# Effects of Explanation Strategies to Resolve Failures in Human-Robot Collaboration

Parag Khanna<sup>\*1</sup>, Elmira Yadollahi<sup>\*1</sup>, Mårten Björkman<sup>1</sup>, Iolanda Leite<sup>1</sup>, and Christian Smith<sup>1</sup>

Abstract-Despite significant improvements in robot capabilities, they are likely to fail in human-robot collaborative tasks due to high unpredictability in human environments and varying human expectations. In this work, we explore the role of explanation of failures by a robot in a humanrobot collaborative task. We present a user study incorporating common failures in collaborative tasks with human assistance to resolve the failure. In the study, a robot and a human work together to fill a shelf with objects. Upon encountering a failure, the robot explains the failure and the resolution to overcome the failure, either through handovers or humans completing the task. The study is conducted using different levels of robotic explanation based on the failure action, failure cause, and action history, and different strategies in providing the explanation over the course of repeated interaction. Our results show that the success in resolving the failures is not only a function of the level of explanation but also the type of failures. Furthermore, while novice users rate the robot higher overall in terms of their satisfaction with the explanation, their satisfaction is not only a function of the robot's explanation level at a certain round but also the prior information they received from the robot.

#### I. INTRODUCTION

Robots and artificial agents' capabilities rapidly grow as they are deployed in real-world environments like factories, hospitals, and schools. Nevertheless, failures inevitably occur during task execution and collaboration [1], and with the increasing use of robots in in-the-wild environments, where robots are more prone to collaborate with novice and nonexpert users, the study of failures, and mitigating their impact becomes imminent. While in many failure scenarios, robots can recover by themselves, there are cases where human assistance is required to resolve failures for task continuity [2]. For a non-expert user, understanding *why* a robot failure has occurred and if and *how* they could contribute to the recovery is essential for smooth human-robot collaboration.

The emergence of studies on robot failures attests to the evolution of research from exploring people's perception and resolution of failures to the robot's role in identifying and mitigating them [3]. While the topic has expanded to include the use of holistic approaches such as explanation, apology, denial, and promise, that identify, resolve, and mitigate failures for untrained users [1], [4], [5], few have studied the effect of these approaches in repeated interactions to the best of our knowledge [6]. Providing an explanation is a practical approach to mitigating failures in collaborative scenarios, particularly when failures require human intervention or



Fig. 1. The pick and place task: Human places the objects on the table, then the robot's goal is to place them on the shelf while providing an explanation in case a failure occurs. The zoomed views show (top right) two levels of the shelf and (bottom right) the markers for object placement on the table.

assistance. Advances in the field of Explainable AI (XAI) [7], and its extension to goal-driven explanations [8] for robots and agents contribute to research on explanation generation for failures. Currently, the literature on the topic of resolving failures via explanations focuses on determining *what* type of information should be presented [4] and *how* the explanations should be automatically generated [9], [10]. In our research, we address the missing link between explanation generation and participant satisfaction in repeated interactions with recurring failures, for example, do we need to be consistent with the explanations as the failures reoccur, or should we provide more details early on and reduce as the interaction continues?

As a result, we developed a study to understand how different strategies of providing explanations in repeated interaction influence non-expert users' performance and satisfaction. We developed a collaborative pick-and-place task, where the robot and human had to place objects from four baskets on a shelf. We counted each basket as one round of interaction and aimed to have four rounds of interaction. We designed two types of strategies for providing the explanations: 1) maintaining the details of the explanation during the rounds, i.e. fixed strategy and 2) reducing the details of the explanation, i.e. *decaying strategy*. To develop the strategies, we first defined the explanation levels inspired by the previous work by [4] and labeled them as low, mid, and high. Subsequently, we conducted a between-subject user study where participants experienced either of these strategies in four rounds of interaction. We aimed at evaluating how participants' performance in resolving failures and satisfaction with the explanations were impacted by the explanation levels and strategies which lead us to the following research questions:

<sup>&</sup>lt;sup>1</sup> Division of Robotics, Perception and Learning (RPL), EECS, KTH Royal Institute of Technology, Sweden {paragk, elmiry, celle, iolanda, ccs}@kth.se

<sup>\*</sup> Authors contributed equally

- *RQ1:* How does explanation level impact participants' performance in the task and satisfaction with the explanation?
- *RQ2*: Which explanation strategy (fixed vs. decaying) leads to better performance and satisfaction in participants?

# II. RELATED WORK

Several studies in the fields of human-robot interaction (HRI) and collaboration (HRC) have addressed the importance of understanding the effect of failures on trust [11] and perception [12], and mitigating its impact through failure recovery [13], explainability [4], and promise [5]. In [11], a user study was designed to investigate the effects of a collaborative robot's failure on human trust and the impact of justification strategies. Altogether, the results indicated that a faulty robot is regarded as far less trustworthy. It is also shown that the impact of failures was reduced with justifications when the consequence of failure was less significant. With the change of trend in using holistic approaches to identify and resolve failures, one of the approaches used more commonly in recent years is generating and providing explanations, studied on both the computational front, e.g., XAI [4], [14], and the social front, e.g., behavioral [15]. Several of the research in explainability has been inspired by the sociocognitive definitions of explanations in various fields and their social implications [15]. A recent review by Wallkotter et al. [16] identified three research directions on the topic that contribute to understanding the explainability mechanisms and how they can be integrated into the interaction context with occasional overlaps with the field of XAI. On the topic of studying explainability in robot failures, a study by Das et. al. investigated the types of explanation that helped non-experts to identify robot failures and assist the recovery by extending the XAIP algorithms via introducing failure explanation [4]. The goal was to produce explanations for unexpected failures in a pick-and-place task for a robot in a household environment. Failure and solution identification has been observed to be most effective when explanations include the context of the failure action and the history of previous actions. Another study in [14] used machine-learning models to predict robot grasp failures and study the tradeoff between accuracy using black-box models and interpretability using explainable models. They showed an explanation of predicted faults could contribute to the efficiency of designing the robot and avoiding future failures. Diel et al. proposed a causal-based method to develop explanations for robot failures in collaborative scenarios [10]. Their approach incorporated learning from a causal

Bayesian network that enabled the robot to generate the explanation by contrasting a failure state against the closest successful state and by using a breadth-first search. Beyond the studies focusing on generating explanation, the effects of different types and amounts of explanation by an XAI system on human understanding of the system were discussed in [17], where an increase in the information contained in the explanation resulted in the users' better understanding and prediction of the system behavior, as well as increased user performance. However, this came at a cost of increased time and attention needed by users to comprehend the explanation.

#### III. DESIGN

#### A. Collaborative Task Design

We designed a pick-and-place task where a Baxter robot and a human had the goal of collaboratively picking objects from a basket and placing them on the shelf (Fig. 1). We created four baskets, (numbered 1 to 4), each including a combination of four household items presented in Fig 2. This resulted in a total of 16 objects that needed to be placed on the shelf during the whole duration of the experiment. In our design, each round of the experiment started by picking the items from one basket, putting them on the table, and placing them on the shelf, when the task was successfully executed. The placement of an object was deemed unsuccessful if it was not placed on the shelf.

We marked each object in the basket with an A, B, C, or D tag on one face and a fiducial tag [18] on the other to let the robot detect the object. At the start of each round, the human collaborator placed all objects from the basket in corresponding positions as they are marked (see Fig. 1). For handling each object, the robot executed the following steps: *detect* the object, *pick* it up, *carry* it, and finally *place* it on the shelf. A possible failure could happen at each step during collaboration with the robot. As result, we defined the following failures and possible resolutions that could help complete the task despite a failure. In the next section, the explanations generated based on these failures and resolutions are provided.

- 1) **Detect Failure**  $(f_0)$ : Robot failed to detect the object on the table, e.g. not being able to scan the tag. **Resolution Action**  $(r_0)$  Human moves or rotate the object to ensure the tag is visible to the robot.
- 2) **Pick Failure**  $(f_1)$ : Robot failed to pick up an object, e.g. not fitting in the gripper, based on its placement or size. **Resolution Action**  $(r_1)$ : Human picks up and hands over the object to the robot.



(a) Sponge 10g (left), Toy 25g (right)



(b) Pen box 500g (left),<br/>Pen box-heavy 750g (right)(c) Heavy box 1 725g (left),<br/>Heavy box 2 1000g (right)





(d) Cloth box 710g, Flat box 130g

(e) Random Box 2 580g (left) Random Box 1-3 35g (right)

Fig. 2. Objects as they were required to be placed in front of the robot



Fig. 3. Description of human-robot collaborative task with the robot and human action spaces. Arrows in green represent transitions due to action success. Arrow in red represents transitions due to action failure.

- 3) **Carry Failure**  $(f_2)$ : Robot failed to carry an object, e.g. weight beyond the limit robot can handle. **Resolution Action**  $(r_2)$ : Robot hands over the object to the human and the human places it on the shelf.
- 4) **Place Failure**  $(f_3)$ : Robot failed to place an object, e.g. the desired destination is beyond the robot's reach. **Resolution Action**  $(r_3)$ : Robot hands over the object to the human, and they place it at the desired location.

Fig. 3 shows the workflow for placement of an object with possible failures denoted in red, which was accompanied by an explanation from the robot and resolved from the human side. In the task design, we included 7 objects for which the robot successfully executes all steps and 9 objects involving some robotic failures, spread out across the four rounds as shown in Table II. We are not intentionally incorporating the detection failures ( $f_0$ ), but as they might occur due to the way the object is placed on the table, we provide the appropriate resolution. For any other unintended failure ( $f_4$ ),

ROUND-WISE DESCRIPTION									
Round	Object	Object	Robotic Action-Success			Resolution			
No.	Туре		Pick	Carry	Place				
	А	Sponge	~	~	$\checkmark$	None			
1	В	Cloth-Bag	Х	Х	-	$r_1 \& r_2$			
1	С	Random-Box1	$\checkmark$	$\checkmark$	$\checkmark$	None			
	D	Pen-Box	$\checkmark$	$\checkmark$	Х	$r_3$			
	А	Heavy-Box1	~	Х	-	$r_2$			
	В	Random-Box2	$\checkmark$	$\checkmark$	$\checkmark$	None			
2	С	Flat-Box	Х	$\checkmark$	$\checkmark$	$r_1$			
	D	Toy	$\checkmark$	$\checkmark$	$\checkmark$	None			
	А	Pen-Box	~	~	Х	$r_3$			
2	В	Pen-Box-Heavy	$\checkmark$	Х	-	$r_2$			
3	С	Toy	$\checkmark$	$\checkmark$	$\checkmark$	None			
	D	Random-Box3	$\checkmark$	$\checkmark$	$\checkmark$	None			
	А	Flat-Box	Х	$\checkmark$	$\checkmark$	$r_1$			
4	В	Sponge	$\checkmark$	$\checkmark$	$\checkmark$	None			
4	С	Pen-Box	$\checkmark$	$\checkmark$	Х	$r_3$			
	D	Heavy-Box2	Х	Х	-	$r_1 \& r_2$			

TABLE II

the resolution  $(r_4)$  in the form of asking the human to place the object on the shelf was also integrated.

## B. Explainability Mechanisms

We considered three verbal explanation levels: low, medium, and high. Additionally, we included a nonverbal baseline to complement the explanations as a result of initial pilot studies where we noticed users needed some baseline behaviors to understand the failures, particularly when given low-level explanations. As a result, we designed the following explanation levels inspired by [4] and Table I presents each explanation for each failure type.

- Low Level: Based on *action-based* explanation in [19]. After the failure, the robot states the failure action and its resolution.
- Medium Level: Based on *context-based* explanation in [19]. Post failure, the robot states the failed action and the cause of failure, followed by a resolution statement.

Level	$f_0$	$f_1$	$f_2$	$f_3$	f4 Shakes head		
Zero (Non- verbal)	Shakes its head to show unable to find the object	Tries to get the object in its gripper, shakes its head at failure, and moves the arm to the handover position	Moves the arm down with the object in the gripper, conveying it fails to carry it, shakes head and moves to the handover position	Stops arm near the lower level of the shelf, shakes head and move the arm to the handover position			
Resolution	Nothing	Nothing	Nothing	Nothing	Nothing		
Low	"I can't detect the object"	"I can't pick up the object"	"I can't carry the object"	"I can't place the object"	"I failed to handle the ob- ject"		
Resolution	"Move it" "Hand it to me"		"Carry it for me"	"Place it for me"	"Place it for me"		
Medium Resolution	"I can't detect the object because the tag is not visi- ble to me" "Can you move it within my field of view?"	"I can't pick up the object because it doesn't fit in my gripper" "Can you hand it over to me?"	"I can't carry the object because it is too heavy for my arm" "Can you carry the object for me?"	"I can't place the object because the destination is out of my arm's reach" "Can you place the object for me?"	"I failed to handle the object because an unex- pected failure happened" "Can you place the object for me?"		
High	"I scanned all the objects and can't detect the object I am looking for, because probably the tag is not vis- ible to me"	"I can detect the object, but I can't pick it up because it doesn't fit in my gripper"	"I can pick up the object, but I can't carry it because it is too heavy for my arm"	"I can carry the object, but I can't place it because the destination is out of my arm's reach"	"I can detect the object but I failed to handle it be- cause an unexpected fail- ure happened"		
Resolution	"Can you move it within my field of view to make sure I see the tag?"	"Can you hand it over to me by placing it in my gripper?"	"Can you carry the object for me and place it on the shelf?"	"Can you place the object on the shelf location that is out of my reach?"	"Can you finish placing the object for me?"		

TABLE I FAILURE-WISE EXPLANATION LEVELS

• High Level: Based on *context-based* + *history-based* explanation in [19]. After failure, the robot states the previous successfully completed action, the current failure action, and its cause. The resolution statement also includes the resolution action.

Informed by our pilot studies, we included a nonverbal baseline to help with identifying the failure in lower explanation levels.

• Zero (Nonverbal): This only includes the robot head shaking at each failure with more specific robotic actions based on the failure type.

## C. Interaction Details

The Baxter robot was programmed in ROS and only used its left arm. More detail on the technical developments and the interaction is available in [20] and the accompanying video with this work. Each round started with the robot receiving verbal confirmation that all objects are placed on the table, where the robot proceeded to pick up the objects by following the action sequence depicted in Fig. 3. Once a failure occurred, the robot exhibited non-verbal actions described in Table I. followed by an explanation based on the current strategy and waiting for the human to resolve the failure before moving to the next step. If the failure was not resolved in a predefined amount of time, the robot repeated itself up to five times spaced with three-second intervals. The system is autonomous, but the experimenter (unbeknown to the participant) made the decision to move to the next step when they failed to complete the task after five repetitions (something that might happen in low explanation cases). To avoid handover failures, the human-to-robot handover was completed after the robot received a verbal confirmation to close its gripper after the human handed over the object, and the robot-to-human handover used sufficient pull-force, in line with a prior study [21].

#### IV. METHODOLOGY

## A. Experiment Design

We investigated two explanation strategies (fixed and decaying) using the three levels of explanations (low, mid, and high). For the fixed explanation strategy, we tested each explanation level using the three conditions: C1, C2, and C3 presented in Table III. For the decaying explanation strategy, we focused on the rate of decay. Given four rounds of interactions, we defined two types of decay: *slow* (D1) and *rapid* (D2). Slow decay was implemented by reducing the level of explanation once per round, which resulted in the following combination: high, medium, low, and none. In Rapid decay, the explanation was reduced from high to low and keep it in a low level as presented in Table III.

#### B. Hypotheses

Prior research on the topic of XAI and explainability in robotics has shown mixed results in how humans perceive explanations. In the study by Das et. al. [4], explanations that encompassed context and history of past successful interactions were able to improve failure identification and

TABLE III EXPERIMENTAL CONDITIONS

ID	Details	Round 1	Round 2	Round 3	Round 4					
C1	Fixed-Low	Low Mid	Low Mid	Low Mid	Low					
C2 C3	Fixed-High	High	High	High	High					
D1	Decay-Slow	High	Mid	Low	None					
D2	Decay-Rapid	High	Low	Low	Low					

failure. Their context-based including history corresponds to our high-level explanation. On the other hand, with regard to our first research question, we have the following hypotheses:

- *H1a:* participants show better performance e.g. shorter task resolution time and successfully resolving the failure in the high explanation level compared to low and mid-levels.
- *H1b:* participants are more satisfied when given more detailed explanations compared to lower or intermediate explanations.

Given our second research question, we hypothesize:

- H2a: In final rounds, participants' performance and satisfaction in decaying conditions (with low explanations) is better than the fixed-low explanation condition.
- H2b: In final rounds, participants have comparable performance and satisfaction in decaying conditions (with low explanations) compared to fixed-high explanation conditions (with high explanations).

For H2a, we specifically focus on Low-level explanations in round 3 and expect participants to perform better in the decaying conditions (D1, D2) compared to the fixed low explanation condition C1 as they were given higher explanations in the previous rounds. We also compare the performance in round 3 of C3 with (D1, D2) for which we expect participants to have similar perceptions and performances compared to C3, despite the low level of explanation, as they have already been exposed to higher explanations in earlier rounds (H2b). The level of explanation is a betweensubject variable for fixed strategy conditions (C1, C2, and C3). The level of explanation also varies within the decaying strategy conditions (D1 and D2) as it changes both between conditions and within the decaying conditions. The dependent variables are participants' performance in the task and their explanation satisfaction rating.

#### C. Measures

In line with our hypotheses, the measures for this experiment were *participants' performance* and *participants' satisfaction*, collected through multiple variables and after the completion of each round of interaction.

1) Participants' performance: We measure the performance over two dimensions corresponding to the instances where failure occurs, 1) the time they take to intervene and resolve a failure, 2) their success rate in resolving the failure, e.g. placing the object on the shelf.

Failure resolution time:  $T_{res}$  is calculated from when the robot completes the explanation statement to when the participant completes the resolution, e.g. placing the object on the shelf. **Success rate of failure resolution:** This is a measure of the successful resolution of each failure and it is measured differently depending on the type of failure as presented in the Collaborative Task Design section.

2) Participants perception: The participant's perception was measured using an explanation satisfaction survey and some task-related questions. The task-related questions included more open-ended questions, designed to understand participants' approaches to resolving the failure beyond the robot's explanations. We are not reporting the qualitative analyses of the responses in this paper.

**Explanation satisfaction scale:** To measure how participants were satisfied with the explanations at each round, we asked them to respond to 8 questions after completing each round. The questions were originally introduced and evaluated in [22]. They define explanation satisfaction as "the degree to which users feel that they understand the AI system or process being explained to them". The questions were derived from the psychological literature on explanation and include several key attributes of explanations: *understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, trustworthiness.* 

## D. Participants and Procedure

We recruited sixty-nine participants via advertisement on campus. Our main criterion was that the participants had no prior experience in physical collaboration with a robot. Twelve participants had to be excluded from the analysis due to unaccounted robot failures beyond the failures designed for the experiment. The final sample size was N = 55 (M =26.63, SD = 7.42) (21 Female, 33 Male, 1 Other) resulting in 11 participants per condition. At the start, the participants filled out the consent form for data and video collection and reading procedural instructions. They were briefed about their role to place objects on the table and the robot's role to pick them up and place them on the shelf; however, no mention of the possible failures and related resolutions was presented. After the completion of the experiment, they were given a debriefing sheet describing the aim of the study.

#### V. RESULTS

To prepare the data, first, we evaluated the internal consistency of the questionnaires using Cronbach's alpha. The *explanation satisfaction* questionnaire presented high internal consistency with Cronbach's  $\alpha = 0.79$ ,  $\alpha = 0.91$ ,  $\alpha = 0.92$ , and  $\alpha = 0.92$  for each round, respectively.

# A. Impact of Explanation Level

To investigate *H1a* and *H1b*, we only looked at the first round of interaction and grouped participants into groups of low, mid, and high explanation levels. This implied grouping the participants in conditions C3, D1, and D2 into *High-level*, C2 into *Mid-level*, and C1 into *Low-level*. This decision was made to get a baseline for the explanation levels, additionally, to analyze strategies we need multiple rounds of interaction which we address in the next section.

Given that each failure type required a different resolution and intervention to resolve that failure successfully, we analyzed the performances separately for each failure type. Table IV(a) shows the success rate in resolving the failures for each failure type in all three levels. For carry failures, Fisher's exact test p = 0.0023 shows a significant difference between the low, mid, and high explanation levels in successfully resolving the failure (Fig. 4a). According to post hoc tests p = 0.0022 this difference is significant between *Highlevel* compared to the *Mid-level*. For place failure (Fig 4c), according to Fisher's exact test p = 0.0339, participants that received the *High-level* explanation were significantly more successful than the ones receiving the *Low-level* explanation.

For our second measure of analyzing performance, we looked at the time participants took to resolve failure cases. For pick and place failures, we observed no significant difference in the resolution times based on the explanation levels. For carry failures, Kruskal-Wallis chi-squared test H(2) = x, p = 0.0075 indicated that the resolution time significantly differed based on the explanation level. Post hoc tests and Figure 4b show the difference is significant between Low-level and Mid-level p = 0.0061, and Mid-level and High-level p = 0.045.

**Results for** *H1a*: Overall, the results partially support *H1a*, where we expected participants to perform better in High-level explanations compared to Mid and Low-level. However, the analyses show that failure type and how much the immediate resolution could be inferred from the environment, irrespective of the explanation, are important factors in the participants' performance.

**Results for** *H1b***:** Regarding *H1b*, we analyzed participants' responses to the explanation satisfaction questionnaire



Fig. 4. Performance in terms of success rate and resolution time for round 1

TABLE IV									
SUCCESS	RATE	IN	FAILURE	RESOLUTIO	N				

				Employedian	Employed		E-11		Englandian	Employed		E-il.	
Evalenction	Esilenee		Explanation Explanation		Failures		Explanation Explanation		Failures				
Explanation	D' 1	Fanures		Strategy	Level	Pick	Carry	Place	Strategy	Level	Pick	Carry	Place
Level	evel Pick Carry	Carry	Place	C1	Low	100%	100%	18.18%	C1	Low	100%	90.91%	18.18%
Low	100%	81.82%	18.18%	C2	Medium	100%	100%	45.45%	C2	Medium	100%	100%	36.36%
Medium	100%	54.55%	27.27%	C3	High	100%	100%	72.73%	C3	High	100%	100%	72.73%
High	100%	96.97%	60.61%	D1	Low	100%	72.72%	18.18%	D1	Low	100%	100%	45.45%
0	(a) <b>D</b> au	(-) <b>D</b> d 1		D2	Low	100%	81.82%	36.36%	D2	Low	100%	90.91%	63.63%
(a) Round 1													

(b) Round 3

(c) Round 4

after the first round. Kruskal-Wallis chi-squared test indicated no significant difference in the explanation satisfaction between the explanation levels H(2) = 2.47, p = 0.2903, rejecting our hypothesis. The distribution of the satisfaction rating in round 1 is presented in Fig. 7a.

## B. Impact of Explanation Strategy

To analyze the impact of the explanation strategy, we looked at participants' performance and satisfaction ratings in rounds 3 and 4 for conditions C1, C2, C3, D1, and D2. In H2a, we are comparing the final round performances in the *decaying* conditions, i.e. D1 and D2, versus the *fixed* condition C1. In round three of these conditions, participants are receiving Low-level explanations with different a priori. In round four, participants are receiving Low-level explanations in conditions C1 and D2, and Zero-level explanations in D1.

The percentages for the success rates in rounds 3 and round 4 are presented in Table IV-(b),(c). For pick and carry failures, participants showed success rates above 80% in all conditions. For place failures, while we observed better

performances in D1 and D2 conditions compared to C1 as shown in Fig. 5c, the difference was not significantly different. Regarding the failure resolution times in round 3, no significant difference was observed for pick and carry failures between C1, D1, and D2 conditions. However, for place failures, Kruskal-Wallis chi-squared test indicated that there was an overall difference in the resolution times between the three conditions H(2) = 2.47, p = 0.2903. The pairwise comparison confirmed that this difference was significant between C1 and D2 conditions H(2) = 2.47, p = 0.0386.

Furthermore, we explored the data in round 4, where the explanation level for condition D1 was reduced to baseline or none. As shown in Fig. 6, the performances for the D1 condition have decreased for all failure types, with a significant difference for place failure cases. By only looking at the performances for place failures in condition D1, we observe that after three rounds of interaction, participants were still not ready to resolve the failures without any explanation. Furthermore, the explanation satisfaction ratings for rounds 3 and 4 are presented in Fig. 7b, and 7c, and show





no significant difference between the discussed conditions.

**Results for** H2a and H2b: Overall, based on the performance and satisfaction results in the last rounds, we reject H2a. However, we can accept H2b, proving that participants in the last rounds of decaying explanation conditions; i.e. D1 and D2, showed comparable performances to the fixed-high explanation, i.e. C3.

## VI. DISCUSSION

#### A. Impact of Explanation Level

We observed that there is a significant effect of explanation level on the participants' performance. Participants showed an overall higher success rate in resolving failures when given context-based high-level explanations with the history of past successful actions which was also evident from the shorter time in resolving the failures and completing the task. This is aligned with the results from [4], where participants watched videos of the failures and respective explanations and their performance was evaluated based on success in identifying the cause of the failure and its resolution. Nevertheless, we noticed that the results are not generalizable over different types of failures. For example, participants have shown above 80% success rates in resolving pick failures irrespective of the given level of explanation. One reason lies in the nature of the failure to pick an object, which regardless of its cause, e.g. size, shape, or slippery edges, can be easily detected by a collaborator.

On the other hand, the performances in resolving carry and place failures exhibited some significant differences based on explanation level. We identify that in carry failure cases, the cause was not explicit, i.e. object weight was beyond the robot arm's limit. However, in contrast to our expectations, participants in Mid-level conditions, had the worst performances compared to Low and High-levels (Fig. 4a, 4b). This finding contributes to the argument that giving additional information without pointing to a cause or resolution can hinder human performance which is also aligned with Thagard's theory of explanatory coherence [23], where people prefer simpler explanations with fewer causes and more general explanations. In place failure cases, we observed significant performance improvement with the increase of explanation (Fig. 4c, 4d). Several factors could contribute to this, including the harder detection of the resolution without receiving the appropriate explanation. It is plausible that

participants understood the robot's failure to place the object on the shelf, however, they missed the exact reason, i.e., the inaccessibility of the lower shelf, and managed to place the object on the upper shelf, which was not the goal.

Overall our findings guide us to further investigate factors such as *failure type*, with respect to its severity and *information availability* as critical factors in estimating the need for explanation and generating the appropriate explanation upon failure. While the literature on robot failures and trust evaluation considers failure severity to be an important factor that influences trust [24], we further observe that situational awareness [25] and the information availability play an important role too. As a result, we conclude that: 1) if people can understand the failure and its resolution from the onset of failure, their performance is not influenced by the amount of provided explanations, and 2) more explanation does not automatically lead to better performance.

## B. Impact of Explanation Strategy

To understand how different explanation strategies performed, we analyzed participants' performances in later rounds, i.e., 3 and 4. In round 3, conditions C1, D1, and D2 had low-level explanations with different prior explanation levels. We observed that for carry failures, performances were not significantly impacted by the explanation strategy. At this point, participants were already familiarized with the cause of this type of failure and resolution, and given their quite high success rate in the first round, they just kept improving. On the other hand, we noticed that for place failure, where the High-level explanation was crucial to understanding the resolution, having a prior High-level explanation in conditions D1 and D2 improved the success rate. Consequently, the same improvement was observed in completing the task in a shorter time which was significant between conditions C1 and D2. Overall, we conclude that in a repeated interaction scenario, a user responds better to a low level of explanation after being exposed to a higher level of explanation in prior rounds. This presents a strong justification for explanation strategies that reduce the level of explanations which reduces the overall task completion time. Considering the results in condition D1, which included a Zero-level explanation in round 4, we conclude that not only the rate of reducing the explanation is important, but also the baseline where the explanation level reduces to.

#### C. Limitations and Future Work

Due to the exploratory nature of the study, we limited the number of possible conditions via pilot testing. Nevertheless, testing 5 conditions with 55 participants restricted us from drawing firm conclusions. While we observed some trends in the satisfaction ratings, having more participants wille enable us to surpass participants' personal differences. This study was the first step in identifying the variables involved in how non-expert users perceive explanations after robot failures and the findings help us to improve our understanding of robotic failures and explanations strategies. Next, we plan to isolate some of these variables to determine the optimal adaptation that leads to higher human satisfaction and performance. to better evaluate how humans perceive the explanation and what type of adaptation is needed. We are extending the research by analyzing the dataset composed of participants' behaviors from their participation in the study when encountering failures. We are aiming at using social cues to recognize if participants have detected the failures [26] and utilize that information in a closed-loop system to adapt the explanation in response to the human's reaction to the failure. Furthermore, we plan to conduct more user studies, investigating the conditions showing high varieties in performance and satisfaction ratings in more detail, including a detailed comparison of conditions C1, D1, and D2.

# VII. CONCLUSION

In this work, we investigate what levels of explanation and what explanation strategies in repeated interactions help non-experts to assist a robot to recover from failures in a collaborative task. We introduce two types of explanation strategies in the context of repeated interactions i.e. fixed and decaying and designed a collaborative task i.e., picking and placing objects from a table to shelves, where we incorporated three types of commonly occurring failures in such tasks. A user study with 55 participants evaluated three variations of the fixed and two variations of the decaying strategies, with failures in four rounds of interaction. The results portrayed a bigger picture of how participants' performances in resolving the failures and their satisfaction with the robot's explanation is a function of types of failure, level of explanation, and strategy. We observed, for failures with a more explicit resolution, the level of explanation did not influence participants' performance or satisfaction. However, for failures where the cause of the failure contributed to resolving it, performance in the task and satisfaction were directly impacted by the context of the explanation. With regard to explanation strategies, we noticed that specifically for complex failures that can be resolved with the high explanation, we can aim for decaying strategies, where we can avoid repetitions and reduce overall collaboration times. However, more modalities could be incorporated to decide the reduction rates, e.g. success rate in the previous rounds.

#### ACKNOWLEDGMENT

This work was partially funded by Digital Futures Research Center and Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH.

#### REFERENCES

- S. van Waveren, C. Pek, J. Tumova, and I. Leite, "Correct me if i'm wrong: Using non-experts to repair reinforcement learning policies," in *HRI*. ACM/IEEE, 2022, pp. 493–501.
- [2] A. Bauer, D. Wollherr, and M. Buss, "Human-robot collaboration: a survey," *Int. J. Humanoid Robot.*, vol. 5, no. 01, pp. 47–66, 2008.
- [3] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018.
- [4] D. Das, S. Banerjee, and S. Chernova, "Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery," in *HRI*. ACM/IEEE, 2021, pp. 351–360.
  [5] U. B. Karli, S. Cao, and C.-M. Huang, ""what if it is wrong": Effects
- [5] U. B. Karli, S. Cao, and C.-M. Huang, ""what if it is wrong": Effects of power dynamics and trust repair strategy on trust and compliance in hri," in *HRI*. ACM/IEEE, 2023.
- [6] C. Esterwood and L. P. Robert Jr, "Three strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness," *Computers in Human Behavior*, vol. 142, 2023.
- [7] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," AI magazine, vol. 40, no. 2, pp. 44–58, 2019.
- [8] T. Sakai and T. Nagai, "Explainable autonomous robots: a survey and perspective," Advanced Robotics, vol. 36, no. 5-6, pp. 219–238, 2022.
- [9] M. Eder and G. Steinbauer-Wagner, "A fast method for explanations of failures in optimization-based robot motion planning," in *Advances in Service and Industrial Robotics: RAAD 2022.* Springer, pp. 114–121.
- [10] M. Diehl and K. Ramirez-Amaro, "Why did i fail? a causal-based method to find explanations for robot failures," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8925–8932, 2022.
- [11] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, "Exploring the impact of fault justification in human-robot trust," in AAMAS, 2018, p. 507–513.
- [12] S. van der Woerdt and P. Haselager, "Lack of effort or lack of ability? robot failures and human perception of agency and responsibility," in *BNAIC 2016: Artificial Intelligence.* Springer, 2017, pp. 155–168.
- [13] S. Reig, E. J. Carter, T. Fong, J. Forlizzi, and A. Steinfeld, "Flailing, hailing, prevailing: Perceptions of multi-robot failure recovery strategies," in *HRI*. ACM/IEEE, 2021, pp. 158–167.
- [14] A. Alvanpour, S. K. Das, C. K. Robinson, O. Nasraoui, and D. Popa, "Robot failure mode prediction with explainable machine learning," in *CASE*. IEEE, 2020, pp. 61–66.
- [15] T. Miller, "Contrastive explanation: A structural-model approach," *The Knowledge Engineering Review*, vol. 36, 2021.
- [16] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani, "Explainable embodied agents through social cues: a review," *Transactions on Human-Robot Interaction*, vol. 10, no. 3, pp. 1–24, 2021.
- [17] R. Linder, S. Mohseni, F. Yang, S. K. Pentyala, E. D. Ragan, and X. B. Hu, "How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking," *Applied AI Letters*, vol. 2, no. 4, p. e49, 2021.
  [18] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in
- [18] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *ICRA*. IEEE, 2011, pp. 3400–3407.
- [19] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "The emerging landscape of explainable ai planning and decision making," arXiv preprint arXiv:2002.11697, 2020.
- [20] P. Khanna, E. Yadollahi, M. Björkman, I. Leite, and C. Smith, "User study exploring the role of explanation of failures by robots in human robot collaboration tasks," *arXiv preprint arXiv:2303.16010*, 2023.
- [21] P. Khanna, M. Björkman, and C. Smith, "Human inspired grip-release technique for robot-human handovers," in *Int. Conf. on Humanoid Robots (Humanoids)*. IEEE/RAS, 2022, pp. 694–701.
- [22] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint* arXiv:1812.04608, 2018.
- [23] P. Thagard, "Extending explanatory coherence," *Behavioral and brain sciences*, vol. 12, no. 3, pp. 490–502, 1989.
- [24] C. G. M. Garza, "Failure is an option: How the severity of robot errors affects human-robot interaction," *PhD Thesis, Pittsburgh: Carnegie Mellon University*, 2018.
- [25] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.
- [26] A. Bremers, A. Pabst, M. T. Parreira, and W. Ju, "Using social cues to recognize task failures for hri: A review of current research and future directions," arXiv preprint arXiv:2301.11972, 2023.