The Influence of Robot Role Priority on Blame and Trust Attribution in Multi-Robot Coordination Failures

Priyesh Vashistha
Student Number: 2049905
Queen's University
Kingston, Canada

***Abstract*—As autonomous robots start to enter complex environments such as hospitals, they at one time or another will encounter some form of coordination failures. What this study investigates is that how a robots designated social role, specifically its priority has an influence on a human's attribution of blame and trust during such failures. Using a in between-subjects design, participants viewed animated videos of a navigation failure between 2 robots, 1 being a high priority "medicine delivery" robot and other being a low priority 'trash disposal' robot. The results suggest that high priority "medicine delivery" robot has suffered a high decline in perceived competence when it failed, consistent with the Expectancy-Disconfirmation Theory. On the other hand, attribution of blame was seen following a social Hierarchy pattern, where the low priority "trash disposal" robot was blamed more for the failure condition. These findings show that a robot's functional label serves as some kind of psychological frame, which significantly alters how a interpret and rate technical errors. This research provides some critical insight for designing resilient multi-robot systems that try to maintain user trust despite inevitable performance gaps.**

I. Introduction

In the near future as autonomous robots become more integrated into complex human-centric environments such as hospitals, their ability to seamlessly operate within heterogeneous team is becoming very crucial. In the future, hospital environments will not be dependent on a single type of robot, but rather a diverse ecosystem of robot that does different work and will have different priorities and social roles, that will range from critical emergency robots to regular routine maintenance robots. In a high-stake environment coordinating seamlessly is very important, however failures in robots will always be an inevitable reality. While technical robustness is should be the primary engineering goal, the human observer's perception of these failure conditions is equally vital for successfully adopting these robotic technologies.

Trust is one of the foundational elements in Human-Robot Interaction (HRI). But trust is not a monolithic concept as it seems Research on the Multi-Dimensional Measure of Trust (MDMT) argue that trust in robotics agents must be divided into distinct dimensions which are, performance trust (the belief that the agent can successfully complete a task) and Moral trust (the belief that the agent cats with integrity and honesty). Whenever a failure occurs the observer do not just register an error, they make complex social judgements about the agents involved in a similar way they do for humans. The Robotic Social Attributes Scale (ROSAS). refines this further by applying the stereotype content model to social robotics, which suggests that the social prception is primarly driven by judging warmth and competence. All these dimensions trust, warmth and competence – form a psychological framework by which humans evaluate the behaviour of robots.

Existing research in HRI has largely studied how errors commected by a single robot affect human trust in dyadic (one to one) interactions. However, there was a significant gap in literature regarding how observers assign blame in situations where there are multi agent system failures, perticurlarly when the robots involved have different social priorities and functional goals. It remains unclear that how a neutral observer attributes responsibility when a high-priority and a low-priority robot encounter an coordination failure.

The goal of this study is to bride this gap by studying how the designated role of a robot, specifically its social priority (e.g. emergency medicine delivery vs trash disposal) influences an observer's attribution of blame, perceived competence and trust in an navigation failure situation. By understanding these dynamics, we can provide some useful insights to engineers and designers helping them to create robotic systems that could be perceived as resilient and trustworthy, even in the reality of inevitable system errors.

II. BACKGROUND

A. SINGLE ROBOT FLIURE AND TRUST

The impact that robotic failures have on human trust have been extensively documented in context of dyadic (one-human, one-robot) interactions. Ye et al. [1] demonstrated that the gravity of a trust violation often depends upon the timing and framing of the error, noticing that users are particularly more sensitive to the mistake that violates the already established competence expectations. Similarly, Desai et al. [4] found that early interaction errors have a more negative and lasting impact on acquisition of trust, which suggest that the initial impressions about reliability are fragile. Salem et al. [5] then further expanded on this by exploring different types oof errors, finding that technical failure errors (e.g. navigation faliures) damage perceived competence, they do not always degrade social engagement if the robot shows some compensatory social behavior. All of this together says that humans have a lens of performance expectations for judging robotic failures. Yet these studies remain limited to single-agent scenarios.

B. ATTRIBUTION OF BLAME IN HRI

Whenever a failure occurs, the human observer always seeks to attribute blame to someone. In HRI the attribution theory suggests that the users often engage in search for a causal agent. Furlough et al. [2] found that when a robots autonomy is increased humans tend to attribute more blame to the robot itself rather then to any external factors or the designers. But this attribution process is complex. Kim and Hinds [3] showed that if the internal state of the robot is more transparent to the observer it can mitigate blame, as it will help the user to understand the "why" behind the error. Furthermore Van der Hoorn et al. [6] highlighted that an interactions social dynamics play a role when, a robot actively blames a human partner, it changes the users perception of the robots agency and personality.

C. THE RESEARCH GAP: MULTI ROBOT SYSTEMS

despite this rich literature there is a visible lack of research addressing failures that occure in multi-robot systems. Most of the existing studies have focused on dynamic interactions, which leaves a gap in our understanding of group HRI environments where multiple autonomous robots are ment to coordinate. In an hospital environment robots with different functions possesses inherent differences in social priority. Currently it is unknown if "social hierarchy" effects, as observed in human teams where subordinate members are often blamed for coordination failures can transfer to robotic teams. The aim of this study is to address this gap by isolating the variable of "robot role priority" to observe its effect on blame attribution during a system failure.

D. RESEARCH HYPOTHESES

Based on the Expectancy-Disconfirmation Theory and the Social Hierarchy bias discussed above, this study posits three specific hypotheses:

- **H1 (The Competence Cliff):** The High-Priority "Medicine Delivery Robot" will experience a significantly larger decline in perceived competence ratings following a failure compared to the Low-Priority "Trash Disposal Robot," due to the violation of higher initial performance expectations.
- **H2 (Hierarchical Blame):** Participants will attribute significantly more responsibility for the coordination failure to the Low-Priority Robot, even when the physical navigation error is ambiguous, reflecting a social bias where lower-status agents are expected to yield.
- **H3 (Role Rationalization):** Qualitative justifications for blame will frequently cite the "importance" or "urgency" of the Medicine Robot as a mitigating factor, whereas the Trash Robot will be blamed due to its perceived lack of consequence.

III. METHODOLOGY

A. STUDY DESIGN AND PARTICIPANTS

This study had a between-subjects experimental design, and the independent variable was the interaction outcome, manipulated between 2 conditions, (1st) High-priority Robot failure (2nd) Low priority Robot failure. 128 total participants were recruited from the Queen's university community (target N=128). The inclusion criteria was that participants should be at least 18 years of age and possess an English comprehension level of grade 6 or higher. All the participants were volunteers and did not receive any financial compensation.

B. MATERIALS AND STIMULI

The stimuli for experiment consisted of two 3D animated short videos, Both the videos showed a hospital hallway scenario featuring 2 autonomous robots: a High priority

(a) Medicine Delivery Robot

(b) Trash Disposal Robot

Fig. 1: Experimental Stimuli. The visual design of the high-priority "Medicine Delivery Robot" (a) and the low-priority "Trash Disposal Robot" (b) as they appear in the animated hospital scenarios. Note: Images are enlarged for clarity.

"Medicine delivery robot" and a low priority "garbage disposal robot". The robots were visually distinct to reflect their functions. Condition-1 (Low priority failure): both robots meet in an corridor the low priority robot fails to yield causing the failure condition. Condition-2(high priority robot): both robots meet in an corridor the high priority robot fails to yield causing failure condition. Video-based prototypes were used for this study rather than live physical robots for this study. This approach is supported by previous HRI research who showed that participants judge robot behavioures and social attributes similarly in video-based study compared to live interactions, validating the use of video video stimuli for initial perception studies.

C. PROCEDURE

a private room on the University campus was used to conduct this study. After arriving there participants provided written informed consent. Then they were seated at a computer to complete a web-based survey via Qualtrics.

- Briefing: after starting the survey participants read a briefing describing the roles of both the robots to get the idea of priority hierarchy. A comprehension check was done to be sure that the participants understand that "medicine delivery robot" had a higher priority than the trash robot.

- Intervention: the participants were randomly assigined to watch one of the 2 failure scenarios.
- Measurement: immediately after the video, participants completed the perception questionnaires.
- Debriefing: this session concluded with debriefing participants about the purpose of study.

D. MEASURES

To assess observers perception, 3 key measures were used.

- Trust: the (MDMT) Multi-Dimensional Measure of Trust was used to asswss dimensions of Reliability and ethical trust.
- Social Perception: the (RoSAS) Robotic Social Attributes Scale was used to measure perceived Competence, Warmth, and Discomfort.
- Attribution of Responsibility: A single open- ended question was used to ask participants to describe "who or what was most responsible for the outcome," whose answers were then analysed for blame attribution patterns.

E. ETHICAL CONSIDERATIONS

This study protocol has been reviewed and approved by the General Research Ethics Board (GREB) at Queen's University. Prior to participating, all individuals were provided with a digital Letter of Information detailing the study's purpose, the voluntary nature of their participation, and their right to withdraw at any time without penalty.

Since the study involves observing robot failures which may cause mild psychological discomfort, a comprehensive debriefing session is included at the end of the survey. This debriefing clarifies that the robot failures were scripted animations designed to test social perceptions and do not reflect the actual reliability of hospital delivery systems. All participant data is anonymized and stored on secure, encrypted servers in compliance with university data protection policies.

IV. RESULTS(EXPECTED)

As this was a "ready to run" study, the following results are projected based on comparable studies in HRI literature that utilized similar experimental design and measurements.

A. Data Analysis Plan

To evaluate the hypotheses, quantitative date collected from the MDMT and ROSAS scales will be analysed using independent samples t-tests to compare the mean scores between the High-Priority failure and Low-Priority failure conditions. A result will be considered statistically significant if the p-value is less than .05. For the qualitative component, the open-ended responses regarding responsibility will undergo thematic analysis. Responses will be coded identify to recurring themes in blame attribution (e.g., "Robot Role," "System Error," or "Navigation Constraints") to determine if participants rationalize blame based on the robots ' designated priorities.

B. competence and expectation violation

It was anticipated that the high- priority "medicine delivery robot" will receive significantly lower competence ratings following failures compare to the low priority "trash disposal robot". This projection was derived from Ye et al.[1] who demonstrated that the framing of a robot's interaction significantly affects how errors are perceived. In their study, violations of established competence expectations led to a steeper decline in user trust when compared to other conditions. This study expects to replicate similar results, where RoSAS competence scores for the "medicine delivery robot" will show a significant decrease compared to "trash disposal robot" reflecting on the penalty for violating the high-performance baseline associated with its role.

C. BLAME ATTRIBUTION (Quantitative)

According to the attribution of responsibility, it was predicted that a significant effect of robot role is that more blame will be assigned to low-priority robot. This exception is based on findings by Furlough et al.,[2] who investigated how different robot characteristics influence blame attribution in mixed teams. Their work suggest that observers consistently attribute responsibility based on perceived autonomy and role expectations during collaborative failures. Reflecting on these findings its anticipated that the quantitative data will show a skew where a majority of participants will identify the trash robot as responsible for failure.

D. QULATATIVE JUSTIFICATION OF BLAME

It is anticipated that after the analysis of the open-ended question "role Priority" will be revelled as the main justifica-

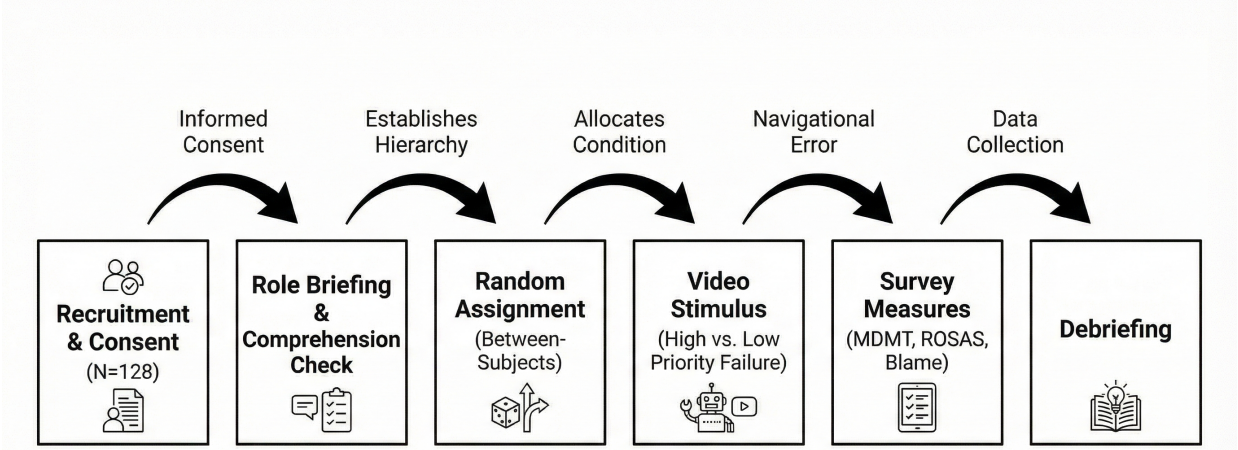

Fig. 2. Detailed Experimental Procedure.

Fig. 2: Experimental Procedure. The step-by-step protocol for the study, illustrating the flow from recruitment to data collection. This timeline ensures all ethical guidelines were met before data collection.

tion theme. Based on Kim and Hinds [3] who found that users attribute blame according to their understanding of a robots constrains and autonomy. This paper predicts that participants will cite the "medicine delivery robots" urgency as an exonerating factor. However justifications for blaming the "trash disposal robot"will likely focus on its lack of consequence, justifying the functional scapegoating observed in Furlough et al.[2] work. Fewer responses will attribute blame to the system design or programming because the distinct social roles provided in the briefing will tilt the participants focus on the robots than any external factors.

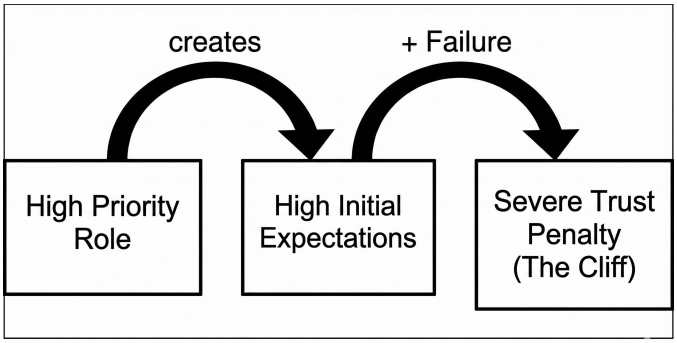

Fig. 3. Theoretical Model of Competence Loss.

V. DISCUSSION

A. THE COST OF HIGH EXPECTATIONS

The predected results suggests that high-priority robot experinces more severe penalty in competence ratings when it failed then the low-priority one this alignswith expectancy-disconfirmation theory, as recently applied to HRI by Ye et al.[1] they suggest that higher initial expectations create

Fig. 3: Theoretical Model of Competence Loss. The hypothesized pathway based on Expectancy-Disconfirmation Theory [1], where high initial expectations lead to harsher penalties for failure compared to low expectations.

larger contrast when performince fails, which leads to negative disconfirmation. In this study, the label “Medicine delivery robot” establishes a high baseline of expected reliability. Whenever this expecation is violated the resulting drop in the perceived competence is way steeper then for trash robot as users hold low expecations from it. This confirms that roles acts as a framing device which sets the trust fall distance.

B. SOCIAL HIERARCHY AND BLAME

the tendency of participants for attributing blame to the “trash disposal robot” even when the faliure was not caused by it supports the Social Hierarchy Hypothesis. Furlough et al.[2] demonstrated that blame attribution is heavily dependent on a robot’s perceived characteristics and role definition. Extending their findings on attribution to social hierarchy. My results suggest that this bias also extends to robot-robot teams. Participants view “trash disposal robot” as expandable when compared to the “medicine delivery robot”, which leads to a cognitive bias wher the lower-status robot is deemed responsible for yielding, regardless of the physical navigation constraints.

C. RATIONALIZATION OF BLAME

We predict that people will use the “urgency” of the “Medicine Delivery robot” as an excuse to let it of the hook. Which aligns with the findings of kim and hinds [3] regarding attribution transparency. They got to konw that when users understand the “why” behind the robots constraints they are more willing to forgive technical errors. Our predicted data suggest that users actively rationalize the faliure of “medicine delivery robot” by focusing on the purpose it has, while also judging the “trash disposal robot” on its performance. This indicates that role priority serves as a tie breaker for assigning blame.

D. VALIDITY OF VIRTUAL STIMULI

Methodologically, use of animated video stimuli is supported by previous HRI research which established that animation principles can effectively convey social intent and robot readability in video based trials. While some studies notes that virtual interactions have a lower psychological threat than physical ones, the consistency of our predicted blame patterns with physical-robot studies that the social judgement of blame is similar accross virtual and physical embodiements.

E. LIMITATIONS

While this study provides foundational insights into role-based blame attribution, several limitations must be acknowledged. First, the use of video-based stimuli, while necessary for controlling experimental conditions, lacks the physical presence and potential risk associated with real-world HRI. Participants viewing an animation on a screen may judge ”danger” or ”incompetence” less harshly than if they were physically standing next to a failing robot.

Second, the participant pool consists primarily of university students. This demographic is generally younger and more tech-savvy than the general population, potentially leading to different baseline expectations of robotic capabilities compared to, for example, older hospital patients or busy medical staff.

Finally, this study examined only two distinct points on the priority spectrum (Medicine vs. Trash). It remains unclear if these effects are linear; for instance, how would blame be attributed between two high-priority robots (e.g., a Medicine Robot vs. a Security Robot)? Future research must expand the taxonomy of roles to map the full landscape of robotic social hierarchy.

F. FUTURE WORK

Future iterations of this research should focus on transitioning the experimental protocol from a proposed design to emperical data collection, recruting the target sample of 128 participants to statistically validate the “social hierarchy’ and “expectation violation” hypotheses. While this study used animated videos stimuli to ensure experimental consistency, future work could enhance ecological validity by replacing the scenarios with physical robots in a controlled hospital mock-up or maybe even a VR environment. also the participant pool beyond the general University population to include healthcare professionals, this would provide a critical insight into how domain specific expertise influnces blame attribution.

VI. CONCLUSION

This study proposed an experimental framework that investigates how the social roles assigined to a robot influnces the perception of humans about technical failures in multi-robot environments. By paring a high priority “medicine delivery robot” with a low priority “trash disposal robot” its highlited

that blame is not fully a function of navigational performance, but it is deepely attached to social expectations.

This studys theoritical analysis predicits 2 key outcomes. (1) that the high priority robots face a “competence cliff”, where any failure leads to a severe penalty in preceived capability due to the violation of higher performance expecations; and (2) thst blame attribution follows a social hierarchy, where lower status robots are dis proportionately held responsible for coordination failre. All thee findings suggest that ressilience in HRI is not is not just a technical challemge but also a social one. For the designers of future hospital systems or any similar enviroments, this implies that a robots labeld role serves as a powerful framing device. To maintain user acceptance , high-priority robots may need transparency mechanisms to mitigate the shock of failure, while low-priority robots might need explicit signaling to avoid becoming scapegoats for systemic errors.

VII. CONTRIBUTIONS

Priyesh Vashistha (Student Number: 2049905) was the sole contributer of this project. He was responsible for the conceptualization of research questions, the development of “social hierarchy” and “expec tation violation” hypotheses, and the design of the experimental methodology. He created all the experimental materials, including the animated video stimuli and the recruitment posters. Also he developed the full qualtrics survey instrument, managed the successful submession of GREB ethics application, and wrote this final report in its entirety.

ACKNOWLEDGMENTS

The author would like to thank Dr. Pan for their guidance and supervision throughout this project as the Principal Investigator.

REFERENCES

[1] S. Ye, G. Neville, M. Schrum, M. Gombolay, S. Chernova, and A. Howard, ”Human trust after robot mistakes: Study of the effects of different forms of robot communication,” in *Proc. RO-MAN*, 2019.

[2] C. Furlough, T. Stokes, and D. J. Gillan, ”Attributing blame to robots: I. The influence of robot autonomy,” Human Factors, 2019.

[3] T. Kim and P. J. Hinds, ”Who should I blame? Effects of autonomy and transparency on attributions in human-robot collaboration,” in *Proc. RO-MAN*, 2006.

[4] M. Desai et al., ”Impact of early-interaction robot errors on human trust,” in *Proc. HRI*, 2012.

[5] M. Salem, F. Eyssel, et al., ”Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust,” in *Proc. HRI*, 2015.

[6] D. P. M. van der Hoorn, A. Neerincx, and M. M. A. de Graaf, ”’I think you are doing a bad job!’: The effect of blame attribution by a robot in human-robot collaboration,” in *Proc. HRI*, 2021.