Analysis of Reactions Towards Failures and Recovery Strategies for Autonomous Robots

Daniel J. Brooks, Momotaz Begum, and Holly A. Yanco

Abstract— **Human-robot interaction involving the failure of autonomous robots is not yet well understood. We conducted two online surveys with a total of 1200 participants in which people assessed situations where an autonomous robot experienced different kinds of failure. This information was used to construct a measurement scale of people's reaction to failure where positive values correspond with increasingly positive reactions and negative values with negative reactions. We then used this scale to compare different kinds of failure situations, including the severity of the failures, the context risk involved, and the effectiveness of different kinds of recovery strategies. We found evidence that the effectiveness of recovery strategies depends on the task, context, and severity of failure.**

I. INTRODUCTION

Fully autonomous robots are progressively becoming capable of operating in the unstructured environments of everyday life. To date, very few fully autonomous robots have been deployed outside of the industrial sector, where people's access to such robots is usually restricted during operation to prevent injuries. In contrast, robots such as Rethink Robotics' Baxter, Aethon's TUG, and Google's self driving cars are designed to operate in the presence of people and interact with them. The expanding presence of such systems will dramatically increase the occurrences of non-expert human-robot interactions as self-driving cars, delivery drones, robot vacuums, and more become integrated into society.

The impending proliferation of autonomous robots raises questions about what happens when they malfunction. Our ability to build dependable systems is constantly improving thanks to research in sensor technology, artificial intelligence, error detection, fault tolerant software architectures, and fault prevention [1], [2]. However, even the most reliable of these systems will not be immune to occasional failures, and the manner in which they fail can seriously effect users' perception of those systems and the services they provide. Despite this, relatively little work has focused on investigating human-robot interactions involving failure. For example, it is still unknown whether people can reliably tell if a robot is operating properly or failing, how they will react or behave after encountering a failing robot, or what can be done to mitigate feelings of frustration, anxiety, anger, disgust, or resentment that might result. What we do know is studies indicate that failure by a robotic service make the robot seem less capable, lowers users' trust, and can make people reluctant to use the service again [3], [4], [5].

This work was supported in part by the National Science Foundation (IIS-1552228). The authors are with the Department of Computer Science at the University of Massachusetts Lowell, Lowell, MA, United States
{dan, mbegum, holly}@cs.uml.edu

This raises questions about what can be done to mitigate the consequences after a failure occurs, referred to as recovery strategies. Providing users with advanced warnings of potential problems has been shown to improve users' evaluation of a system after a failure, and activities such as offering an apology can sometimes make the robot seem more competent [5]. Analysis of real-time user trust with an autonomous robot found that the robot could provide the operator with confidence feedback on its current performance to encourage better control allocation without altering the user's level of trust in the system [4]. Researchers have explored having robots seek out nearby people to ask for help [6]. Work on generating failure-specific natural language requests for help based on the robot's task indicated that users had a more enjoyable experience compared to more generic methods of requesting help [7].

Unfortunately, not all recovery strategies work the way they are intended. If people are not made aware of why a robot is behaving in a particular way it can lead to confusion. In one such case, workers at a hospital were documented blaming each other for having "messed up" an autonomous delivery robot after observing inexplicable behavior [8]. However, providing users with information about the cause of a failure could also make the situation worse. Experiments have shown that users respond very negatively when a robot blames them for causing a failure, compared to when blame was collectively assigned to both the user and robot as a team (e.g. using "we" statements), even in cases where the human was likely aware that they were the primary source of the problem [9]. That said, having the robot blame anyone (even itself) for a failure has been shown to cause users to lose trust in the system [10].

Taxonomies have been described by [11] and [12] which categorize faults and provide insight into the many complex ways a system could fail. Attributes of a "dependable" system have been described by [1] as availability, reliability, safety, confidentiality, integrity, and maintainability. Concepts taken from consumer market research have been shown to have analogous effects in robotic services [5]. However, to the best of our knowledge there is no theoretical model that characterizes failures of autonomous robots in order to predict people's reaction to various situations.

In this paper, we demonstrate a method for comparing the detrimental impact of various failures and how effective different types of recovery strategies are at mitigating the resulting negative effects, as perceived by users. We performed a survey experiment looking at different types of failures occurring in various situations. This information was

used to construct a measurement scale of people's reaction to failure, which was then used to compare how the severity of the failures, the context risk involved, and the effectiveness of recovery strategies impact people's reactions. For the purposes of this experiment, we grouped recovery strategies into two categories, *task support* and *human support*.

One of the consequences of a failure occurring in a fully autonomous robot system is a deterioration in the task performance, possibly to the point that the task can no longer be performed. However, an autonomous robot may still be able to take actions that can assist in furthering the task towards completion even in conditions where a failure has rendered the system incapable of carrying it out on its own. We call recovery strategies using proactive behaviors taken by a failed robot that continue to support the completion of the task for which the human operator is responsible *task support*.

A robot operator's situation awareness (SA) is their ability to perceive information related to the state of the system and its surroundings, comprehend that knowledge within the robot's current context, and project or anticipate future events [13]. As autonomy increases, the time the system can run while being ignored by the operator (known as neglect-time) also increases [14]. This decreases the amount of attention the operator pays to the system at any given moment until ultimately the system is considered unsupervised. When a problem occurs in such a system, the person or people responsible for the robot's operation find themselves lacking sufficient SA to either understand the current problem or identify the appropriate actions that need to be taken - a phenomenon known as the out-of-the-loop problem [15]. When an autonomous system has been designed to provide information to people that supports or improves their SA with respect to the failure and the status of the task being performed, we say the system is providing *human support*.

II. EXPERIMENT

A previous investigation of failure mitigation strategies looked at using recovery strategies from the context of consumer research to improve users' satisfaction after a robot fails [5]. This included giving users advanced warning that the robot might fail due to the difficulty of a task, having the robot apologize, offering compensation (such as a refund), and offering alternative options. Their success with these techniques may be related to attribution theory - that consumers try to infer the cause of a failure, and their conclusions drive their expectations for how a situation should be handled [16]. This can lead to further dissatisfaction if the way a situation is handled does not match the consumer's expectations [17], and suggests that satisfaction with how a situation is handled can be controlled by ensuring that people have good situation awareness about the cause of failure.

Hypothesis 1: *Providing human support will help mitigate the negative effects caused by failure.*

When using a fully autonomous robot, an operator entrusts a task or responsibility to the system which they expect to be carried out. The relationship between the operator and the system can be thought of as a form of delegation since many tasks require the use of some level of discretion while being carried out. Thus, behaviors that work towards the completion of the task should be viewed favorably, especially if the robot is otherwise unable to complete the task itself.

Hypothesis 2: *Providing task support will help mitigate the negative effects caused by failure.*

Human and task support could have unintended consequences. Human support implemented using speech could result in unrealistic expectations that the robot is also capable of some form of task support [3]. Moreover, performing task support without providing sufficient human support could cause confusion. Combining the two techniques should minimize these kinds of problems without negative side effects.

Hypothesis 3: *A combination of both human and task support will help mitigate the negative effects caused by failure.*

As the negative effects of a failure are reduced, positive sentiments towards the robotic service should increase.

Hypothesis 4: *Recovery strategies which reduce the negative effects of a failure will also increase the likelihood of users wanting to use the system again.*

A. Survey Design

We conducted two between-subjects survey studies, approved by the Institutional Review Board at the University of Massachusetts Lowell, to test our hypotheses. Our studies were modeled on the technique used in [5]. Participants were presented with a short two part story about a fictional character "Chris" who in one survey used a vacuum cleaner robot and in the other a self-driving taxi. The first part gave a brief background of Chris and included a short history of Chris' previous experience with the robot (reported in a positive manner). The second part described Chris' most recent encounter with the robot and the results of that interaction.

B. Independent Variables

Four independent variables were manipulated in this study: *context risk*, *failure severity*, *task support*, and *human support*. *Context risk* (risk) referred to how undesirable a failure by the robot would be in a particular context or setting, and was either "high" or "low." *Failure severity* (severity) referred to the type of failure the robot experienced and the extent to which it would be an inconvenience. It was either "none" (no failure occurs), "low," or "high." The robot either had *task support* and/or *human support* capabilities, or it did not. Combinations that did not involve failure but included *task support* were not tested, as we were unable to conceptualize any scenarios in which this combination made sense. These variables were combined into twenty survey conditions for each robot scenario.

The variables were represented in the story text in different forms in order to make the scenarios realistic. In the vacuum scenario, Chris was simply experimenting with new settings on the robot to expand the area it would clean for "low" context risk. For "high" risk, Chris was portrayed as a "neat-freak" relying on the robot to clean the house before having guests arrive, despite having never previously attempting this. When the failure severity was "None," the vacuum worked as Chris intended it to. "Low" failure severity was manifested

by the robot not having enough battery to complete the job and Chris returning home to find the floors only partially cleaned. Finally, "High" failure severity depicted the robot creating an additional mess by knocking over a house plant. In scenarios where the robot did not have enough battery to complete cleaning, task support allowed the robot to return to its charger and later (some time after Chris had returned home) resume cleaning from where it left off. The robot without task support would simply clean as long as possible until it ran out of batteries and died in the middle of the floor. In scenarios where the robot knocked over the house plant, the robot with task support would continue cleaning but avoid the area immediately around the accident so as not to make matters worse. In contrast, the robot without task support would attempt to drive through the area resulting in further damage to the plant (tearing off leaves) and spreading mud around the carpet. Human support was implemented by allowing the robot to send status updates about its progress to Chris. The method by which the robot communicated was intentionally omitted and left to the reader's imagination, with the exception of the robot being depicted as able to remotely notify Chris at work.

For the taxi scenario, Chris was going to the grocery store for "low" risk and to the airport to catch a flight for "high" risk. When the failure severity was "none," the vehicle worked exactly as Chris anticipated. During the "low" severity condition, the vehicle attempts to pass a slowly moving vehicle ahead of it while on the highway and misses the exit it was supposed to take. In the "high" severity condition, severe weather interrupts the vehicle's ability to drive and it pulls over on the side of the road. Task support during "low" severity conditions has the vehicle reroute along the next fastest available route to the destination. Without task support the vehicle reroutes itself to turn around and go back to location it originally got off route at, despite a faster route being available. In the "high" severity condition, the vehicle with task support automatically calls for a human-driven vehicle to come to the location the vehicle is stopped to take the passenger to their destination. Without task support, Chris has to summon a new ride. When the vehicle has human support, a map with route information and the vehicle's location is displayed, an estimated arrival time is shown and updates are provided (after the failure occurs), and information about recovery actions being taken are reported. Additionally, in the "high" severity condition human support provides a warning message stating that the vehicle is unable to operate in severe weather, and (if not combined with task support) informs the passengers that they need to find another ride.

C. Dependent Variables

We measured 9 dependent variables using a series of 7 point Likert scale questions regarding how participants believed the character (Chris) felt about the robot following the second half of the story. Participants were asked how *satisfied, pleased*, and *disappointed* Chris was with the service. They were asked how *reliable, dependable, competent, responsible*, and *trustworthy* Chris believed the robot to be.

TABLE I

VACUUM SCENARIO QUESTIONS

How much more or less …
…satisfied is Chris with the robot's performance now compared to previous experiences?
…pleased is Chris with the robot's most recent results compared to previous experiences?
…does Chris trust the robot now compared to prior use?
…trust does Chris now have in the robot, compared to previous experiences?
…will Chris rely on the robot to clean the floors in the future?
…dependable does Chris believe the robot to be compared to before?
…competent does Chris believe the robot to be compared to before?
…certain is Chris that the robot will be able to clean the whole house in the future, given this latest experience?
…responsible does Chris believe the robot to be compared to before?

Possible Responses: *Much Less, Less, Somewhat Less, About the Same, Somewhat More, More,* and *Much More*

Finally, they were asked how *risky* it would be for Chris to use the robot in the future (see Table I). Anticipating that any kind of failure might overpower the effects of the other independent variables, participants were asked to compare Chris' latest experience relative to previous experience with the robot using the scale *Much Less, Less, Somewhat Less, About the Same, Somewhat More, More,* and *Much More* for each dependent variable. Each variable was measured twice using two differently worded questions. The wording of the questions was kept consistent between scenarios, with the exception of context relevant words.

Participants were also asked two questions related to how they personally felt about the robot. These included whether they would want to use the robot described in the story, and if they would recommend the robot in the story to a friend.

D. Manipulation and Attention Checks

Four "attention check" questions where included to check that participants were paying careful attention to the survey. After reading each of the two parts of the story, participants were asked a multiple choice question the answer to which would be obvious to anyone that had read the story - such as "What was the name of the character in the story?" In addition, two attention check questions were included in the bank of Likert questions to ensure people were carefully reading the questions. The answers to these questions were included in the question itself, such as "How much more or less does Chris take pictures? Please answer 'less' to this question." Failure to answer any of the attention check questions resulted in disqualification of the data for analysis.

Participants were also asked to answer six true or false style questions about things mentioned during the story, called manipulation check questions. The questions asked about details in the story related to the four independent variables. Participant's needed to answer all six of these questions correctly in order to demonstrate they had correctly perceived the various important aspects of the story, and have their data included for analysis.

III. RESULTS

Data was gathered using Amazon's Mechanical Turk, with each participant being paid $0.90 for their work. Participants consisted of self-selected MTurk workers who lived in the United States and had previously performed at least 1000 Human Intelligence Tasks (HITs) with at least a 95% approval

rating. We collected 30 participants worth of complete data for each condition in each scenario, totaling 600 participants for each type of robot and a combined total of 1200 participants. We were able to facilitate a between-subjects study due to MTurk workers being required to register their tax information with their account, MTurk providing unique workers for each HIT, and disallowing individual IP addresses from completing each scenario more then once. While 68 of the 1200 people involved (5.6%) participated in both the taxi and vacuum scenarios, each scenario was analyzed independently. See Table II for demographic details.

A. Measuring reaction to failure

Each of the dependent variables reflected different aspects of participants' overall perception of the character's (Chris') reaction to the robot's latest performance. We performed a exploratory factor analysis of the Likert scale questions for each scenario. A Scree test concluded that in both cases there was a single latent variable. The factor analysis accounted for 77% of the variance in the vacuum data and 66% of the variance in the taxi data. Variables in the taxi scenario had a Chronbach's $\alpha = 0.97$ and variables in the vacuum scenario had a Chronbach's $\alpha = 0.98$. All variables loaded the single factor, which we call REACTION, in both cases. Responses to questions with negative wordings were inverted prior to analysis; however, the negative attributes "disappointed" and "risky" were not inverted and subsequently received negative loadings. Thus, positive scores represent positive reactions to the robot's behavior while negative scores represent negative reactions. REACTION scores are shown in Figure 1.

A two-way ANOVA was performed to compare the influence of failure severity and context risk on REACTION. There was a significant main effect of failure severity on REACTION in both the taxi $[F(2) = 287.1284, p < 0.001, \eta_p^2 = 0.491]$ and vacuum $[F(2) = 410.4056, p < 0.001, \eta_p^2 = 0.58]$ surveys. A significant main effect of context risk on REACTION was found in the taxi survey $[F(1) = 13.6936, p < 0.001, \eta_p^2 = 0.039]$, but not in the vacuum survey $[F(1) = 0.9546, p = 0.33, \eta_p^2 = 0.0007]$. There was a significant interaction between context risk and failure severity in taxi survey$[F(2) = 7.6941, p < 0.001, \eta_p^2 = 0.025]$, but not in the vacuum survey $[F(2) = 0.8814, p = 0.415, \eta_p^2 = 0.0029]$.

Most participants who experienced the robot failing without any support had a negative REACTION (*taxi:* 92%, n=120; *vacuum:* 90%, n=120), while nearly everyone who experienced the robot without any failure (both with and without support) had a positive REACTION (*taxi:* 99%, n=120; *vacuum:* 96%, n=120).

B. Effect of support on reaction

A one-way ANOVA was performed to compare the influence of support type on people's reactions for each risk-failure combination. There was a significant effect of support type on REACTION in all conditions in both robot scenarios at a significance level of $\alpha = 0.05$. In the taxi scenario, a significant main effect was found in the low-risk, low-failure condition $[F(3, 116) = 12.61, p < 0.001, \eta^2 = 0.246]$, the low-risk, high-failure condition $[F(3, 116) = 15.27, p < 0.001, \eta^2 = 0.283]$, the high-risk, low-failure condition $[F(3, 116) = 3.805, p < 0.05, \eta^2 = 0.089]$, and the high-risk, high-failure condition $[F(3, 116) = 12.19, p < 0.001, \eta^2 = 0.239]$. In the vacuum scenario, a significant main effect was found in the low-risk, low-failure condition $[F(3, 116) = 13.16, p < 0.001, \eta^2 = 0.254]$, the low-risk, high-failure condition $[F(3, 116) = 17.27, p < 0.001, \eta^2 = 0.309]$, the high-risk, low-failure condition $[F(3, 116) = 38.06, p < 0.001, \eta^2 = 0.496]$, and the high-risk, high-failure condition $[F(3, 116) = 21.42, p < 0.001, \eta^2 = 0.356]$. For each of these conditions, a post-hoc Tukey's HSD test was used to determine significant differences between human support (HUMAN), task support (TASK), combined human and task support (COMBINED) and no support (NONE). The results of the post-hoc tests are shown in Figure 1.

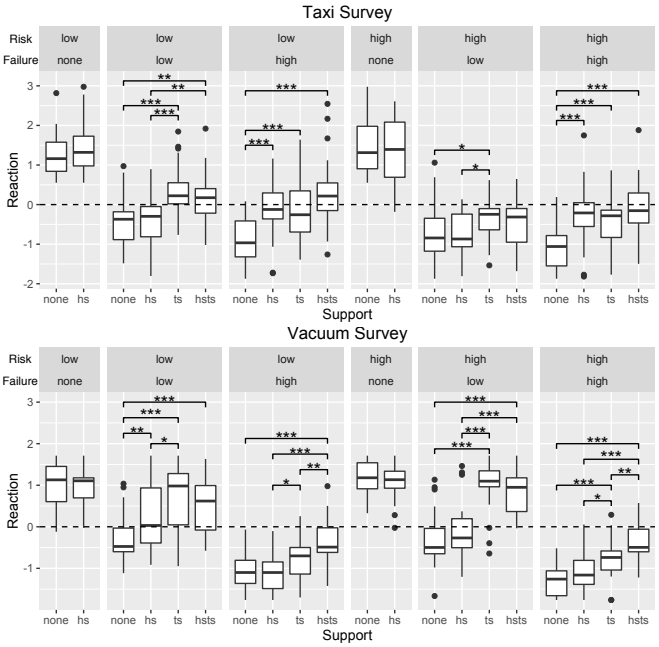

Fig. 1. Participants' REACTION scores grouped by risk, failure, and support type. *hs:* HUMAN, *ts:* TASK, *hsts:* COMBINED. $n = 30$ for each bar. Results from post-hoc Tukey's HSD tests: $\star p \leq 0.05$, $\star\star p < 0.01$, $\star\star\star p < 0.001$.

C. Effect of support on wanting to use the robot

A one-way ANOVA was performed to compare the influences of support on people's responses to wanting to use the robot described in the scenario they read. There was a significant main effect of support on wanting to use the robot in all conditions. In the vacuum scenario, there was a significant main effect $[F(3, 116) = 4.53, p < 0.01, \eta^2 = 0.105]$ in the low-risk, low-failure condition. A post-hoc test showed significant differences between TASK and NONE ($p < 0.05$), and COMBINED and NONE ($p < 0.01$). There was a significant main effect $[F(3, 116) = 4.25, p < 0.01, \eta^2 = 0.099]$ in the high-risk, low-failure condition. A post-hoc test showed significant differences between TASK and NONE ($p < 0.05$), and COMBINED and NONE ($p < 0.05$). There was a significant main effect $[F(3, 116) = 3.876, p = 0.01, \eta^2 = 0.091]$ in the low-risk, high-failure condition. A post-hoc test showed significant

TABLE II

DEMOGRAPHICS

Age

	Vacuum	Taxi
18-21	15	13
22-34	247	281
35-44	150	148
45-54	103	98
55-64	65	49
65-over	20	11

Gender

	Vacuum	Taxi
Male	294	289
Female	304	306
Other	2	5

Education

	Vacuum	Taxi
< HS	4	3
HS	70	49
Vocational	19	26
In College	148	130
2 Yr Deg	74	63
4 Yr Deg	211	238
Grad Deg	74	91

differences between TASK and HUMAN ($p = 0.05$), and COMBINED and HUMAN ($p = 0.01$). There was a significant main effect $[F(3, 116) = 4.905, p < 0.01, \eta^2 = 0.112]$ in the high-risk, high-failure condition. A post-hoc test showed significant differences between COMBINED and NONE ($p < 0.01$), and COMBINED and HUMAN ($p < 0.05$).

In the taxi scenario, there was a significant main effect $[F(3, 116) = 3.339, p = 0.02, \eta^2 = 0.079]$ in the low-risk, low-failure condition. A post-hoc test showed significant differences between TASK and NONE ($p = 0.02$). There was a significant main effect $[F(3, 116) = 4.695, p < 0.01, \eta^2 = 0.108]$ in the high-risk, low-failure condition. A post-hoc test showed significant differences between TASK and NONE ($p < 0.05$), COMBINED and NONE ($p < 0.05$), TASK and HUMAN ($p < 0.05$), and COMBINED and HUMAN ($p < 0.05$). There was a significant main effect $[F(3, 116) = 5.139, p < 0.01, \eta^2 = 0.117]$ in the low-risk, high-failure condition. A post-hoc test showed significant differences between HUMAN and NONE ($p < 0.01$), and COMBINED and NONE ($p = 0.01$). There was a significant main effect $[F(3, 116) = 7.016, p < 0.001, \eta^2 = 0.153]$ in the high-risk, high-failure condition. A post-hoc test showed significant differences between HUMAN and NONE ($p < 0.01$), COMBINED and NONE ($p < 0.001$), and COMBINED and TASK ($p < 0.05$).

The REACTION score of each participant was compared to their response for "I would want to use this robot/vehicle." 95% (251/264) of participants in the vacuum survey and 77% (200/258) of participants in the taxi survey who had a positive REACTION score responded with some level of agreement. Of participants who had a negative REACTION, only 58% (194/336) of participants in the vacuum survey and 42% (144/342) of participants in the taxi survey responded with some level of agreement.

IV. ANALYSIS OF RESULTS

Participants' REACTION was significantly influenced by failure severity in both the taxi and vacuum surveys, and by context risk in the taxi survey. The REACTION scale correctly divided people who experienced the robot operating successfully from people who experienced the robot failing (with no support) by whether or not their score was positive or negative with 94% accuracy ($n = 480$). The magnitude of people's REACTION was significantly influenced by the severity level of the failure in both surveys.

The REACTION scale also highlights the variability by which recovery strategies can alter a person's response to a failure, ranging from having no measurable effect to being indistinguishable from not having failed. Further, it indicates that the effectiveness of recovery strategies (which in general improved people's REACTION to failure) seems to be influenced by the task, context risk, and severity or type of failure.

Hypothesis 1: *Providing human support will help mitigate the negative effects caused by failure.*

Human support significantly ($p < 0.01$) improved people's REACTION in several scenarios. The amount it influenced people's REACTION varied by the task, severity of failure, and context risk. However, the significance of human support seems to be better correlated to whether the information conveyed could be used by the person to effect the outcome of the situation. In the high severity condition of the taxi scenario, the car informed the passenger they needed to call for another ride which significantly improved people's REACTION. When the taxi missed a turn in the low severity condition, the support information allowed the user to predict but not effect the outcome, and had almost no effect. There were no conditions in the vacuum scenario in which the human support was used to alter the outcome of the situation. However, it still significantly improved people's reactions in the low-risk, low-failure scenario. This could be interpreted as Chris knowing that he would need to clean the floor himself when he got home - something he would have the chance to do in the low risk scenario but not the high risk scenario.

Hypothesis 2: *Providing task support will help mitigate the negative effects caused by failure.*

Using task support significantly improved people's REACTION ($p < 0.05$) in all but one scenario (vacuum, low-risk, high-failure severity, $p = 0.06$). One particularly interesting data point is the extremely positive REACTION to task support in the high-risk, low-failure condition of the vacuum scenario. The robot's behavior in this case was to return to its charger before the battery ran out, and resume cleaning where it left off when it had recharged - thus eventually completing the task. While the completed task certainly contributed to the high REACTION, the response to the same behavior in the corresponding low-risk condition had a much higher variance. One possible explanation is that the difference in variance may be the result of people being less certain about the significance of the failure in the low-risk condition compared to the high-risk condition. However, this would suggest there should also be higher variances in the low-risk, high-failure condition, which was not the case. Another possible explanation is that the difference in risk changed the way people imagined Chris perceiving the way the task was completed. In the low-risk condition Chris was portrayed as experimenting with the robot's capabilities which may have prompted a more critical view of the results, while in the high-risk condition Chris was hoping for a particular result despite the lack of a precedent, making the robot's performance a pleasant surprise.

Hypothesis 3: *A combination of both human and task support will help mitigate the negative effects caused by failure.*

Combined support significantly ($p < 0.001$) improved people's REACTION in all but one scenario (taxi, high-risk, low-failure severity, $p = 0.16$). In one case (vacuum, high-risk, high-failure), combined support was significantly

better ($p < 0.01$) than using task support, which was itself significantly better ($p < 0.001$) than no support. However, a non-significant trend can be seen in which combined support performed better than task support in high severity situations, but worse in low severity situations. One possible explanation for this is that in certain situations the additional information is regarded as too verbose or possibly annoying, while in others the information is welcomed. Unfortunately, this logic would be better supported if the trend corresponded with differences in context risk rather than the failure severity.

Hypothesis 4: *Recovery strategies which reduce the negative effects of a failure will also increase the likelihood of users wanting to use the system again.*

The percentage of people who wanted to use the robot was much higher among people who had a positive REACTION score than among those who had negative scores. Both human and task support effected how much people wanted to use the robot, although neither had a ubiquitous effect.

V. Limitations

Survey experiments have some inherent flaws [18]. The self-selection of participants may have introduced non-response error into the data, and the extensive use of rating scales in our survey may have caused some people to mark multiple questions with the same answer (non-differentiation). Responses could be biased for various reasons such as acquiescence response bias (tendency to agree regardless of the question) or question wording incidentally cueing a particular response. Our survey was only available to people residing in the US, and may not reflect the way people in other parts of the world would behave in similar situations. Furthermore, prior work has shown the MTurk population does not perfectly match the US population (it was also not extremely different) [19]. Thus, the experiment could benefit from being repeated with other populations.

Finally, the third person perspective of both the story and questions was chosen over a first person perspective to allow participants to distance themselves from the situation, reducing the effects of subconsciously biased responses such as from people trying to portray themselves in a particular manner [20]. However, reading about a hypothetical situation someone else is experiencing is not the same as experiencing the same situation for one's self in real life. Thus, a laboratory experiment in which participants experience failures in-person is needed to verify these results.

VI. Conclusion and Future Work

The REACTION scale captures the main characteristics of failure by autonomous robots, while also highlighting the nuanced complexity of the situation. We have demonstrated its use by comparing successful and failed operation of robots with various recovery strategies. In doing so we found evidence that while human support and task support can both be used to mitigate failures, the type/severity of failure and context risk influence their effectiveness.

In this study, we only compared results of the REACTION scale within individual robots due to the use of separate exploratory factor analysis for each study. Similarities observed between factor loadings of the two analysis suggest we should be able to refine the REACTION scale into a generic question bank that will be task and platform independent, potentially allowing it to be used as a comparison between robots. Other future work includes determining if additional variables need to be controlled and adapting the scale for use by interaction roles [21] other than the operator.

References

[1] B. Lussier, R. Chatila, F. Ingrand, M.-O. Killijian, and D. Powell, "On fault tolerance and robustness in autonomous systems," in *3rd IARP-IEEE/RAS-EURON Joint Workshop on Technical Challenges for Dependable Robots in Human Environments*, Sept 2004, pp. 351–358.

[2] D. W. Payton, D. Keirsey, D. M. Kimble, K. Jimmy, and J. K. Rosenblatt, "Do whatever works: A robust approach to fault-tolerant autonomous control," *Applied Intelligence*, vol. 2, no. 3, May 1992.

[3] E. Cha, A. Dragan, and S. Srinivasa, "Perceived robot capability," in *24th IEEE International Symposium on Robot and Human Interactive Communication*, Kobe, Japan, Aug 2015.

[4] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proc. of the 8th ACM/IEEE Intl. Conf. on Human-Robot Interaction*, 2013.

[5] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *Proc. of the 5th ACM/IEEE Intl. Conf. on Human-Robot Interaction*, 2010.

[6] S. Rosenthal, M. Veloso, and A. K. Dey, "Is someone in this office available to help me?" *Journal of Intelligent & Robotic Systems*, vol. 66, no. 1, pp. 205–221, 2012.

[7] R. A. Knepper, S. Tellex, A. Li, N. Roy, and D. Rus, "Recovering from failure by asking for help," *Autonomous Robots*, vol. 39, no. 3, pp. 347–362, 2015.

[8] T. Kim and P. Hinds, "Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction," in *Proc. of the 15th IEEE Intl. Symposium on Robot and Human Interactive Communication*, 2006.

[9] V. Groom, J. Chen, T. Johnson, F. A. Kara, and C. Nass, "Critic, compatriot, or chump?: Responses to robot blame attribution," in *Proc. of the 5th ACM/IEEE Intl. Conf. on Human-robot Interaction*, 2010.

[10] P. Kaniarasu and A. Steinfeld, "Effects of blame on trust in human robot interaction," in *The 23rd IEEE Intl. Symposium on Robot and Human Interactive Communication*, Aug 2014.

[11] J. Carlson and R. R. Murphy, "How ugvs physically fail in the field," *IEEE Transactions on Robotics*, vol. 21, no. 3, 2005.

[12] G. Steinbauer, "A survey about faults of robots used in robocup," in *RoboCup 2012: Robot Soccer World Cup XVI*. Springer, 2013.

[13] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, no. 1, pp. 32–64, 1995.

[14] D. R. Olsen, Jr. and S. B. Wood, "Fan-out: measuring human control of multiple robots," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2004.

[15] D. B. Kaber and M. R. Endsley, "Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety," *Process Safety Progress*, vol. 16, no. 3, pp. 126–131, 1997.

[16] V. S. Folkes, "Consumer reactions to product failure: An attributional approach," *Journal of Consumer Research*, vol. 10, no. 4, 1984.

[17] T. W. Andreassen, "Antecedents to satisfaction with service recovery," *European Journal of Marketing*, vol. 34, no. 1/2, pp. 156–175, 2000.

[18] P. S. Visser, J. A. Krosnick, and P. J. Lavrakas, "Survey research," in *Handbook of research methods in social and personality psychology*, H. T. Reis and C. M. Judd, Eds. New York, NY, US: Cambridge University Press, 2000, vol. xii, ch. 9, pp. 223–252.

[19] A. J. Berinsky, G. A. Huber, and G. S. Lenz, "Evaluating online labor markets for experimental research: Amazon. com's mechanical turk," *Political Analysis*, vol. 20, no. 3, pp. 351–368, 2012.

[20] R. E. Nisbett, C. Caputo, P. Legant, and J. Marecek, "Behavior as seen by the actor and as seen by the observer." *Journal of Personality and Social Psychology*, vol. 27, no. 2, p. 154, 1973.

[21] H. A. Yanco and J. L. Drury, "Classifying human-robot interaction: an updated taxonomy." in *Proc. of the IEEE Conf. on Systems, Man and Cybernetics*, Oct 2004.