-ended survey questions with open-ended ones was particularly useful given this setup. Analyzing short-answer explanations of closed-ended questions facilitated the discovery of qualitative insights that might have emerged through interviews or observations in an in-person, laboratory setting. These insights helped us develop Study 2, which was instrumental to the conclusions we drew from this research. Still, future work is needed to examine how people react and respond to multi-robot failures and recoveries during real-life interactions. Additionally, future work should further explore the relationship among perceived agency, robot resilience, and trust that we began to identify in this work.

9 CONCLUSION

Real-world human-robot interaction is messy, and failures are bound to occur. When these failures happen, a robot's immediate response can have critical and lasting effects on people's perceptions. Multi-robot systems have a number of options for how to recover from failures in ways that repair trust and other aspects of human-robot relationships. Our findings have implications for human-robot interaction design during instances of failure as well as for human-multirobot-interactions more broadly.

10 ACKNOWLEDGMENTS

This work was supported by NSF grant SES-1734456 and NASA grant 80NSSC19K1133. We also thank Zhi Tan.

REFERENCES

[1] [n. d.]. Exploring and designing our future robot companions | Lirec *. http://lirec.eu/

[2] [n. d.]. Meet Vector. https://www.digitaldreamlabs.com/pages/meet-vector

[3] 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*. 403–403.

[4] Elliot Aronson, Ben Willerman, and Joanne Floyd. 1966. The effect of a pratfall on increasing interpersonal attractiveness. *Psychonomic Science* 4, 6 (1966), 227–228. Publisher: Springer.

[5] A. M. Aroyo, F. Rea, G. Sandini, and A. Sciutti. 2018. Trust and Social Engineering in Human Robot Interaction: Will a Robot Make You Disclose Sensitive Information, Conform to Its Recommendations or Gamble? *IEEE Robotics and Automation Letters* 3, 4 (Oct. 2018), 3701–3708. https://doi.org/10.1109/LRA.2018.2856272 Conference Name: IEEE Robotics and Automation Letters.

[6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (Jan. 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[7] Chip R. Bell and Ron E. Zemke. 1987. Service breakdown: the road to recovery. *Management review* 76, 10 (1987), 32. Publisher: American Management Association.

[8] Curt Bererton and Pradeep Khosla. 2002. An analysis of cooperative repair capabilities in a team of robots. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, Vol. 1. IEEE, 476–482.

[9] Serena Booth, James Tompkin, Hanspeter Pfister, Jim Waldo, Krzysztof Gajos, and Radhika Nagpal. 2017. Piggybacking robots: Human-robot overtrust in university dormitory security. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 426–434.

[10] Daniel J. Brooks, Dalton J. Curtin, James T. Kuczynski, Joshua J. Rodriguez, Aaron Steinfeld, and Holly A. Yanco. 2020. Chapter 14 - A communication paradigm for human-robot interaction during robot failure scenarios. In *Human-Machine Shared Contexts*, William F. Lawless, Ranjeev Mittu, and Donald A. Sofge (Eds.). Academic Press, 277–306. https://doi.org/10.1016/B978-0-12-820543-3.00014-6

[11] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262.

[12] Filipa Correia, Samuel F. Mascarenhas, Samuel Gomes, Patrícia Arriaga, Iolanda Leite, Rui Prada, Francisco S. Melo, and Ana Paiva. 2019. Exploring prosociality in human-robot teams. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 143–151.

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

[14] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. 2012. Effects of changing reliability on trust of robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12)*. Association for Computing Machinery, New York, NY, USA, 73–80. https://doi.org/10.1145/2157689.2157702

[15] Brian R. Duffy, Gregory MP O'Hare, Alan N. Martin, John F. Bradley, and Bianca Schon. 2003. Agent chameleons: Agent minds and bodies. In *Proceedings 11th IEEE International Workshop on Program Comprehension*. IEEE, 118–125.

[16] Neta Ezer. 2008. *Is a robot an appliance, teammate, or friend? Age-related differences in expectations of and attitudes toward personal home-based robots*. PhD Thesis. Georgia Institute of Technology.

[17] Terrence Fong. 2001. Collaborative Control: A Robot-Centric Model for Vehicle Teleoperation. *Technical Report CMU-RI-TR-01-34*. (2001), 198.

[18] Marlena R. Fraune, Benjamin C. Oisted, Catherine E. Sembrowski, Kathryn A. Gates, Margaret M. Krupp, and Selma Šabanović. 2020. Effects of robot-human versus robot-robot behavior and entitativity on anthropomorphism and willingness to interact. *Computers in Human Behavior* 105 (2020), 106220. Publisher: Elsevier.

[19] Marlena R. Fraune and Selma Šabanović. 2014. Negative attitudes toward minimalistic robots with intragroup communication styles. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 1116–1121.

[20] Marlena R. Fraune, Selma Šabanović, Eliot R. Smith, Yusaku Nishiwaki, and Michio Okada. 2017. Threatening flocks and mindful snowflakes: How group entitativity affects perceptions of robots. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 205–213.

[21] Felix Gervits, Terry W. Fong, and Matthias Scheutz. 2018. Shared Mental Models to Support Distributed Human-Robot Teaming in Space. In *2018 AIAA SPACE and Astronautics Forum and Exposition*. 5340.

[22] Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. 2020. Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 429–437.

[23] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie YC Chen, Ewart J. De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527. Publisher: Sage Publications Sage CA: Los Angeles, CA.

[24] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[25] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna Van Riemsdijk, and Maarten Sierhuis. 2014. Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction* 3, 1 (March 2014), 43. https://doi.org/10.5898/JHRI.3.1.Johnson

[26] Scott W. Kelley, K. Douglas Hoffman, and Mark A. Davis. 1993. A typology of retail failures and recoveries. *Journal of Retailing; Greenwich* 69, 4 (1993), 429. https://search.proquest.com/docview/228647115/abstract/C8F62E3641184CBFPQ/1

[27] Chyon Hae Kim, Yumiko Yamazaki, Shunsuke Nagahama, and Shigeki Sugano. 2013. Recognition for psychological boundary of robot. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 161–162.

[28] Taemie Kim and Pamela Hinds. 2006. Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Univ. of Hertfordshire, Hatfield, UK, 80–85. https://doi.org/10.1109/ROMAN.2006.314398

[29] 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. Publisher: Springer.

[30] Kheng Lee Koay, Dag Sverre Syrdal, Wan Ching Ho, and Kerstin Dautenhahn. 2016. Prototyping realistic long-term human-robot interaction for the study of agent migration. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 809–816.

[31] Kheng Lee Koay, Dag Sverre Syrdal, Michael L. Walters, and Kerstin Dautenhahn. 2009. A user study on visualization of agent migration between two companion robots. In *In Thirteenth International Conference on Human Computer Interaction*.

[32] Megan N. Kozak, Abigail A. Marsh, and Daniel M. Wegner. 2006. What do i think you're doing? Action identification and mind attribution. *Journal of Personality and Social Psychology* 90, 4 (2006), 543–555. https://doi.org/10.1037/0022-3514.90.4.543

[33] Michael Kriegel, Ruth Aylett, Pedro Cuba, Marco Vala, and Ana Paiva. 2011. Robots meet IVAs: a mind-body interface for migrating artificial intelligent agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 282–295.

[34] 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*. 87–95.

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

[36] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 633–644.

[37] Alan Martin, Gregory MP O'hare, Brian R. Duffy, Bianca Schön, and John F. Bradley. 2005. Maintaining the identity of dynamically embodied agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 454–465.

[38] Erika Mason, Anusha Nagabandi, Aaron Steinfeld, and Christian Bruggeman. 2013. Trust during robot-assisted navigation. In *2013 AAAI Spring Symposium Series*.

[39] Michael A. McCollough, Leonard L. Berry, and Manjit S. Yadav. 2000. An Empirical Investigation of Customer Satisfaction after Service Failure and Recovery. *Journal of Service Research* 3, 2 (Nov. 2000), 121–137. https://doi.org/10.1177/109467050032002 Publisher: SAGE Publications Inc.

[40] Lanssie Mingyue Ma, Terrence Fong, Mark J. Micire, Yun Kyung Kim, and Karen Feigh. 2018. Human-Robot Teaming: Concepts and Components for Design. In *Field and Service Robotics*, Marco Hutter and Roland Siegwart (Eds.). Vol. 5. Springer International Publishing, Cham, 649–663. https://doi.org/10.1007/978-3-319-67361-5_42 Series Title: Springer Proceedings in Advanced Robotics.

[41] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21. Publisher: Frontiers.

[42] Cecilia G. Morales, Elizabeth J. Carter, Xiang Zhi Tan, and Aaron Steinfeld. 2019. Interaction Needs and Opportunities for Failing Robots. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. ACM, San Diego CA USA, 659–670. https://doi.org/10.1145/3322276.3322345

[43] Bonita Marlene Muir. 1989. *Operators' Trust in and Use of Automatic Controllers in a Supervisory Process Control Task*. University of Toronto. Google-Books-ID:

lESrjgEACAAJ.

[44] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[45] Kohei Ogawa and Tetsuo Ono. 2008. ITACO: Effects to interactions by relationships between humans and artifacts. In *International Workshop on Intelligent Virtual Agents*. Springer, 296–307.

[46] Benjamin C. Oistad, Catherine E. Sembroski, Kathryn A. Gates, Margaret M. Krupp, Marlena R. Fraune, and Selma Šabanović. 2016. Colleague or Tool? Interactivity Increases Positive Perceptions of and Willingness to Interact with a Robotic Co-worker. In *International Conference on Social Robotics*. Springer, 774–785.

[47] Christine Packard, Tayler Boelk, James Andres, Chad Edwards, Autumn Edwards, and Patric R. Spence. 2019. The Pratfall Effect and Interpersonal Impressions of a Robot that Forgets and Apologizes. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Daegu, Korea (South), 524–525. https://doi.org/10.1109/HRI.2019.8673101

[48] Ana Paiva, Fernando P. Santos, and Francisco C. Santos. 2018. Engineering prosociality with autonomous agents. In *Thirty-second AAAI conference on artificial intelligence*.

[49] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O. Arras. 2016. Errare humanum est: Erroneous robots in human-robot interaction. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 501–506.

[50] Samantha Reig, Michal Luria, Janet Z. Wang, Danielle Oltman, Elizabeth Jeanne Carter, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2020. Not Some Random Agent: Multi-person Interaction with a Personalizing Service Robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, Cambridge, United Kingdom, 289–297. https://doi.org/10.1145/3319502.3374795

[51] 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*. 198–209.

[52] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 101–108.

[53] 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 *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 1–8.

[54] Nicole Salomons, Michael van der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. 2018. Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Chicago IL USA, 187–195. https://doi.org/10.1145/3171221.3171282

[55] Kristin E. Schaefer, Jessie YC Chen, James L. Szalma, and Peter A. Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400. Publisher: Sage Publications Sage CA: Los Angeles, CA.

[56] Kristin E. Schaefer, Tracy L. Sanders, Ryan E. Yordon, Deborah R. Billings, and Peter A. Hancock. 2012. Classification of robot form: Factors predicting perceived trustworthiness. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 56. SAGE Publications Sage CA: Los Angeles, CA, 1548–1552. Issue: 1.

[57] Matthias Scheutz, Scott A. DeLoach, and Julie A. Adams. 2017. A framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making* 11, 3 (2017), 203–224. Publisher: Sage Publications Sage CA: Los Angeles, CA.

[58] Peter Warren Singer. 2009. *Wired for war: The robotics revolution and conflict in the 21st century*. Penguin.

[59] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 178–186.

[60] Dag Sverre Syrdal, Kheng Lee Koay, Michael L. Walters, and Kerstin Dautenhahn. 2009. The boy-robot should bark!-children's impressions of agent migration into diverse embodiments. In *Proceedings: New Frontiers of Human-Robot Interaction, a symposium at AISB*.

[61] Xiang Zhi Tan, Samantha Reig, Elizabeth J. Carter, and Aaron Steinfeld. 2019. From one to another: how robot-robot interaction affects users' perceptions following a transition between robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 114–122.

[62] Ravi Tejwani, Felipe Moreno, Sooyeon Jeong, Hae Won Park, and Cynthia Breazeal. 2020. Migratable AI. *arXiv:2007.05801 [cs]* (July 2020). http://arxiv.org/abs/2007.05801 arXiv: 2007.05801.

[63] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. (2014). Publisher: Robotics: Science and Systems Foundation.

[64] Astrid Weiss, Judith Igelsböck, Manfred Tscheligi, Andrea Bauer, Kolja Kühnlenz, Dirk Wollherr, and Martin Buss. 2010. Robots asking for directions—The willingness of passers-by to support robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 23–30.

[65] Tom Williams, Priscilla Briggs, and Matthias Scheutz. 2015. Covert robot-robot communication: Human perceptions and implications for human-robot interaction. *Journal of Human-Robot Interaction* 4, 2 (2015), 24–49.