2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Impact of Different Failures on a Robot's Perceived Reliability

Andrew Violette[1], Zhanxin Wu[1], Haruki Nishimura[2], Masha Itkina[2],
Leticia Priebe Rocha[2], Mark Zolotas[2], Guy Hoffman[3], Hadas Kress-Gazit[3]

***Abstract*— Robots fail, potentially leading to a loss in the robot's perceived reliability (PR), a measure correlated with trustworthiness. In this study we examine how various kinds of failures affect the PR of the robot differently, and how this measure recovers without explicit social repair actions by the robot. In a preregistered and controlled online video study, participants were asked to predict a robot's success in a pick-and-place task. We examined manipulation failures (slips), freezing (lapses), and three types of incorrect picked objects or place goals (mistakes). Participants were shown one of 11 videos—one of five types of failure, one of five types of failure followed by a successful execution in the same video, or a successful execution video. This was followed by two additional successful execution videos. Participants bet money either on the robot or on a coin toss after each video. People's betting patterns along with a qualitative analysis of their survey responses highlight that mistakes are less damaging to PR than slips or lapses, and some mistakes are even perceived as successes. We also see that successes immediately following a failure have the same effect on PR as successes without a preceding failure. Finally, we show that successful executions recover PR after a failure. Our findings highlight which robot failures are in higher need of repair in a human-robot interaction, and how trust could be recovered by robot successes.**

I. Introduction

Robots are increasingly used in the workplace, working alongside humans. Application areas include logistics [1], manufacturing [2], [3], and service industries [4], [5]. For example, shipment warehouses have introduced robots handing objects to humans who pack orders [6], and robots routinely perform deliveries in cities [7].

As these robots operate, they will invariably fail. Limitations to software, hardware, and the robots' interaction with the environment all lead to potential failures [8]. This can include misidentifying an object, selecting an inappropriate action, and misinterpreting the task. For example, when instructed to grasp a red mug on the table, the robot might instead pick up a yellow mug or push the mug rather than grasp it.

In this work, we classify robot failures based on taxonomies of human failure, in line with prior work in human-robot interaction (HRI) [9]. Reason et al. categorized human error into *slips*, *mistakes*, and *lapses* [10]. Slips occur when the plan is correct, but the agent suffers a mechanical failure during execution. Mistakes occur when the plan is wrong—manipulating the wrong object, or putting the correct object in the wrong location. Lapses occur when the plan is initially correct, but is 'lost' during execution—spending extended periods of time making no progress towards the goal.

Our research question is: do different types of failure have different effects on how reliable the robot seems to users? To answer this question, we showed users videos of different types of robot failures in a pick-and-place task. Each participant viewed one of 11 videos explained in Fig. 1—five types of Failure, five types of Failure+Success (a failure followed by a successful execution in the same video), or a successful execution video. This experimental manipulation was followed by two successful execution videos.

After each video, participants were asked to choose between betting on the robot's success or on a fair (50%) coin. If they chose the robot and the robot failed, or they lost the coin lottery, participants would lose $2 of their study compensation. This study design allowed us to evaluate the perceived reliability (PR) of the robot using an objective measure, rather than self-report. PR is a measure correlated to trustworthiness [11] and we evaluate it instead of trust in the context of HRI, as it can be construed without the need for anthropomorphizing the robot or introducing vulnerability.

We find that (a) mistakes are the least damaging of failures, (b) that success immediately after failure has the same impact on PR as success without failure, and (c) participants return to perceive the robot as reliable if it eventually succeeds, regardless of failure type.

Our findings can inform which types of failures are in higher need of repair strategies. Slips and lapses, in particular, produce large drops in PR, suggesting that they indicate an important flaw on the robot's part. In contrast, fluent physical execution of a wrong plan, on a wrong object, or with a wrong task location do not seem to faze participants as much. People also seem to be forgiving of robots that fail before eventually succeeding, even without explicit trust repair strategies. These findings can have important implications on the design of failure recovery techniques in HRI.

II. Related Work

Taxonomies of robot failure are delineated by cause and severity [12]. Cause-based taxonomies analyze the source of the failure by technical module (e.g., software, hardware) or agent (robot, human, interaction between the two) [8], [13]. Severity definitions measure failure based on the service being provided [14] and whether the intended service can be recovered [13], [15]. In this work, we adapt the taxonomy

[1]Department of Computer Science, Cornell University
[2]Toyota Research Institute (TRI).
[3] Sibley School of Mechanical and Aerospace Engineering, Cornell University

TRI provided funds to assist the authors with their research; this article solely reflects the opinions and conclusions of its authors, and not TRI or any other Toyota entity.

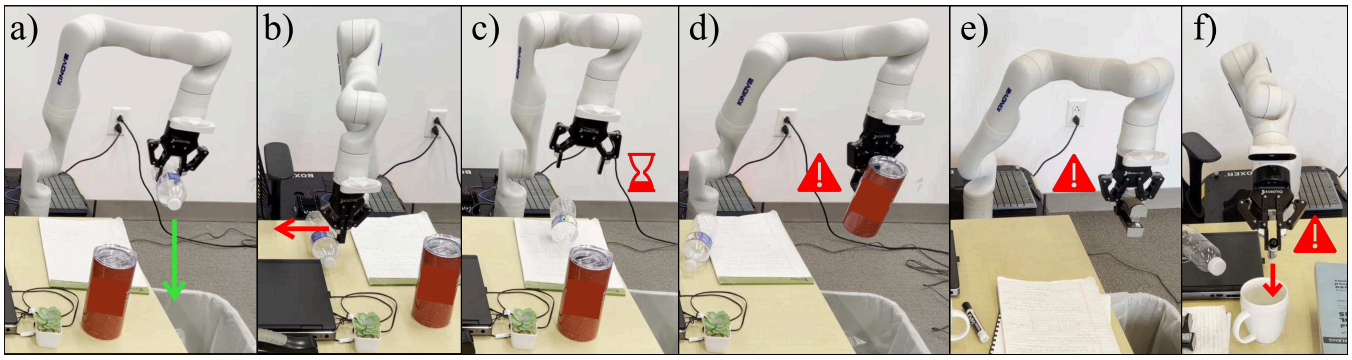

Fig. 1: 6 of the 11 conditions examined in this study. From left to right: a) **Success**. The robot puts the bottle in the trash. b) **Slip**. The robot is unable to pick up the bottle. c) **Lapse**. The robot positions to pick up the bottle, then freezes for 15 seconds. d) **Mistake (Thermos)**. The robot picks up the reusable thermos and places it in the trash. e) **Mistake (Stapler)**. The robot picks up the stapler and places it into the trash. f) **Mistake (Marker)**. The robot picks up the marker and places it into the mug. Not shown: The other five conditions, which are the failure modes (b–f) immediately followed by success.

of human failure from [10], which defines errors according to their primary cognitive processing stage—mistakes occur during planning, lapses occur when the initial plan is correct but is lost during storage, and slips occur when a plan is not executed correctly. [16] highlights these failure modes in executions of visual language action (VLA) models: *mistakes* ("tends to execute the wrong task"), *lapses* ("unable to move away from its starting pose"), and *slips* ("fails when attempting to grasp the object").

Do different types of robot failures have different impacts on PR? The results in the literature are mixed. Slips and mistakes had the same impact on user trust for a social approach task [17], and differed only in perceived benevolence for a risk detection task [18]. However, different types of errors impact the efficacy of different recovery strategies [19], and users express distinct emotional reactions depending on the severity of error [20]. We seek to expand the understanding of different failures by comparing slips, lapses, and multiple types of mistake.

Once mistakes occur, trust recovery techniques have primarily focused on social methods—for example, by apologizing [21], providing a justification for the mistake [22], ignoring the mistake [23], or promising to do better [24]. There is also a large body of work focused on trust repair strategies for specifically slips [19], [20].

Lapses and Failure+Success have only sparse representation in the literature. While delays in execution are well studied in the context of teleoperation [25], and in social interaction [26]–[28], there is limited research into how a delay in execution impacts PR for autonomous manipulation tasks outside of a social context. As the failure type with the second-largest impact on PR, lapses require further study. There is also a gap around *Failure+Success*—failure behavior followed immediately by successful task completion. This is called a physical trust repair strategy, and is most effective when performed automatically [29]. We examine how physical trust repair strategies impact PR for different types of failures without social intervention.

III. Method

Our research questions are: How do different kinds of failures, specifically slips, mistakes, and lapses, affect a robot's PR? How does this change if the robot successfully completes the task after the failure? To investigate these questions, we pre-registered two hypotheses before running the experiment described in this paper:[1]

1) When a user observes an autonomous robot completing a manipulation task, different types of failures, defined as slips, mistakes, and lapses, will each have a different impact on PR, and
2) Successes will have different impacts on PR depending on whether a failure precedes the success and depending on what type of failure occurred.

To address these research questions, we conducted an online video study. Participants (n=326, 161 male, 159 female, 6 other, aged 18-78, median:37) watched videos of a robot attempting to place a disposable plastic water bottle into a trash can as shown in Fig. 2. All videos were of a robot arm being teleoperated to demonstrate the experiment condition, although this was not specified to the participants.

A. Experiment Flow

The experiment flow is shown in Fig. 3. First, participants gave their consent and answered a 5-point scale question on their attitude toward introducing new technology into their lives. Then, they read a description of the robot's task, the betting mechanism, and their payout.

After the introduction, participants saw an alternating sequence of betting prompts and robot execution video screens. Before watching each video, the participant placed a bet on whether the robot would succeed, based on a description of the task and a photo of the workspace. They were asked whether they bet on "[t]he robot to put the water bottle in

[1]The preregistration can be found at the following link: AsPredicted 233286. In the preregistration, we use the phrase "unexpected behaviors" for the construct we call here, for clarity, "failures". We do not use either term in experimental materials presented to participants.

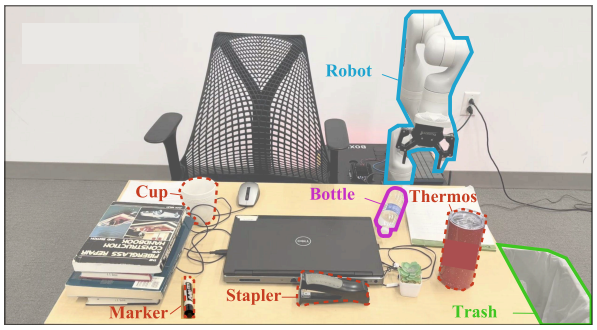

Fig. 2: The experimental setup. The robot (blue) attempted to place the disposable plastic water bottle (pink) into the trash (green). In some videos, the robot interacted with other objects (dotted red). Participants saw an unlabeled version of this before placing bets.

the trash" or "[a] random 50/50 coin flip". Each bet wagered $2 out of their initial $10 participant payout.

When placing their bet, they also answered a 5-point Likert item on how confident they are in their bet, from "not at all confident" (1) to "very confident" (5). In addition, they provided a short open-ended response explaining their reasoning for both the choice and their confidence level.

The experimental manipulation occurred by randomly assigning participants to see one of eleven "condition videos", (third from the left in Fig. 3).

After seeing the condition video, participants also evaluated what happened according to a series of questions:

1) Did the robot successfully move the plastic water bottle from the table into the trash? (Yes or No)[2]
2) Did the robot spend extended periods of time without making any progress? (Yes or No)
3) Did the robot do something other than moving the water bottle into the trash? (Yes or No)
4) Did the robot have difficulty physically moving the water bottle? (Yes or No)
5) If the robot failed to complete the task, can you describe what happened? (Short Answer)[3]

The participant then repeats the same betting procedure with two success videos. After viewing all three videos, the user was asked to evaluate the odds of success of the robot in the future, along with a confidence score and an open question about their reasoning—similar to the betting screen, but without the actual bet.

After every bet, participants were informed about the remaining payout, which decreased by $2 every time they lost betting on the robot. Coin flips were counted throughout the study, but only performed once at the end of the study.

B. Experimental conditions

We examined 11 conditions split into three categories: Failure, failure with subsequent success ("Failure+Success"), and Success. For failure conditions, we examine slip, lapse, and three different types of mistakes. [4] For Failure+Success conditions we examine each of these failures immediately followed by a success. Finally, we examine direct success with no preceding failure. We assign participants evenly to all conditions through balanced randomization. Videos for each condition are attached as supplemental material.

1) Failure Conditions: Participants in the Failure category saw one of the five failure conditions below, also summarized in Fig. 1.

1) **Slip:** attempting the requested task, but failing to mechanically execute it. The robot fails to open its gripper, pushing the bottle instead of picking it up.
2) **Lapse:** extended periods of time spent without making any progress on a task. The robot moves towards the bottle, then freezes, eventually returning to its resting position.
3) **Mistake (Thermos):** completing a different task. The robot places a reusable thermos into the trash can instead of the water bottle.
4) **Mistake (Stapler):** The robot places a stapler in the trash can instead of the water bottle.
5) **Mistake (Marker):** The robot places a dry erase marker into a mug instead of placing the the water bottle into the trash can.

C. Dependent Variable: From Bet to PR

After each video, we are interested in measuring the robot's PR, which we define as the participant's belief that the robot is going to succeed in its task. To recap, we ask users to choose between a "fair coin" lottery, which has a 50% chance of success, and betting on the robot. In each case, if the lottery loses or the robot fails, respectively, the participant loses $2 out of their total $10 study participation payout. Participants were also asked to state their confidence on a scale of 1 to 5 in the choice of their bet.

The PR score used in the remainder of this paper is the participant's "signed confidence" [30], calculated as the binary decision (-1 for lottery, 1 for robot) multiplied by the participant's confidence score. This is equivalent to an 11-point Likert item, excluding level 0, from "very confident in the lottery" (-5) to "very confident in the robot" (5).

While there is significant debate in how to understand and measure a person's confidence in a decision or their belief in an outcome [31], [32], we use existing practices to interpret "signed confidence" [30] as the participant's confidence in the robot's success. Choosing lotteries is an established mechanism in behavioral economics [33], which—under common assumptions of rationality—states that betting on the robot reflects at least a 50% belief that the robot would succeed. The confidence measure then correlates with the

[2]This question was shown for all videos. The others were only shown in the first, condition video.

[3]This question was only shown if the participant answered 'No' to Q1.

[4]The original protocol included three failure conditions—a slip, a lapse, and a mistake in which the robot placed a thermos, instead of the water bottle, into the trash. We found that 42% of participants interpreted the thermos mistake as a successful execution. To better understand how users perceive an actual mistake condition, we added two more versions: placing a stapler in the trash and placing a marker in a cup. This resulted in a total of five failure conditions and eleven total conditions.

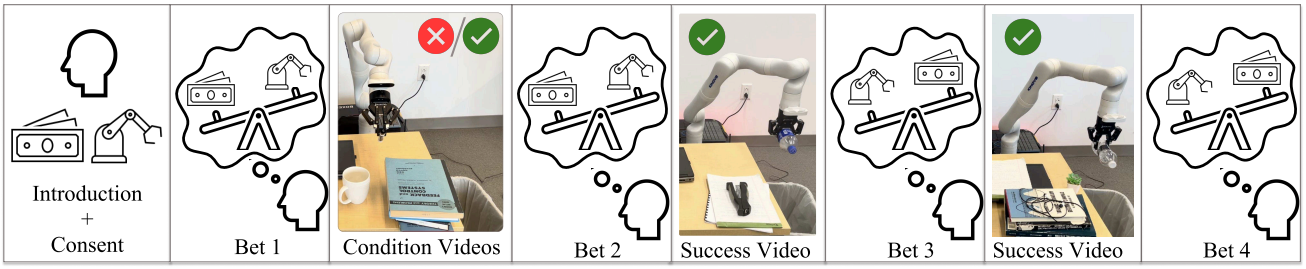

Fig. 3: The flow of the experiment. Participants were asked to place bets on robot success before and after seeing three videos—one of the experimental conditions, then two success videos.

distance from this 50% mark. For more discussion on the signed confidence measure, see, e.g., [34].

In contrast to most HRI studies, which rely primarily on subjective measures without any real stakes for participant choices, our PR metric design uses a strictly objective measure, and one with monetary risk associated with it, similar to [35].

Finally, we measure PR before and after seeing a video with the experimental condition to determine the impact of each condition. This also controls for dispositional difference in each participant by treating the first bet as a baseline.

D. Statistical Testing

We preregistered three different statistical tests on the difference between PR before and after the condition video: 1) a one-way ANOVA comparing all Failure conditions, 2) a one-way ANOVA comparing all Failure+Success and Success conditions, and 3) an independent Welch's t-test comparing conditions without a success (Failure) against the conditions with a success (Success and Failure+Success)[5]. We meet normality assumptions by collecting 30 samples from each condition. Both success and failure ANOVA groups satisfy homoscedasticity assumptions (Levene's test, F(4,140)=0.97, $p = 0.43$ for failure and F(5,175)=0.96, $p = 0.44$ for success). To control for the rate of false positives, we compute q-values by applying Holm-Bonferroni [36] correction to p-values of the three tests. Our threshold for strong evidence of an effect is $q < 0.05$.

In addition to the preregistered tests, we ran an exploratory independent t-test on the final bet (after seeing all videos) to determine if failure has a lasting impact. This test compared the conditions without a success (Failure) against the conditions with a success (Success and Failure+Success). This was not included in the false-positive correction calculation.

E. Affinity Diagramming

To better understand participant betting patterns, we used affinity diagramming [37] on the free response portion of each bet. We capture key observations using quotes from participants' responses, then group quotes into clusters to extract larger trends. Finally, we convert the larger trends into codes, and quantify how many participants fit each trend.

F. Participants

We recruited participants through the online recruitment platform Prolific[6]. We collected data from 357 participants, of which 20 did not complete the survey and 11 were removed for suspected large language model (LLM) usage. Thus, we analyze the data of 326 participants (161 male, 159 female, 6 other, aged 18-78, median:37). Participants started with $10 (USD) in their 'compensation account'. Since participants placed three bets, each worth $2, they received between $4 and $10 depending on their bets, with an average compensation of $8.61.

To eliminate inattentive participants and clean up our qualitative response pool, we screened participants by detecting Generative AI usage based on four criteria: The first is including common parts of LLM output, such as "Great! Here's a sample explanation you can use or adapt [...]" The second is writing in the second person, such as "Opting for the 50/50 coin flip suggests that you preferred a guaranteed probability rather than relying on the robot's past performance." The third is referencing videos that have not been shown. For example, one participant wrote this before seeing any videos: "According to the earlier videos, the robot seemed to be fairly steady in its efforts [...]" The last criteria is talking about robot demonstrations as if they have not happened. For example, one participant wrote this after seeing three robot demonstrations: "After observing the prior two setups and performance patterns (assuming some demonstration occurred), there may now be enough evidence to judge the robot's reliability." Generally speaking, LLM-generated responses seem to be written by a third party with only secondhand knowledge of the experiment.

IV. Findings

Our main pre-registered hypotheses (Section III) propose that different types of failures affect PR in distinct ways, and that successes will have different impacts on PR depending on whether a failure precedes the success and depending on its type. In this section, we present quantitative findings

[5]Code and output for all statistical tests can be found at www.github.com/violetteavi/impact-of-different-failures-analysis

[6]www.prolific.com

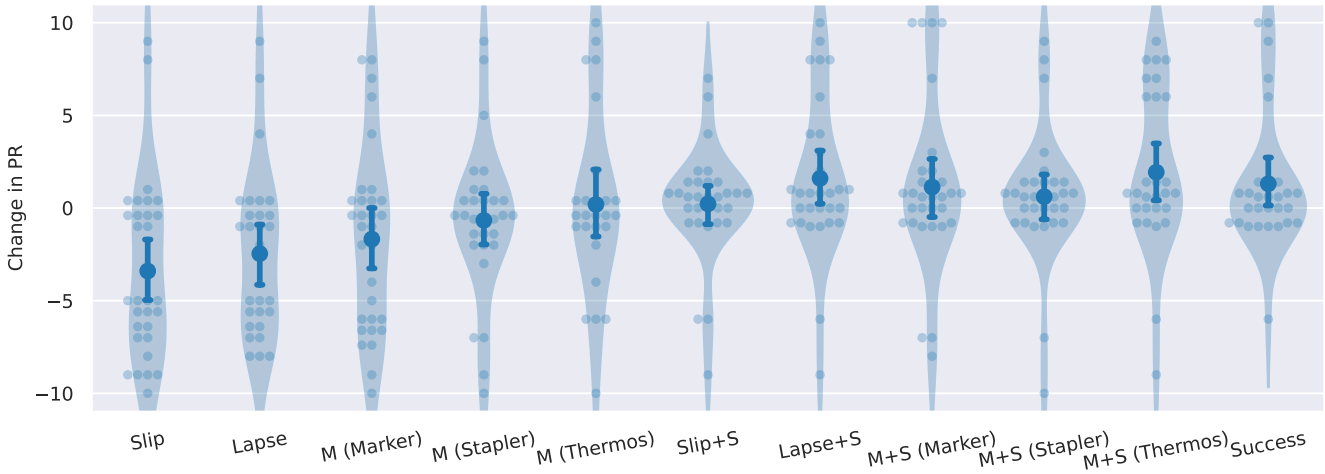

Fig. 4: Change in participant bet after they see the condition video, compared to the bet before. Each light dot represents one participant, the dark dot is the mean of participants, and the whiskers are the 95% confidence interval of the mean. We see evidence for differences in effect of failure conditions (slip, mistake, and lapse), with no evidence for difference in effect if the demonstration contains success. Mistake and Success are abbreviated as M and S, respectively.

	Failure	Failure+Success
Slip	-3.40 ± 0.84	$+0.23 \pm 0.69$
Lapse	-2.46 ± 0.87	$+1.60 \pm 0.70$
Mistake (Marker)	-1.68 ± 0.82	$+1.13 \pm 0.69$
Mistake (Stapler)	-0.67 ± 0.84	$+0.61 \pm 0.69$
Mistake (Thermos)	$+0.19 \pm 0.90$	$+1.93 \pm 0.71$
No Failure		$+1.31 \pm 0.71$

TABLE I: $\Delta_{\text{PR}}^{1\to 2}$, the difference in perceived reliability (PR) after seeing the conditional video across conditions, reported as mean $\pm$ standard error.

testing these hypotheses using the PR measure described above, and contextualize these results with qualitative coding from participants' responses to open-ended questions. We then present findings from our surveys that offer additional insights on how participants perceived robot failures.

A. Type of failure matters

We find that different types of failure had different effects on PR. As preregistered before the study, we calculate $\Delta_{\text{PR}}^{1\to 2}$, the difference in PR score between the second bet (after seeing the conditional video) and the first bet (which can be considered a participant's baseline). This measure gives us the specific change in PR for each experimental condition. Comparing this measure across the five failure conditions, we find weak evidence for a difference (One-Way ANOVA, F(4,140)=2.677, $p = 0.034$, $q = 0.068$). Fig. 4 shows $\Delta_{\text{PR}}^{1\to 2}$ for all conditions; failures are on the left.

Of the failure modes, Slip and Lapse caused the largest decreases in PR (-3.40 ± 0.84 and -2.46 ± 0.87, respectively). Mistake had the least detrimental effect on PR. In the Mistake (Thermos) condition, PR was nearly the same after seeing a failure ($+0.19 \pm 0.90$, see Fig. 4 middle). Mistake (Stapler) and Mistake (Marker) caused more modest decreases in PR than Slip and Lapse (-0.67 ± 0.84 and -1.68 ± 0.82, respectively).

Open responses support these results. Almost half of the participants in the Slip and Lapse condition cited past failure as a reason to bet against the robot (14/30 participants and 12/28, respectively). For example, P88 of the lapse condition said "The robot failed a task very similar to this one in the first trial." In contrast, much fewer participants in the Mistake (Thermos) and Mistake (Stapler) conditions wrote that failure influenced their bet (5/26 for thermos, 3/30 for stapler).

B. Mistakes are not all the same

It is worth noting the difference in $\Delta_{\text{PR}}^{1\to 2}$ between the three types of mistakes. Placing a reusable thermos in the trash resulted in no harm to PR ($+0.19 \pm 0.90$), and putting a stapler in the trash caused a negligible decrease in PR (-0.67 ± 0.84). The biggest decrease in PR for a mistake occurred for placing a marker in the cup. We present possible explanations for this finding in the Discussion.

C. Given eventual success, failures do not matter

Comparing the combined average $\Delta_{\text{PR}}^{1\to 2}$ for all Failure conditions (M=-1.64, SD=3.83) to the average for the combined data including all Failure+Success conditions and the direct Success condition (M=$+1.12$, SD=4.72), we find, as expected, that the latter is markedly better for PR than the former ($\delta = 2.76 \pm 0.48$, t(324)=5.84, $p < 0.001$, $q < 0.001$). Contrast, for example, Fig. 4 left and right. Any success in the video increases the average PR compared to the baseline.

This is also supported by the open response coding—participants cite the robot's eventual performance when justifying their bets: 43% of participants in the Failure+Success cases (78/181) indicated that prior success was the primary reason they bet on the robot. For example, P42 (Lapse+Success) said "After seeing the previous robot being able to remove a water bottle [...] I have faith in the robot's success." In contrast, 37% of participants in the Failure conditions (54/145) indicated that prior failure was

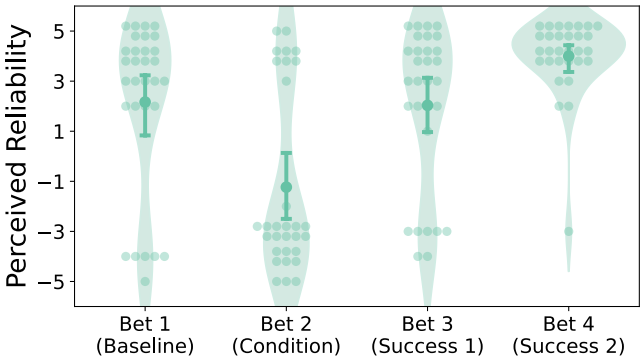

Fig. 5: Participant betting behavior before any videos (bet 1), after the Slip video (bet 2), and after additional success videos (bets 3 and 4). Perceived reliability (PR) drops after seeing the robot slip. PR rises after seeing success, ending above the baseline from Bet 1.

the primary reason they bet against the robot—P88 (Slip) wrote that "The robot failed earlier, let me see whether the coin flip will at least be a better bet."

If the robot succeeds, the different conditions are statistically indistinguishable. Comparing $\Delta_{\text{PR}}^{1\to 2}$ for all Failure+Success conditions along with the direct Success condition, did not suggest any difference (One-Way ANOVA, F(5,175)=0.81, $p = 0.543$, $q > 1$).

D. Success is an effective recovery method

We find that seeing successful execution after the conditional video is an effective recovery technique, matching existing literature [38]. This trend can be seen in Fig. 5, even for the biggest drop in PR (Slip). Nearly all participants (310/326, 95%) bet on the robot with a confidence rating of at least 3/5 after two success videos.

This trend is also reflected in the free response data for the failure cases. Almost no participants in Failure conditions (5/145, 3%) indicated, directly after seeing the failure, past success as the reason they would bet on the robot. This rose to 62% of participants in Failure conditions (90/145) after seeing two successes. P342 (Mistake (Marker)) illustrates this: "The robot has demonstrated twice after the initial failure that it is capable of performing the task. Thus, the error from the first attempt has been resolved." All conditions showed a higher PR than the participants' baseline after two successes. However, there remained a gap in PR between the Failure (M=3.88, SD=1.46) and Failure+Success/Success (M= 4.58, SD=0.87) conditions after two successful robot executions ($\delta = 0.70 \pm 0.13$, $t(324) = 5.34$, $p < 0.001$).

E. Failure Perception

Finally, we analyze the responses to the survey questions in which participants were asked to describe what happened in the video before learning the outcome of their bet. To determine how participants perceive the robot's performance, we compare the experimental conditions as we intended them to a series of questions answered by participants (see Section III-A). Participants tended to agree with the conditions'

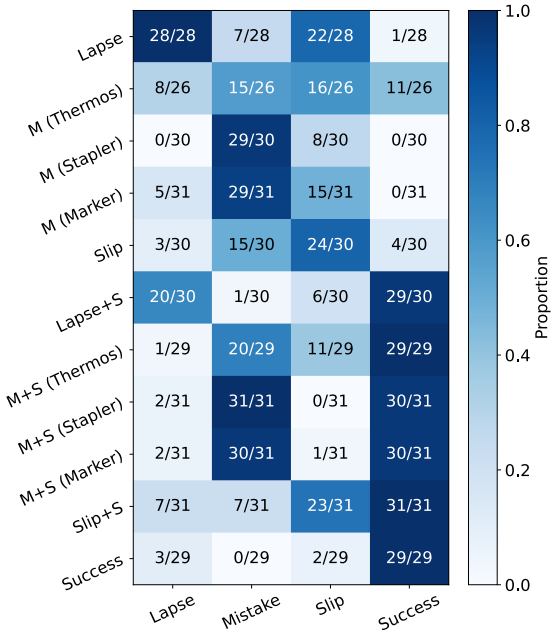

Fig. 6: Participant evaluation (horizontal) matched our labels (vertical), though participants often labeled lapses and mistakes as slips. Mistake (M) and Success (S) are abbreviated.

labels of failure modes, with some notable exceptions. Fig. 6 shows a breakdown of these disagreements.

Participants agreed on the success label, with 94% (307 of 326) participants matching the experimenter's labels of the video. The 'success' label disagreement largely lies in the Mistake (Thermos) condition, in which the robot places a reusable thermos with a handle into the trash. 42% of participants viewed this as a successful execution, despite the disposable plastic bottle remaining on the table.

Another area in which our label did not agree with participants' interpretations is that many applied the statement that the robot had "difficulty physically moving the water bottle"—which we associated with the Slip condition—to Mistakes and Lapses. For 29% of participants (95 out of 326), there is a mismatch between their condition and their response to the above statement. This mismatch primarily comes from participants labeling other failure modes as slips. This number is lower for Lapse (13%) and Mistakes (17%).

V. Discussion

Why do the different types of mistakes lead to different drops in PR? While slips and lapses notably decrease PR, mistakes are overall less harmful to PR, but some mistakes are worse than others. We have limited information, having compared only three types of mistakes, but can speculate possible explanations for the difference between mistakes.

First, the severity of loss in PR may be related to the semantic similarity of the mistaken task relative to the intended task. Participants may view putting a thermos into the trash more positively than putting a stapler into the trash because the thermos, like a bottle, is an object that can contain liquids for drinking.

Alternatively, a mistake in object handled might be less crucial than one in the goal position. So, as long as the goal is correct, even if the wrong object was picked up, the overall task seems doable by the robot. This could explain why putting the wrong object in the trash indicates higher reliability than putting the wrong object in a mug.

A third explanation is an implicit breakdown of the task into smaller pieces—the participant may separate pick-and-place into a pick, and a place. In the Mistake (Thermos) and Mistake (Stapler) conditions, the robot may get partial credit for the "place" part. The Mistake (Marker) condition would not get this credit, as it picks the wrong object and places into the wrong location.

Open responses suggest that mistakes can indeed demonstrate some capability, even if they do not achieve the task. For example, P308 of the Mistake (Stapler) condition wrote: "I saw it successfully move the stapler in the trash, so I believe in its abilities but not so sure it will pick up the right item... I know it is capable but not sure overall."

Some participants indicated that the Mistake (Marker) case occurred due to disobedience instead of a lack of capability. This raises questions around what is considered a robot's motivation versus its ability, a known distinction in human trust theory [39]. When thinking about a robot's ability, do humans assume that correctly processing language instruction is an inherent robot capability, perhaps due to the proliferation of strong natural language systems? Does this expectation extend to other semantic tasks, such as object identification via computer vision?

Finally, the mismatch between condition labels and user interpretations (Section IV-E) suggest that people consider "cognitive" failures, such as lapses or mistakes as "physical" breakdowns. This might indicate, again, that people working with robots may distinguish between what a robot knows and what a robot knows how to do.

In summary, how do failures matter? Though the initial effect of each failure mode was different, successful executions recovered PR. This provides evidence for leniency: As long as the task is eventually completed, participants were accepting of failures that occurred along the way. However, this recovery does not imply robustness to future failure. Recency bias in HRI is well-documented [40]. PR is unlikely to remain high if the robot fails repeatedly. Understanding which failures need mitigation is still important.

A. Limitations and Future Work

Our study has several limitations. First, participants get paid based on whether the study labels the robot behavior as a success, and they see this judgment in real time. Therefore, the betting measure might not necessarily measure the user's perception of robot probability to succeed according to their own standards, but their desire to maximize compensation by guessing what the experiment considers a success. That said, only 19 out of 326 participants (6%) labeled success differently from the experiment design, so it is unlikely that these differences significantly impacted our findings.

Another limitation is that our study was conducted online, which may not generalize to in-person HRI. Some experiments show that effects in online studies are mirrored in physically situated studies, and only vary in the magnitude of the effect (e.g., [41]). Others suggest more qualitative differences between video and in-person interactions with robot (e.g., [42]). In both cases, there may have been effects we were unable to detect.

Finally, while this study illuminates differences in PR after failure, it does not provide participants with the ability to interact with the robot. Interactivity may change the results of the study either through premature termination of the robot's activity or even just through a stronger sense of control over the robot. For example, if the human sees the robot grabbing a stapler instead of the water bottle, they may stop the execution. This would convert a Mistake+Success into a Mistake, as the robot does not have the opportunity to succeed once a failure occurs. The timing of human intervention during these failures could be further studied.

VI. Conclusion

In this study, we compared different robot failure conditions using a pre-registered and controlled online video experiment. Our findings reveal that not all robot failures are perceived equally when it comes to the robot's perceived reliability (PR): physical slips are most detrimental, followed by robot freezes (lapses). The failures that were least harmful for the robot's PR were mistakes that could be attributed to object recognition or plan errors.

All that said, we also find that, when failures are followed by a successful execution of the task, they do not harm PR and are statistically indistinguishable from successes. Moreover, PR recovers rapidly when the robot succeeds after a failure, even in the absence of explicit repair behaviors, such as apologies or explanations.

This result highlights a critical nuance in human-robot interaction: not all robot failures are equal and need to be treated with the same repair mechanism. Recovery from failure does not even necessarily require overt repair mechanisms. Instead, consistent successful performance can restore a robot's reliability perceptions over time. These insights shed light on which types of robot failures matter most in shaping user perceptions and inform priorities when designing interaction strategies that mitigate the impact of failure.

Moreover, our findings motivate existing questions about how people interpret robot failures——particularly in terms of perceived robot ability versus motivation, and the distinctions between physical and cognitive capabilities. Understanding these interpretive frames is essential for developing robots that can effectively manage user expectations and maintain trust in human-robot interaction.

References

[1] C. Eppner, S. Höfer, R. Jonschkowski, R. Martín-Martín, A. Sieverling, V. Wall, and O. Brock, "Lessons from the amazon picking challenge: Four aspects of building robotic systems." in *Robotics: science and systems*, vol. 12, 2016.

[2] J. Arents and M. Greitans, "Smart industrial robot control trends, challenges and opportunities within manufacturing," *Applied Sciences*, vol. 12, no. 2, p. 937, 2022.

[3] L. Liu, F. Guo, Z. Zou, and V. G. Duffy, "Application, development and future opportunities of collaborative robots (cobots) in manufacturing: A literature review," *International Journal of Human–Computer Interaction*, vol. 40, no. 4, pp. 915–932, 2024.

[4] E. Moriuchi and S. Murdy, "The role of robots in the service industry: Factors affecting human-robot interactions," *International Journal of Hospitality Management*, vol. 118, p. 103682, 2024.

[5] J. A. Gonzalez-Aguirre, R. Osorio-Oliveros, K. L. Rodríguez-Hernández, J. Lizárraga-Iturralde, R. Morales Menendez, R. A. Ramirez-Mendoza, M. A. Ramirez-Moreno, and J. d. J. Lozoya-Santos, "Service robots: Trends and technology," *Applied Sciences*, vol. 11, no. 22, p. 10702, 2021.

[6] R. Allgor, T. Cezik, and D. Chen, "Algorithm for robotic picking in amazon fulfillment centers enables humans and robots to work together effectively," *INFORMS Journal on Applied Analytics*, vol. 53, no. 4, pp. 266–282, 2023.

[7] A. Schneider, A. Robinson, C. Grimm, and N. T. Fitter, "How Do Starship Robots Affect Everyday Campus Life? An Exploratory Posting Board Analysis and Interview-Based Study," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, Aug. 2024, pp. 528–534.

[8] G. Steinbauer, "A survey about faults of robots used in robocup," in *Robot Soccer World Cup*. Springer, 2012, pp. 344–355.

[9] N. C. Krämer, A. Von Der Pütten, and S. Eimler, "Human-Agent and Human-Robot Interaction Theory: Similarities to and Differences from Human-Human Interaction," in *Studies in Computational Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 215–240.

[10] J. Reason, *Human Error*. Cambridge University Press, Oct. 1990.

[11] P. A. Hancock, T. T. Kessler, A. D. Kaplan, K. Stowers, J. C. Brill, D. R. Billings, K. E. Schaefer, and J. L. Szalma, "How and why humans trust: A meta-analysis and elaborated model," *Frontiers in Psychology*, vol. 14, Mar. 2023.

[12] 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, Jun. 2018.

[13] J. Carlson and R. Murphy, "How UGVs physically fail in the field," *IEEE Transactions on Robotics*, vol. 21, pp. 423–437, Jun. 2005.

[14] J.-C. Laprie, "Dependability of computer systems: concepts, limits, improvements," in *Proceedings of Sixth International Symposium on Software Reliability Engineering. ISSRE'95*. Toulouse, France: IEEE Comput. Soc. Press, 1995, pp. 2–11.

[15] O'Hare, G. M. P., (Greg M. P.), Collier, Rem, Ross, Robert, "Demonstrating Social Error Recovery with AgentFactory," in *Proceedings of 3rd International Joint Conference on Autonomous Agents and Multi Agent System. AAMAS'04*. IEEE, 2004.

[16] . TRI LBM team, "A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation," 2025, arXiv:2507.05331 [cs].

[17] R. Flook, A. Shrinah, L. Wijnen, K. Eder, C. Melhuish, and S. Lemaignan, "On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy?" *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, vol. 20, no. 3, pp. 455–486, Nov. 2019, publisher: John Benjamins Publishing Company.

[18] E. Kox, M. Hennekens, J. Metcalfe, and J. Kerstholt, "Trust Violations due to Error or Choice: The Differential Effects on Trust Repair in Human–Human and Human–Robot Interaction," *ACM Transactions on Human-Robot Interaction*, no. 4, pp. 1–27, 2025.

[19] X. Zhang, S. K. Lee, H. Maeng, and S. Hahn, "Effects of Failure Types on Trust Repairs in Human–Robot Interactions," *International Journal of Social Robotics*, vol. 15, no. 9-10, pp. 1619–1635, 2023.

[20] M. Stiber and C.-M. Huang, "Not All Errors Are Created Equal: Exploring Human Responses to Robot Errors with Varying Severity," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*. Netherlands: ACM, Oct. 2020, pp. 97–101.

[21] P. Fratczak, Y. M. Goh, P. Kinnell, L. Justham, and A. Soltoggio, "Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction," *International Journal of Industrial Ergonomics*, vol. 82, p. 103078, Mar. 2021.

[22] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, "Exploring the impact of fault justification in human-robot trust," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '18. International Foundation for Autonomous Agents and Multiagent Systems, 2018, p. 507–513.

[23] S. Engelhardt, E. Hansson, and I. Leite, "Better faulty than sorry: Investigating social recovery strategies to minimize the impact of failure in human-robot interaction." in *WCIHAI@ IVA*, 2017, pp. 19–27.

[24] 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 *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '23. ACM, Mar. 2023, pp. 271–280.

[25] E. Yang and M. C. Dorneich, "The Effect of Time Delay on Emotion, Arousal, and Satisfaction in Human-Robot Interaction," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1, pp. 443–447, Sep. 2015, publisher: SAGE Publications.

[26] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "How Quickly Should a Communication Robot Respond? Delaying Strategies and Habituation Effects," *International Journal of Social Robotics*, vol. 1, no. 2, pp. 141–155, Apr. 2009.

[27] D. Kang, C. Nam, and S. S. Kwak, "Robot Feedback Design for Response Delay," *International Journal of Social Robotics*, vol. 16, no. 2, pp. 341–361, Feb. 2024.

[28] H. Pelikan and E. Hofstetter, "Managing Delays in Human-Robot Interaction," *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 4, pp. 1–42, Aug. 2023, publisher: Association for Computing Machinery (ACM).

[29] S. Lane, C. Esterwood, D. Kulić, and N. Robinson, "Robots That Use Physical Repair Strategies After Repeated Errors to Mitigate Trust Decline in Human-Robot Interaction: A Repeated Measures Experiment," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, 2024, pp. 936–943.

[30] L. Aitchison, D. Bang, B. Bahrami, and P. E. Latham, "Doubly bayesian analysis of confidence in perceptual decision-making," *PLOS Computational Biology*, vol. 11, no. 10, p. e1004519, Oct. 2015.

[31] A. Kepecs and Z. F. Mainen, "A computational framework for the study of confidence in humans and animals," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1594, p. 1322–1337, May 2012.

[32] A. Pouget, J. Drugowitsch, and A. Kepecs, "Confidence and certainty: distinct probabilistic quantities for different goals," *Nature Neuroscience*, vol. 19, no. 3, p. 366–374, Feb. 2016.

[33] C. A. Holt and S. K. Laury, "Risk aversion and incentive effects," *American Economic Review*, vol. 92, no. 5, p. 1644–1655, Nov. 2002.

[34] P. Mamassian and V. de Gardelle, "Modeling perceptual confidence and the confidence forced-choice paradigm." *Psychological Review*, vol. 129, no. 5, p. 976–998, Oct. 2022.

[35] A. Kshirsagar, B. Dreyfuss, G. Ishai, O. Heffetz, and G. Hoffman, "Monetary-incentive competition between humans and robots: Experimental results," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Mar. 2019, p. 95–103.

[36] Holm, Sture, "A Simply Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, vol. 6, pp. 65–70, 1979.

[37] A. Lucero, "Using Affinity Diagrams to Evaluate Interactive Prototypes," in *Human-Computer Interaction – INTERACT 2015*, J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, and M. Winckler, Eds. Cham: Springer International Publishing, 2015, vol. 9297, pp. 231–248, series Title: Lecture Notes in Computer Science.

[38] H. Pan, K. Xu, Y. Qin, and Y. Wang, "How does drivers' trust in vehicle automation affect non-driving-related task engagement, vigilance, and initiative takeover performance after experiencing system failure?" *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 98, pp. 73–90, Oct. 2023.

[39] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction*. Elsevier, 2021, pp. 3–25.

[40] M. B. Luebbers, A. Tabrez, K. S. Talanki, and B. Hayes, "Recency Bias in Task Performance History Affects Perceptions of Robot Competence and Trustworthiness," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 274–11 280.

[41] Y. Hu and G. Hoffman, "Using skin texture change to design emotion expression in social robots," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 2–10.

[42] M. Bretan, G. Hoffman, and G. Weinberg, "Emotionally expressive dynamic physical behaviors in robots," *International Journal of Human-Computer Studies*, vol. 78, pp. 1–16, 2015.