Gazing at Failure: Investigating Human Gaze in Response to Robot Failure in Collaborative Tasks

Ramtin Tabatabaei
The University of Melbourne
Melbourne, Australia
stabatabaeim@student.unimelb.edu.au

Vassilis Kostakos
The University of Melbourne
Melbourne, Australia
vassilis.kostakos@unimelb.edu.au

Wafa Johal
The University of Melbourne
Melbourne, Australia
wafa.johal@unimelb.edu.au

***Abstract*—Robots are prone to making errors, which can negatively impact their credibility as teammates during collaborative tasks with human users. Detecting and recovering from these failures is crucial for maintaining effective level of trust from users. However, robots may fail without being aware of it. One way to detect such failures could be by analysing humans' non-verbal behaviours and reactions to failures. This study investigates how human gaze dynamics can signal a robot's failure and examines how different types of failures affect people's perception of robot. We conducted a user study with 27 participants collaborating with a robotic mobile manipulator to solve tangram puzzles. The robot was programmed to experience two types of failures —executional and decisional— occurring either at the beginning or end of the task, with or without acknowledgement of the failure. Our findings reveal that the type and timing of the robot's failure significantly affect participants' gaze behaviour and perception of the robot. Specifically, executional failures led to more gaze shifts and increased focus on the robot, while decisional failures resulted in lower entropy in gaze transitions among areas of interest, particularly when the failure occurred at the end of the task. These results highlight that gaze can serve as a reliable indicator of robot failures and their types, and could also be used to predict the appropriate recovery actions.**

***Index Terms*—Robot Failures, Gaze Dynamics, Human-Robot Collaboration**

I. Introduction

As robotics advances, the potential for robots to assist people in various domains, such as manufacturing [1], [2], domestic assistance [3]–[5] is becoming increasingly evident. One significant application of robotics is in human-robot teaming, which involves collaboration between humans and robotic systems working together to perform joint activities [6]. In human-robot teaming, robots must behave and communicate effectively to maintain alignment within the team [7]. However, as robots become more integrated into our daily lives, ensuring the reliability of these systems is a pressing concern [8], [9]. Robot errors are inevitable, much like human errors, due to the inherent uncertainty of the world and the need to make decisions and act in real-time. If these failures are not managed appropriately, they can negatively impact task success, human safety, trust, and perceptions of the robot's intelligence [10]–[13]. These factors are crucial because the degree to which people trust robots influences their willingness to collaborate with them, which is essential for establishing effective human-robot teams [14]–[16]. However, trust can fluctuate over time; it tends to increase when robots perform well but might drop rapidly when they inevitably make errors. In addition, the type of failure (i.e. at the motion execution or task planning level), might significantly affect trust, and robots that demonstrate awareness of their errors show potential in restoring trust, as the saying goes, 'a fault confessed is half forgiven.' By focusing on moments when human interactions deviate from expected patterns, strategies can be identified to make these interactions more robust.

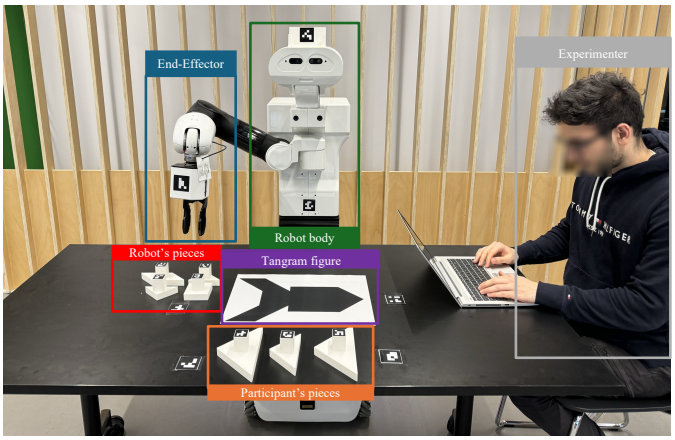

Fig. 1: Different areas of interest in the experiment.

One strategy to enhance human-robot interaction is modelling the user's reactions to robot failures. This user model can be inferred from various signals [17], such as the user's social cues during the period of the failure [18]. One of those social and non-verbal cues is a person's eye gaze [19], which plays a crucial role in conveying attention [20], [21], intentions [20]–[22], and emotional states [20], [23]. Eye gaze has been leveraged in human-robot interactions to enhance the robot's ability to comprehend and to anticipate human actions [21], [24]. Research has also shown that people exhibit consistent gaze patterns while performing specific activities, [25]–[27], making it possible to model these patterns. However, there is a lack of studies on accurate gaze patterns in response to robot failures, which could potentially aid robots in recovering from their failures.

This study examines the impact of robot failures on users' perception and gaze behaviour. Utilising a within-between experimental design, we analysed the effects of failure types, failure timing, and failure acknowledgement in a sample of 27 participants. Participants engaged in a collaborative task involving four Tangram puzzles, during which the robot was programmed to fail. Each participant experienced all combinations of failure type and timing (within-subjects), with one group exposed to the robot acknowledging its failures and the other group receiving no acknowledgement (between-subjects). Our results indicate that users' gaze patterns during failure events differ significantly from

those observed during times of no failure. Furthermore, these gaze behaviours are highly dependent on the type of failure.

II. RELATED WORKS

A. Social Signals to Robot's Failure

Social signals have been found to be reliable indicators of errors, as people react to robot errors socially due to their unexpectedness. Specifically, users display more social signals during situations with errors than those without [28]. Common instinctive responses to robot errors include gaze [29]–[31], facial expressions [31]–[34], verbalization [29], [31], [32], and body movements [32], [34], [35].

Several studies have demonstrated that participants exhibit distinct social signals in response to robot failures. For instance, Aronson et al. [36] observed that participants' gaze patterns deviate from the norm during unexpected robot actions. Wachowiak et al. [19] further found that during failures, participants focus more on the entity they are collaborating with, whereas in error-free scenarios, their gaze is more evenly distributed. Similarly, Peacock et al. [30] noted that gaze initially increases in motion during failures and then stabilizes as users recognize and correct the error. Stiber et al. [28] identified that specific facial muscles, such as those involved in smiling and brow lowering, become more active during robot errors. Kontogiorgos et al. [29], [31] found that robot failures lead to increases in spoken words, utterance duration, and gaze shifts towards the robot, indicating heightened engagement during errors. While there is substantial research on human reactions to interacting with a failing robot, there is limited understanding of how this interaction affects the perception of the robot as a teammate in highly collaborative tasks and its impact on human gaze behaviour. This gap is significant, as existing research indicates that individuals exhibit distinct social reactions depending on the type of robot failure encountered [31], [32].

B. Types of Failure

Robot failures can be classified into different types depending on the nature of the issue. Mirnig et al. [32] identified two main types: technical failures, where the robot fails to perform its task correctly, and social norm violations, which occur when the robot deviates from expected social behaviour. Honig and Oron-Gilad [8] offered a taxonomy distinguishing between (a) technical failures, involving hardware malfunctions and software issues, and (b) interaction failures, arising from uncertainties during interactions with the environment, other agents, or humans. Similarly, Tian et al. [37] categorised errors into performance errors, which affect perceived intelligence and task competence, and social errors, which impact socio-affective competence. Kontogiorgos et al. [29] further classified conversational failures into task-oriented failures, such as incorrect guidance or incomplete instructions, and social protocol violations, like disengagement. Additionally, Morales et al. [38] categorised robot failures into Personal Risk Failures (e.g., throwing objects or erratic movements), Property Risk Failures (e.g., dropping or crushing objects), and an Assistance scenario where the robot seeks participant help without posing direct risks.

C. Timing of Failure

Research has shown that the timing of failures during a task influences people's perceptions of the robot in various ways. Desai et al. [39] found that early failures significantly reduce trust and make it harder to recover compared to failures occurring later in the interaction. Similarly, Rossi et al. [40] observed that participants' trust in the robot did not increase when severe mistakes happened early in the interaction. In contrast, Morales et al. [38] discovered that the order of failures significantly impacts participants' perceptions, with severe failures occurring last leaving a stronger impression and making participants more likely to believe the robot will fail again in future tasks. Lucas et al. [41] also found that early errors can be somewhat recovered from, especially with positive social interaction, but late errors are more damaging. On the other hand, Kontogiorgos et al. [31] demonstrated that reactions to failures remain consistent, regardless of whether they occur early or late in the interaction. Existing research shows contrasting results regarding the effects of robot failure timing on trust, highlighting the need for further study to understand how failure timing affects user perception of the robot. Furthermore, to the best of the author's knowledge, no research has explored the impacts of failure timing on human gaze behaviour.

D. Failure Repair

Previous studies have investigated how different trust repair strategies used by robots influence users' perceptions. For example, LeMasurier et al. [42] considered three strategies for explaining failures: 1) The robot only acknowledges its failure, 2) The robot explains what went wrong and why after the failure, and 3) The robot predicts and explains potential failures before they occur. Their results highlight that both explaining and predicting failures enhance users' perceptions of a robot's intelligence and trustworthiness compared to providing no explanation at all. In [43], four trust repair approaches (promises, denials, explanations, apologies) were compared during a collaborative robot task. Apologies, explanations, and promises were similarly effective and outperformed denials for the ability measure, while apologies and promises were most effective for benevolence. Additionally, Wachowiak et al. [44] found that participants preferred apologies most and silence least when a robot made an error. While previous studies have shown that apologies and explanations for failures help people regain trust in the robot, we wonder if this could also affect their social behaviour, specifically their gaze behaviour.

E. Gaze in HRC

The gaze behaviour in human-robot collaboration has been studied widely, focusing on both the robot's gaze behaviour while collaborating with a human and the human's gaze behaviour while interacting with a robot [45], [46]. Most studies in human-robot collaboration emphasise the human's gaze behaviour, as it can indicate human intent and focus, allowing the robot to determine its next move and adapt its behaviour accordingly.

The literature collectively highlights the significant role of gaze in intent recognition. Huang et al. [16] focused on enabling robots to proactively perform task actions by predicting the task intent of their human partners based on observed gaze patterns. Their anticipatory control method significantly improved task efficiency, allowing the robot to respond faster compared to the reactive method. Additionally, Shi et al. [47] also developed an effective model for accurately determining which object a person intends to focus on during interactions with a robot. The

literature demonstrates that gaze can be a reliable indicator of a person's intent and by extension their anticipation of upcoming actions. This leads us to wonder whether gaze also has the potential to help the robot repair from its failure.

Despite extensive research on human reactions to robot failures, little is known about how such failures influence perceptions of the robot as a teammate or affect human gaze behavior. Moreover, the impact of failure timing and the robot's acknowledgment on user gaze remains unclear. To explore these gaps, we address the following research questions:

- **RQ1** How does human gaze behaviour change in response to different robot failures during a collaborative task?
- **RQ2** How do different robotic failures affect human perception of the robot as a teammate?

III. Methodology

A. Tasks Description

The experiment consists of four distinct tasks, in which one participant and a robot collaboratively solve Tangram puzzles. In each task, participants were required to create a unique shape using Tangram pieces. The sequence of shapes to solve is Rocket, Rabbit, Turtle, and Cat. We chose puzzles of similar difficulty to ensure the difficulty would not affect participants' perception and behaviour towards the robot.

Each Tangram puzzle consisted of seven pieces. The robot handled four pieces (two small triangles, a square, and a parallelogram), while the participant had to place correctly three pieces (a medium-sized triangle and two large triangles). The Tangram pieces were 3D printed, and when assembled, formed a square of 200mm in side. Each piece was 20mm high. Besides, an cube ($32\text{mm} \times 32\text{mm} \times 40\text{mm}$) was attached on top of each piece to act as a handle and to facilitate the robot's ability to pick up the pieces. The puzzles' silhouettes were printed in black on A2 white paper, slightly larger than the Tangram pieces to avoid the need for very precise placement, with approximately 1 cm clearance on each side. These papers were fixed to the table, and the participant and the robot had to place each of their pieces in the correct position. The participant was asked to move a piece only after the robot had completed its action.

The robot always placed the first piece. To reduce confusion about when the participant should place their piece, the robot said: "Now it is your turn.", after placing each piece, except for the last one, when it said: "Now, let's solve the next puzzle." If the participant placed a piece incorrectly, the robot responded,"You have placed the object in the wrong location."

The robot's pieces were placed next to the paper and near the robot, as shown in Figure 1. In each puzzle, the arrangement of the pieces varied from the previous one, and the robot first determined the placement and orientation of each piece before picking it up. To facilitate this process, an ArUco marker was attached to the top of each piece, allowing the robot to accurately locate them. The Tiago robot, programmed using ROS1, then utilised the tf library to transform the pose of the desired object to the coordinate frame associated with its arm, and subsequently employed inverse kinematics to move its arm to the correct location. The robot's head movements were pre-programmed to approximately mimic human gaze behaviour. During its turn, the robot maintained its gaze on the Tangram piece while picking it up and placing it. Once the robot finished placing a piece, it started looking at the participant.

B. Robot Failures

We designed the robot to fail during each task in its interaction with the participants. These failures varied based on their type, timing, and whether the robot acknowledged its failure or not.

1) ***Types of Failures****:* The types of failures in our experiment represent typical robot malfunctions that may occur during interactions and are commonly reported in HRI. In this research, the robot will simulate two distinct types of failures: 1) Executional failure (EF) and 2) Decisional failure (DF).

EF can be categorised as a technical failure, specifically timing and ordering [8]. In this scenario, the robot pauses for 15 seconds just before picking up an object, while keeping the object within its end effector. After the 15-second pause, the robot will resume and complete the task of picking up and placing the object. This type of failure aligns with previous research [19], [29].

DF can also categorized as a technical failure, where the robot performs the correct action incorrectly [8]. In this scenario, after picking up an object, the robot will mistakenly move to the location designated for a different object, place it and pause for 5 seconds. While still holding the object, the robot will then lift the object again and place it in the correct location. This type of failure aligns with previous research, in which the robot attempts to perform the correct action but executes it incorrectly [28], [48].

The procedure for picking up and placing objects during failure events is identical to the procedure when no failure occurs, indicating that the robot shows no signs of committing a failure beforehand. The only distinction in the EF is a pause, which increases the total time for the pick-and-place task by 15 seconds. In the DF, the robot moves its arm to the wrong location, goes down, and comes back up, resulting in an overall increase of 16.5 seconds to the motion.

2) ***Timing of Failures****:* The literature suggests that the timing of a failure—whether it occurs at the beginning or the end of an interaction—can affect a person's perception of the robot differently. In this research, we aim to investigate how the timing of a failure impacts both gaze behaviour and user perceptions. Specifically, the robot may fail either at the beginning of the collaboration when placing its first piece, or towards the end of the interaction when placing its third piece.

3) ***Acknowledgement of Failures****:* A fault confessed is half redressed. Guided by this principle, we explored how the robot's ability to acknowledge its mistakes influences participants' perception and gaze patterns in subsequent failures. We designed two distinct scenarios. In one scenario, the robot demonstrates awareness of its mistakes by acknowledging each failure immediately after they occur. After a DF failure, it says, "Sorry, I made a mistake." and after an EF, it says, "Sorry for the delay." In the other scenario, the robot does not declare any of its failures. In both scenarios, the robot performs physical repairs.

C. Participants

We conducted *a priori* power analysis to calculate the sample size for our experiment using *G*Power* [49]. The calculation was based on a medium effect size of $f = 0.25$, an alpha level of 0.05, and a power of 0.8. As a result, we determined a minimum of 24 participants was required; however, we recruited

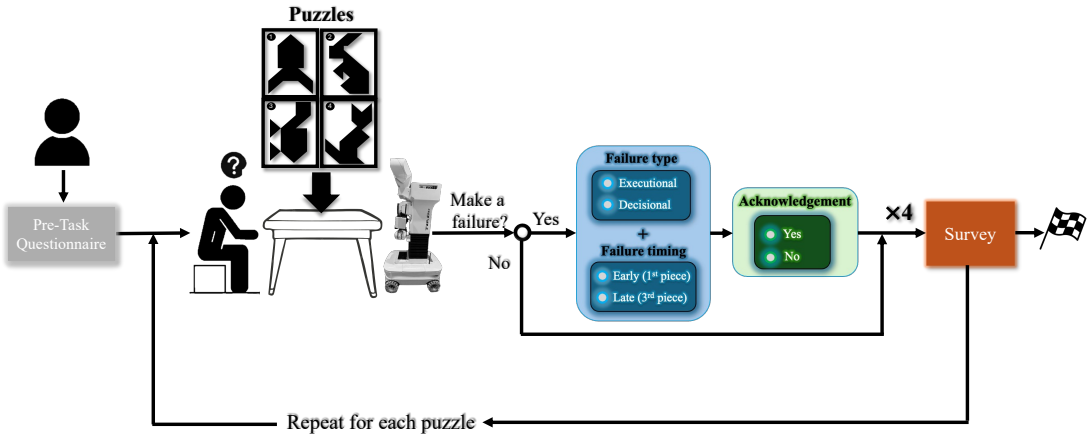

Fig. 2: Experimental diagram showing the process where participants first complete a pre-task questionnaire, followed by collaboratively solving a Tangram puzzle four times.

27 participants (16 females, 10 males, and 1 non-binary) via our university recruitment website. Participants primarily consisted of students and university staff, none of whom had previous experience working with robots. They were compensated with a gift voucher for their participation. The average age was ranging between 18y.o. and 34y.o. (M = 23.26, SD = 4.3). Due to some technical issue, one participant's gaze data was not recorded, and another participant did not complete some questions in one of the after-task questionnaires. Participants signed a consent form before participation and were made aware that their gaze data was recorded during the experiment. At the end of the experiment, participants were informed that the study aimed to observe their responses to the robot's failures.

D. Experiment

Participants were first briefed by an experimenter on how to interact with the robot and the goals of the tasks. They then completed a demographics questionnaire, providing their age and gender. Following this, participants were seated at a table opposite the robot and asked to wear eye-tracking glasses during the experiment. As per the experimental conditions (Table I), during each puzzle, the robot correctly picked and placed three pieces but intentionally made an error with one piece. After each puzzle, participants completed a questionnaire assessing their perception of the robot's performance during that specific puzzle. Participants were unaware that these errors were pre-programmed. This cycle was repeated for all four puzzles. The experimental procedure is illustrated in Figure 2.

The experimenter initiated the robot's turns and intervened when participants made mistakes by triggering the robot's verbal response. This ensured that the robot began its turn immediately after the participant's turn, maintaining a consistent time gap across all participants. The experimenter was seated on the opposite side of the table, near the robot, to ensure safety and to press the emergency button in case of an actual malfunction. For consistency, the same experimenter conducted all sessions and operated the robot throughout the study.

The experiment was conducted in a laboratory on the University of Melbourne campus. The duration of solving each puzzle together with the robot was about 191.12s ± 35.40s. After each puzzle, the experimenter asked the participant to complete a survey and prepared the table for the next puzzle. The gap between each puzzle was about 93.88s ± 38.81s.

A mixed experimental design was used, with failure types (executional and decisional) and failure timings (early and late) as within-subjects factors, and failure acknowledgement as a between-subjects factor. To minimise order effects, the within-subjects factors were counterbalanced using a four-condition balanced Latin Square. Each factor was systematically integrated into the puzzles. The first thirteen participants experienced the failure acknowledgement, while the second fourteen did not.

Participant ID	Puzzle 1	Puzzle 2	Puzzle 3	Puzzle 4	Acknowledgment
1	EF (Early)	EF (Late)	DF (Late)	DF (Early)	Yes
2	EF (Late)	DF (Early)	EF (Early)	DF (Late)	Yes
3	DF (Early)	DF (Late)	EF (Late)	EF (Early)	Yes
4	DF (Late)	EF (Early)	DF (Early)	EF (Late)	Yes
5	EF (Early)	EF (Late)	DF (Late)	DF (Early)	Yes
...	...	...	...	...	...
14	EF (Early)	EF (Late)	DF (Late)	DF (Early)	No
...	...	...	...	...	...

TABLE I: Order of failure type and timing across puzzles with acknowledgement of failure

E. Measures

1) Objective Gaze Measures: For each puzzle and each piece, we recorded the robot's current action— such as moving above the target object, and lowering to pick up the object—along with whether a failure occurred and the type of failure, all based on Unix time. We recorded users' gaze data during the whole experiment.

Gaze data was collected during the tasks as participants collaborated with the Tiago robot to solve the puzzles. In our experiment, the gaze data during the robot's turn was particularly important, from the moment it started moving until it completed its turn. Data was captured using Neon Eye Tracking Glasses from Pupil Labs. The gaze data included the participant's field of view image frame along with the x and y coordinates of their gaze within that frame. This data was recorded in real-time on a computer. The gaze data was captured at a rate of 30 Hz for both the image frames and gaze coordinates.

To facilitate the identification of participants' areas of interest (AoIs), we attached ArUco markers near the areas of interest.

The AoIs in our experiment included the robot body (comprising the robot's face and torso), the Tangram figure, the end effector, the robot's pieces, the participant's pieces, and the experimenter. These areas of interest are illustrated in Figure 1.

We calculated several gaze-related measures to analyse user behaviour during the interaction. These metrics included: (1) the number of gaze shifts toward the robot body, (2) the number of gaze shifts across all AoIs, (3) the proportional distribution of gaze directed toward the robot body, the Tangram figure, and the robot's end effector, and (4) transition and stationary entropy derived from gaze transition matrices [50], [51]. Each of these measures captures different aspects of gaze behaviour. The number of gaze shifts reflects the frequency of visual transitions between specific areas, providing insight into user engagement and focus dynamics. The proportional distribution of gaze indicates how much time users spent looking at each AoI, offering a measure of relative visual attention. Transition entropy quantifies the unpredictability of gaze transitions between AoIs, while stationary entropy measures the overall distribution of gaze within the AoIs, highlighting how scattered or concentrated the gaze behaviour was during the task.

The gaze measures were calculated during a specific time window for both failure and non-failure conditions: from the moment the robot began moving to pick up an object until it placed the object and returned to its initial position. Since failure timing is not applicable in non-failure conditions, the analysis of these measures was conducted in two ways. First, we analysed the data by failure type (no failure, executional failure, decisional failure) and acknowledgement (yes vs. no). Second, we analysed it by failure type (executional failure, decisional failure), timing (early vs. late), and acknowledgement (yes vs. no).

2) Subjective Measures: After each puzzle, participants rated their perceptions of the robot's behaviour in terms of perceived intelligence, perceived safety, and performance trust. Perceived intelligence and safety were measured using items from the Godspeed questionnaire [52], while performance trust was assessed using items from the Multi-Dimensional Measure of Trust (MDMT) questionnaire [53].

To evaluate the level of intelligence participants attributed to the robot, we used three items from the Godspeed questionnaire: "Incompetent/Competent," "Irresponsible/Responsible," and "Foolish/Sensible." For perceived safety, we included one item from the Godspeed questionnaire: "Anxious/Relaxed." To assess performance trust across various robot failures, we utilised the "performance trust" dimension from the MDMT. This included two items from the Reliable subscale ("Reliable" and "Predictable") and two items from the Competent subscale ("Skilled" and "Capable").

The analysis of these measures was conducted based on failure type (executional failure, decisional failure), timing (early vs. late), and acknowledgement (yes vs. no).

IV. RESULTS

A. Behavioural Response

In this section, we address the first research question by analysing participants' gaze behaviour using the measures outlined in section III-E1. Our analysis focuses on how these metrics vary during failure situations. Additionally, we investigate the anticipatory capability of participants' gaze about the placement of objects.

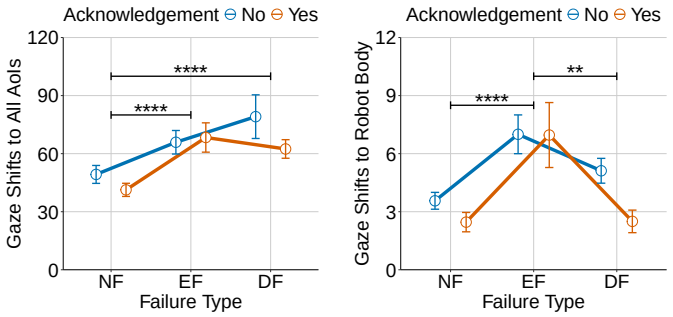

Fig. 3: The average number of gaze shifts across all AoIs (left) and toward the robot body (right) across three different failure situations, with or without the robot acknowledging its failure. Error bars represent the standard error of the mean. Significance levels, based on adjusted p-values, are denoted as follows: ** for $p < .01$, and **** for $p < .0001$.

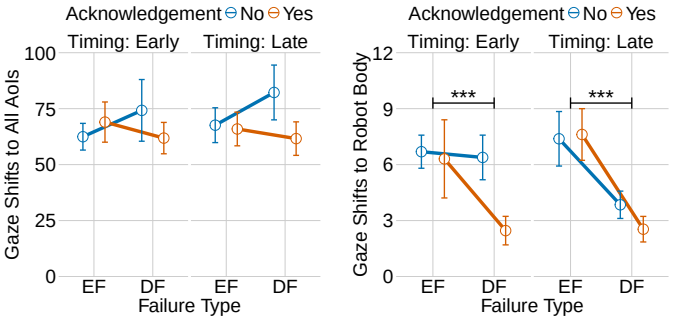

Fig. 4: The average number of gaze shifts across all AoIs and the average number of gaze shifts toward the robot body, comparing failure type, failure timing, and the robot's acknowledgement of its failure. Error bars represent the standard error of the mean. Significance levels are denoted as follows: *** for $p < .001$.

1) Gaze Shift: We conducted a two-way ANOVA to compare gaze patterns during failure versus non-failure robot actions, with having failure type as a within-subjects and the acknowledgement as a between-subjects. For the number of gaze shifts across all AoIs, the results indicated a significant main effect for the factor of failure type ($F(2,48) = 15.16$; $p < .001$; $\eta^2 = 0.39$). Bonferroni-corrected pairwise t-tests revealed significant differences between each type of failure and no failure (NF). For the number of gaze shifts toward the robot's body (i.e., the robot's face and torso), the results again showed significant main effects for the factor of failure type ($F(2,48) = 21.48$; $p < .001$; $\eta^2 = 0.47$). Bonferroni-corrected pairwise t-tests indicated significant differences between EF and DF, as well as between EF and NF. Figure 3 illustrates the average values for each condition.

Subsequently, a three-way ANOVA was conducted to analyse gaze patterns during failure durations, focusing on the effects of failure type and timing as within-subjects factors, and acknowledgement as a between-subjects factor. For the number of gaze shifts across all AoIs, the results showed no significant effects for any of the factors. However, for the number of gaze shifts toward the robot's body, the results revealed a significant effect of failure type ($F(1,24) = 17.780$; $p < .001$; $\eta^2 = 0.43$). Figure 4 shows the average values for each condition.

2) Gaze Distribution: In this section, we compare the proportion of gaze directed toward three AoIs: the robot's end effector, the Tangram figure, and the robot's body, during each task while the robot is performing its actions.

First, we compare the proportion of gaze directed during failure events to that during non-failure events. The results of the two-way ANOVA revealed significant differences in failure type but no differences in acknowledgement across all measures. Specifically, significant differences were observed for the end effector ($F(2,48) = 6.13$; $p = .009$; $\eta^2 = 0.20$), the Tangram figure ($F(2,48) = 17.71$; $p < .001$; $\eta^2 = 0.42$), and the robot's body ($F(2,48) = 14.35$; $p < .001$; $\eta^2 = 0.37$). Bonferroni-corrected pairwise tests indicated significant differences between EF and NF, as well as between EF and DF for all measures. Figure 5 shows the average values for each measure.

We subsequently conducted a three-way ANOVA with failure type and timing as within-subjects factors, and acknowledgement as a between-subjects factor. The results, as presented in Table II, demonstrated significant differences in failure type and timing across the end effector, Tangram figure, and Robot body. Notably, for the Robot body, we also observed significant interactions between acknowledgement and timing. Our analysis showed that participants looked at the Tangram figure more when the failure occurred early in the interaction compared to late failures, while they focused more on the robot's body and end effector during late failures than early ones.

3) Gaze Transition Matrix: Based on the AoIs, we created transition matrices, focusing exclusively on transitions between different AoIs while excluding self-repeating transitions. We then compared the transition matrices using transition entropy and stationary entropy.

We conducted a two-way ANOVA to compare transition matrices during failure versus non-failure robot actions. The results showed significant differences in failure type for both entropies. Specifically, significant differences were observed for transition entropy ($F(2,48) = 13.90$; $p < .001$; $\eta^2 = 0.37$), and stationary entropy ($F(2,48) = 11.01$; $p < .001$; $\eta^2 = 0.31$). No significant differences were found for acknowledgement. Further, Bonferroni-corrected pairwise t-tests indicated significant differences in transition entropy between EF and NF, and between NF and DF. For stationary entropy, significant differences were observed between EF and DF, and between NF and DF. The mean values for transition entropy indicate that NF has the highest value, while DF has the lowest. In contrast, for stationary entropy, EF has the highest value, and DF the lowest. In all conditions where the robot acknowledges its failure, both entropy values are lower. The transition matrices are shown in Figure 6.

Subsequently, we performed the Wilcoxon Signed-Rank Test for failure type and timing, and the Wilcoxon Rank-Sum Test for acknowledgement. The results (Table III) indicate significant differences in failure type for both entropies. Additionally, significant differences in timing were observed for stationary entropy. No significant differences were found for acknowledgement. The median values show that when the failure type is DF, the timing is late, or the robot acknowledges its failure, both entropy values are lower.

4) Goal Anticipation by Gaze Analysis: In this section, we examine the proportion of time participants spent looking at the correct goal location for object placement within the Tangram figure, compared to the total time spent looking inside the Tangram figure. The purpose of this analysis is to determine whether participants exhibited anticipatory gaze behaviour to assist the robot in recovering from its failures. Specifically, we assessed the average percentage of time participants looked at the goal, as well as the average number of gaze shifts towards the goal during each puzzle and failure period. The duration considered for each puzzle spanned from the moment the robot's end effector was positioned above the object it intended to pick up until it was positioned above the designated placement location for that object. During failure periods, we focused on the time from when the robot initiated a failure until it began its repair.

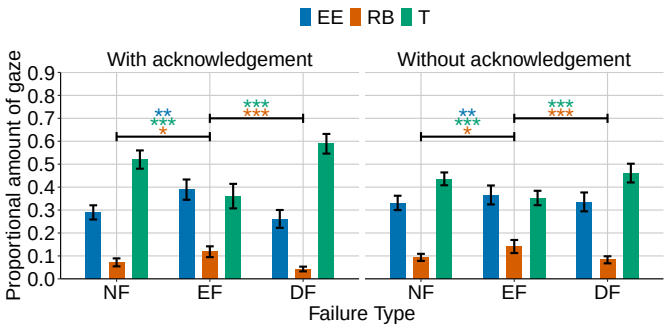

Fig. 5: The average proportion of participant gazes directed at the end effector (EE), robot body/face (RB), and Tangram figure (T) during puzzle solving, across three different failure situations, with or without the robot acknowledging its failure. Error bars represent the standard error of the mean. Significance levels, based on adjusted p-values, are denoted as follows: * for $p < .05$, ** for $p < .01$, and *** for $p < .001$.

The results indicate that as participants progressed through the puzzles, the proportion of time spent looking at the goal decreased. Specifically, in Puzzle 1, participants looked at the goal for 47% (±14%) of the time, followed by 44% (±13%) in Puzzle 2, 38% (±16%) in Puzzle 3, and 22% (±9%) in Puzzle 4. Additionally, the total number of gaze shifts towards the goal also almost decreased as participants advanced through the puzzles. The average number of gaze shifts per piece was 10.33 (±5.35) in Puzzle 1, 10.65 (±4.57) in Puzzle 2, 7.66 (±4.79) in Puzzle 3, and 5.47 (±3.18) in Puzzle 4.

When analysing the failure periods, the results show that participants spent 35% (±26%) of their task-related gaze time looking at the goal during EF. This percentage was higher, at 41% (±15%), during DF. Additionally, participants exhibited an average of 3.42 (±3.36) gaze shifts for each EF, compared to a substantially higher average of 14.84 (±7.11) gaze shifts for each DF.

B. Subjective Measures

In this section, we address the second research question by analysing participants' subjective behaviour using the measures outlined in section III-E2. To achieve this, we conducted a three-way ANOVA for each subjective scale to examine the effects of failure type, timing, and acknowledgement. For the Competent scale, significant interaction effects were found for type*timing ($F(1,24) = 6.79$, $p = .016$, $\eta^2 = 0.22$). The Sensible scale showed a significant main effect of timing ($F(1,24) = 5.79$, $p = .024$, $\eta^2 = 0.19$). In the Anxious/Relaxed (Self) scale, significant main effects were observed for timing

Scale	Measure	Type	Timing	Acknowledgement	[Type*Timing]	[Type*Acknowledgement]	[Timing*Acknowledgement]	[Type*Timing*Acknowledgement]
End Effector	df	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)
	F value	7.79	4.49	0.31	<0.001	3.85	0.55	0.75
	p value	**.010**	**.447**	.586	.980	.061	.465	.396
	η^2	0.24	0.16	0.01	<0.0001	0.14	0.02	0.03
Tangram figure	df	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)
	F value	25.86	4.85	2.02	0.03	4.09	0.80	2.09
	p value	**<.001**	**.038**	.168	.868	.054	.379	.161
	η^2	0.52	0.17	0.08	<0.01	0.15	0.03	0.08
Robot Body	df	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)
	F value	23.01	5.29	1.20	2.80	0.46	5.71	1.34
	p value	**<.001**	**.030**	.284	.108	.502	**.025**	.258
	η^2	0.49	0.18	0.05	0.10	0.02	0.19	0.05

TABLE II: Results of the three-way mixed ANOVA for gaze distribution

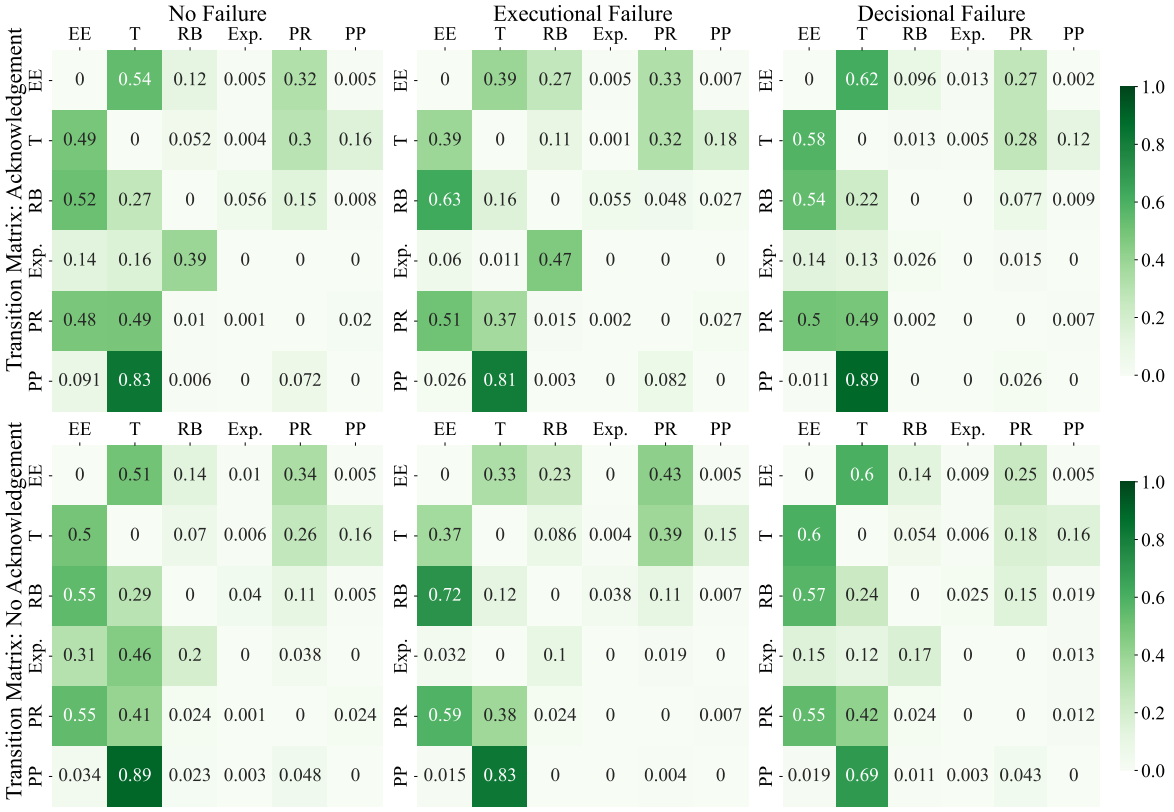

Fig. 6: The transition matrices for three different interaction scenarios—NF, EF, and DF—are presented, both for cases where the robot acknowledges its failure and where it does not. The vertical axis represents the current states, while the horizontal axis represents the next states. 'EE' stands for End Effector, 'T' for Tangram figure, 'RB' for Robot Body, 'Exp.' for Experimenter, 'PR' for Pieces (Robot), and 'PP' for Pieces (Participant). The transition matrices are displayed as heat maps.

Scale	Measure	Type	Timing	Acknowledgement
Transition Entropy	N	52	52	52
	W	1007	864	1583
	p-value	**.004**	.112	.134
Stationary Entropy	N	52	52	52
	W	1197	1035	1522
	p-value	**<.001**	**.002**	.270

TABLE III: Results of the Wilcoxon tests for the entropy of the transition matrices, where N represents the sample size for each condition and W is the test statistic.

($F(1,24)= 7.80$, $p = .010$, $\eta^2 = 0.24$) and acknowledgement ($F(1,24)= 5.50$, $p = .027$, $\eta^2 = 0.18$). The Predictable scale had a significant interaction effect for type*acknowledgement ($F(1,25)=5.38$, $p=.029$, $\eta^2=0.18$). The Skilled scale showed a significant main effect of type ($F(1,25)= 4.98$, $p = .035$, $\eta^2 = 0.17$). Finally, the Capable scale revealed a significant three-way interaction of type*timing*acknowledgement ($F(1,25)=6.99$, $p=.014$, $\eta^2=0.22$). The results showed that participants rated the robot higher on measures of perceived intelligence and trust in the questionnaire when the failure was executional, occurred early, or when the robot acknowledged its failure. However, for feelings of safety, ratings were higher when the failure occurred late and the robot did not acknowledge it. More information can be found in the Appendix.

V. DISCUSSION

This study compared behavioural responses to robot failures, focusing on how individuals reacted and perceived the robot. Failures varied by type, timing, and acknowledgement. The findings revealed that robot failures affect user gaze and perceptions. These findings are discussed further in the following section.

A. Behavioural Response

To address the first research question, we analysed user gaze behaviour in multiple ways: the number of gaze shifts, gaze distribution during puzzle-solving, and gaze entropy based on

transition matrices. These measures allowed us to examine how the type and timing of failures, as well as whether the robot acknowledged its failure, influenced user gaze patterns and whether gaze behaviour varied across different failure scenarios. Our results showed that user gaze is a reliable indicator of robot failures. When the robot made a failure, participants exhibited more frequent gaze shifts between different AoIs, likely due to confusion and an attempt to understand what was happening. This finding is similar to the results of Kontogiorgos et al. [29], who found that people tend to gaze more at the robot when it makes a mistake. The literature suggests that different types of failures influence user perceptions of the robot [38], and our findings support this by showing that users exhibit distinct gaze behaviours in response to various failure types. For example, when the failure was executional, the number of gaze shifts towards the robot was significantly higher compared to when the failure was decisional. Moreover, during executional failures, the proportion of time spent looking at the robot was much higher compared to decisional failures. It is crucial for the robot to recognize the type of failure it has made so that it can determine the appropriate strategy for recovery and regain the user's trust.

The timing of the failure is also crucial for the robot, as it requires different approaches for recovery and repair. In our research, while the timing of the failure—whether at the start or end of the interaction—did not significantly affect gaze shifts, it did influence gaze transition matrices, and gaze distribution across AoIs. Failures at the beginning of the interaction led to higher median gaze transition values, indicating more randomness early on. Additionally, participants' focus on the Tangram figure was more when the failure occurred at the beginning of the interaction compared to later ones, while their focus on the robot's body or end effector was more during late failures than early ones.

In our research, after committing a failure, the robot could either acknowledge the failure and then continue its action, or proceed without acknowledgement. We could not find significant differences in users' gaze behaviour when the robot acknowledged its failure and when it did not. As the literature suggests [43], [44], [54], there are other verbal approaches to failure recovery, such as promises and technical explanations, which might influence users' gaze differently. Verbal failure recovery is important for robots, as it demonstrates an awareness of mistakes. This, in turn, can make the robot appear more intelligent and encourage users to engage with it more.

Our study also explored changes in users' anticipatory gaze behaviour during the task and its potential role in assisting the robot to recover from failures. Participants frequently anticipated the placement of the object before the robot executed the action, even when the robot made an error. This anticipatory gaze behaviour could serve as a valuable cue for the robot to detect its failures and initiate appropriate recovery strategies. However, we observed a decrease in participants' anticipatory gaze behaviour as the number of tasks increased. This decline may indicate reduced engagement over time, with participants being more actively collaborative at the beginning of the interaction. It also suggests that users' gaze behaviour might change throughout the interaction. These findings highlight the dynamic nature of gaze behaviour throughout the interaction.

B. Subjective Measures

To address the second research question, we examined user perceptions of the robot in three areas: perceived intelligence, sense of safety, and trust during failures. The analysis revealed how these measures varied with the type and timing of failure and whether the robot acknowledged its mistake.

The results of the subjective evaluation revealed that users' perceptions of the robot's intelligence and safety were not significantly influenced by the type of failure. However, users exhibited higher levels of trust in the robot during executional failures compared to decisional failures, suggesting that placing an object in an incorrect location reduces trust more than making an incorrect decision. Additionally, we observed interesting findings regarding the timing of the robot's failures. When failures occurred early in the interaction, users rated the robot as more intelligent and trustworthy compared to failures that occurred later. For the measure of "Sensible," this difference was statistically significant. These findings are consistent with previous research by Morales et al. [38] and Lucas et al. [41]. Interestingly, users reported feeling more relaxed when failures occurred later in the interaction, aligning with results from Desai et al. [39] and Rossi et al. [40].

When the robot acknowledged its failures, users perceived it as slightly more intelligent and trustworthy but also experienced increased anxiety. This finding may be explained by the robot's consistent physical repair actions a few seconds after each failure. When the robot did not explicitly acknowledge its failures, users might not have interpreted these actions as errors, reducing their perception of failure events.

C. Limitations and Future Work

There were instances where participants were preoccupied with determining the placement of their next piece, which occasionally led them to overlook the robot's movements. However, these occurrences were minimal. Another limitation is the restriction to only two types of failure and whether the robot acknowledges its failure or not. The effect size in our study was medium; however, to obtain more robust results, a larger sample size would be beneficial. Furthermore, for safety reasons, the robot's arm movement was slowed and the experimenter was in the room, which may have influenced participants' perceptions. Future research could address these limitations by exploring a broader range of failure types and incorporating explanatory feedback from the robot.

VI. CONCLUSION

This study examines how robotic failures affect human gaze dynamics and perceptions during collaborative tasks, offering insights into using gaze as a failure indicator to assist in repair. The findings reveal that executional failures lead to more gaze shifts toward the robot, indicating user confusion, while decisional failures result in lower entropy in gaze transitions among areas of interest. Failures at the beginning of the interaction lead to more randomness in gaze shifts across AoIs. The timing of the failure during the task also affects users' gaze distribution across AoIs. Finally, acknowledgement of failure does not seem to affect gaze behaviour or users' perception. Our work contributes to a better understanding of how gaze behaviour can be leveraged in HRC to design more effective and reliable human-robot interaction systems.

References

[1] A. Sauppé and B. Mutlu, "The Social Impact of a Robot Co-Worker in Industrial Settings," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, (New York, NY, USA), pp. 3613–3622, Association for Computing Machinery, Apr. 2015.

[2] Y. Terzioğlu, B. Mutlu, and E. Şahin, "Designing Social Cues for Collaborative Robots: The Role of Gaze and Breathing in Human-Robot Collaboration," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, (New York, NY, USA), pp. 343–357, Association for Computing Machinery, Mar. 2020.

[3] F. Babel, A. Vogt, P. Hock, J. Kraus, F. Angerer, T. Seufert, and M. Baumann, "Step Aside! VR-Based Evaluation of Adaptive Robot Conflict Resolution Strategies for Domestic Service Robots," *International Journal of Social Robotics*, vol. 14, pp. 1239–1260, July 2022.

[4] E. Schneiders, A. M. Kanstrup, J. Kjeldskov, and M. B. Skov, "Domestic Robots and the Dream of Automation: Understanding Human Interaction and Intervention," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, (New York, NY, USA), pp. 1–13, Association for Computing Machinery, May 2021.

[5] S. Chatterjee, R. Chaudhuri, and D. Vrontis, "Usage Intention of Social Robots for Domestic Purpose: From Security, Privacy, and Legal Perspectives," *Information Systems Frontiers*, vol. 26, pp. 121–136, Feb. 2024.

[6] L. Mingyue Ma, T. Fong, M. J. Micire, Y. K. Kim, and K. Feigh, "Human-Robot Teaming: Concepts and Components for Design," in *Field and Service Robotics* (M. Hutter and R. Siegwart, eds.), (Cham), pp. 649–663, Springer International Publishing, 2018.

[7] T. Chakraborti, S. Kambhampati, M. Scheutz, and Y. Zhang, "AI Challenges in Human-Robot Cognitive Teaming," Aug. 2017. arXiv:1707.04775 [cs].

[8] S. Honig and T. Oron-Gilad, "Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development," *Frontiers in Psychology*, vol. 9, 2018.

[9] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of changing reliability on trust of robot systems," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, HRI '12, (New York, NY, USA), pp. 73–80, Association for Computing Machinery, Mar. 2012.

[10] K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock, "A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 58, pp. 377–400, May 2016.

[11] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, (New York, NY, USA), pp. 141–148, Association for Computing Machinery, Mar. 2015.

[12] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, ""I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 57–65, Mar. 2019. ISSN: 2167-2148.

[13] X. Lei and P.-L. P. Rau, "Should I Blame the Human or the Robot? Attribution Within a Human–Robot Group," *International Journal of Social Robotics*, vol. 13, pp. 363–377, Apr. 2021.

[14] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social Robotics," in *Springer Handbook of Robotics* (B. Siciliano and O. Khatib, eds.), pp. 1935–1972, Cham: Springer International Publishing, 2016.

[15] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "A matter of consequences: Understanding the effects of robot errors on people's trust in HRI," *Interaction Studies*, vol. 24, pp. 380–421, Dec. 2023. Publisher: John Benjamins.

[16] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 83–90, Mar. 2016. ISSN: 2167-2148.

[17] S. Rossi, F. Ferland, and A. Tapus, "User profiling and behavioral adaptation for HRI: A survey," *Pattern Recognition Letters*, vol. 99, pp. 3–12, Nov. 2017.

[18] N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine Theory of Mind," Mar. 2018. arXiv:1802.07740 [cs].

[19] L. Wachowiak, P. Tisnikar, G. Canal, A. Coles, M. Leonetti, and O. Celiktutan, "Analysing Eye Gaze Patterns during Confusion and Errors in Human–Agent Collaborations," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, (Napoli, Italy), pp. 224–229, IEEE, Aug. 2022.

[20] B. M. Velichkovsky, A. Kotov, N. Arinkin, L. Zaidelman, A. Zinina, and K. Kivva, "From Social Gaze to Indirect Speech Constructions: How to Induce the Impression That Your Companion Robot Is a Conscious Creature," *Applied Sciences*, vol. 11, p. 10255, Jan. 2021. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.

[21] Y. Fang, J. Tang, W. Shen, W. Shen, X. Gu, L. Song, and G. Zhai, "Dual Attention Guided Gaze Target Detection in the Wild," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11385–11394, June 2021. ISSN: 2575-7075.

[22] E. Rubies, J. Palacín, and E. Clotet, "Enhancing the Sense of Attention from an Assistance Mobile Robot by Improving Eye-Gaze Contact from Its Iconic Face Displayed on a Flat Screen," *Sensors (Basel, Switzerland)*, vol. 22, p. 4282, June 2022.

[23] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration," *Frontiers in Psychology*, vol. 6, July 2015. Publisher: Frontiers.

[24] E. Mwangi, E. I. Barakova, M. Díaz, A. C. Mallofré, and M. Rauterberg, "Dyadic Gaze Patterns During Child-Robot Collaborative Gameplay in a Tutoring Interaction," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 856–861, Aug. 2018. ISSN: 1944-9437.

[25] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan, "Eye-hand coordination in object manipulation," *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, vol. 21, pp. 6917–6932, Sept. 2001.

[26] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?," *Vision Research*, vol. 41, pp. 3559–3565, Nov. 2001.

[27] J. S. Matthis, J. L. Yates, and M. M. Hayhoe, "Gaze and the Control of Foot Placement When Walking in Natural Terrain," *Current Biology*, vol. 28, pp. 1224–1233.e5, Apr. 2018.

[28] M. Stiber, R. H. Taylor, and C.-M. Huang, "On Using Social Signals to Enable Flexible Error-Aware HRI," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, (New York, NY, USA), pp. 222–230, Association for Computing Machinery, Mar. 2023.

[29] D. Kontogiorgos, S. van Waveren, O. Wallberg, A. Pereira, I. Leite, and J. Gustafson, "Embodiment Effects in Interactions with Failing Robots," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), pp. 1–14, Association for Computing Machinery, Apr. 2020.

[30] C. E. Peacock, B. Lafreniere, T. Zhang, S. Santosa, H. Benko, and T. R. Jonker, "Gaze as an Indicator of Input Recognition Errors," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 142:1–142:18, May 2022.

[31] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani, "A Systematic Cross-Corpus Analysis of Human Reactions to Robot Conversational Failures," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, (New York, NY, USA), pp. 112–120, Association for Computing Machinery, Oct. 2021.

[32] N. Mirnig, M. Giuliani, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Impact of Robot Actions on Social Signals and Reaction Times in HRI Error Situations," in *Social Robotics* (A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, eds.), (Cham), pp. 461–471, Springer International Publishing, 2015.

[33] M. Stiber, R. Taylor, and C.-M. Huang, "Modeling Human Response to Robot Errors for Timely Error Detection," July 2022. arXiv:2208.00565 [cs].

[34] L. Wachowiak, P. Tisnikar, A. Coles, G. Canal, and O. Celiktutan, "A Time Series Classification Pipeline for Detecting Interaction Ruptures in HRI Based on User Reactions," in *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24)*, ACM, Aug. 2024.

[35] P. Trung, M. Giuliani, M. Miksch, G. Stollnberger, S. Stadler, N. Mirnig, and M. Tscheligi, "Head and shoulders: automatic error detection in human-robot interaction," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, (New York, NY, USA), pp. 181–188, Association for Computing Machinery, Nov. 2017.

[36] R. M. Aronson and H. Admoni, "Gaze for error detection during human-robot shared manipulation," in *Fundamentals of Joint Action workshop, Robotics: Science and Systems*, p. 5, 2018.

[37] L. Tian and S. Oviatt, "A Taxonomy of Social Errors in Human-Robot Interaction," *J. Hum.-Robot Interact.*, vol. 10, pp. 13:1–13:32, Feb. 2021.

[38] C. G. Morales, E. J. Carter, X. Z. Tan, and A. Steinfeld, "Interaction Needs and Opportunities for Failing Robots," in *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, (New York, NY, USA), pp. 659–670, Association for Computing Machinery, June 2019.

[39] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 251–258, Mar. 2013. ISSN: 2167-2148.

[40] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario," in *Social Robotics* (A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, eds.), (Cham), pp. 42–52, Springer International Publishing, 2017.

[41] G. M. Lucas, J. Boberg, D. Traum, R. Artstein, J. Gratch, A. Gainer, E. Johnson, A. Leuski, and M. Nakano, "Getting to Know Each Other: The Role of Social Dialogue in Recovery from Errors in Social Robots," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, (New York, NY, USA), pp. 344–351, Association for Computing Machinery, Feb. 2018.

[42] G. LeMasurier, A. Gautam, Z. Han, J. W. Crandall, and H. A. Yanco, "Reactive or Proactive? How Robots Should Explain Failures," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, (New York, NY, USA), pp. 413–422, Association for Computing Machinery, Mar. 2024.

[43] C. Esterwood and L. P. Robert, "Do You Still Trust Me? Human-Robot Trust Repair Strategies," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 183–188, Aug. 2021. ISSN: 1944-9437.

[44] L. Wachowiak, A. Fenn, H. Kamran, A. Coles, O. Celiktutan, and G. Canal, "When Do People Want an Explanation from a Robot?," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, (New York, NY, USA), pp. 752–761, Association for Computing Machinery, Mar. 2024.

[45] V. Srinivasan and R. Murphy, "A survey of social gaze," in *Proceedings of the 6th international conference on Human-robot interaction*, HRI '11, (New York, NY, USA), pp. 253–254, Association for Computing Machinery, Mar. 2011.

[46] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *J. Hum.-Robot Interact.*, vol. 6, pp. 25–63, May 2017.

[47] L. Shi, C. Copot, and S. Vanlanduit, "GazeEMD: Detecting Visual Intention in Gaze-Based Human-Robot Interaction," *Robotics*, vol. 10, p. 68, June 2021. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[48] A. Inceoglu, E. E. Aksoy, A. C. Ak, and S. Sariel, "FINO-Net: A Deep Multimodal Sensor Fusion Framework for Manipulation Failure Detection," July 2021. arXiv:2011.05817 [cs].

[49] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, vol. 39, pp. 175–191, May 2007.

[50] K. Krejtz, A. Duchowski, T. Szmidt, I. Krejtz, F. González Perilli, A. Pires, A. Vilaro, and N. Villalobos, "Gaze Transition Entropy," *ACM Trans. Appl. Percept.*, vol. 13, pp. 4:1–4:20, Dec. 2015.

[51] I. A. Ebeid, N. Bhattacharya, J. Gwizdka, and A. Sarkar, "Analyzing gaze transition behavior using bayesian mixed effects Markov models," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, (New York, NY, USA), pp. 1–5, Association for Computing Machinery, June 2019.

[52] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009. Place: Germany Publisher: Springer.

[53] D. Ullman and B. F. Malle, "MDMT: Multi-Dimensional Measure of Trust v2," 2023.

[54] 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*, HRI '23, (New York, NY, USA), pp. 271–280, Association for Computing Machinery, Mar. 2023.

APPENDIX

Results of the three-way mixed ANOVA for subjective measures

Scale	Measure	Type	Timing	Acknowledgement	[Type*Timing]	[Type*Acknowledgement]	[Timing*Acknowledgement]	[Type*Timing*Acknowledgement]
Competent	df	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)
	F value	1.51	0.03	0.70	6.79	2.49	0.03	2.75
	p value	.231	.855	.413	**.016**	.127	.855	.110
	η^2	0.06	<0.01	0.03	0.22	0.09	<0.01	0.10
Sensible	df	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)
	F value	2.63	5.79	0.53	0.44	2.62	1.79	2.41
	p value	.118	**.024**	.475	.512	.118	.194	.134
	η^2	0.10	0.19	0.02	0.02	0.10	0.07	0.09
Responsible	df	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)
	F value	0.58	0.04	1.74	0.03	0.21	0.34	0.29
	p value	.453	.848	.200	.859	.651	.567	.595
	η^2	0.02	<0.01	0.07	<0.01	<0.01	0.01	0.01
Anxious/Relaxed (Self)	df	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)	(1,24)
	F value	0.07	7.80	2.23	0.26	5.50	0.84	0.01
	p value	.792	**.010**	.148	.613	**.027**	.369	.905
	η^2	<0.01	0.24	0.08	0.01	0.18	0.03	<0.001
Reliable	df	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)
	F value	1.61	0.83	1.28	1.50	1.00	0.35	0.99
	p value	.216	.370	.268	.233	.328	.561	.330
	η^2	0.06	0.03	0.05	0.06	0.04	0.01	0.04
Predictable	df	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)
	F value	2.98	1.43	0.94	3.97	5.38	0.26	3.97
	p value	.097	.243	.340	.057	**.029**	.616	.057
	η^2	0.11	0.05	0.04	0.14	0.18	0.01	0.14
Skilled	df	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)
	F value	4.98	2.93	0.43	0.13	1.65	0.11	3.46
	p value	**.035**	.099	.516	.719	.210	.741	.075
	η^2	0.17	0.11	0.02	<0.01	0.06	<0.01	0.12
Capable	df	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)	(1,25)
	F value	1.71	3.02	0.80	1.15	0.002	0.41	6.99
	p value	.203	.095	.380	.293	.962	.528	**.014**
	η^2	0.06	0.11	0.03	0.04	<0.001	0.02	0.22