Are Explanations All We Need? Investigating the Impact of XAI on Robot Failures and Trust Recovery in Human-Robot Cooperation Tasks

ANONYMOUS AUTHOR(S)*

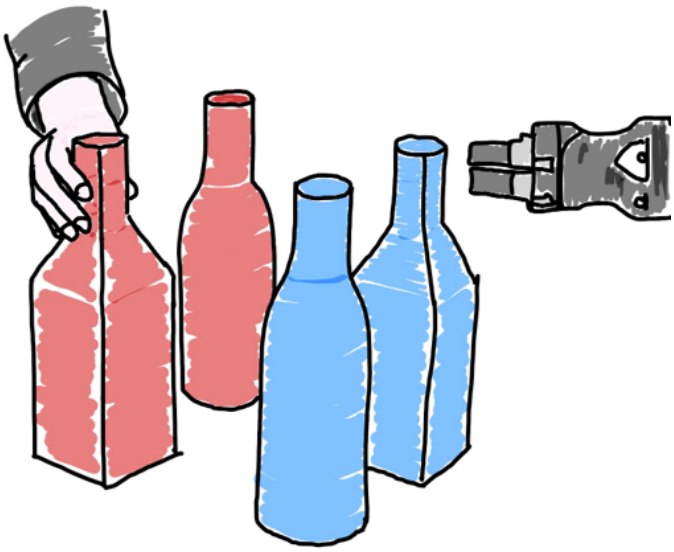

Fig. 1. Cooperative human-robot interaction task: Both, robot and human, sorting bottles regarding their shapes.

Failures are part of our lives. We have learned to deal with failures and fix them in private and work-life. However, with the emerging use of robots in an industrial context, we have to learn to deal with new error originators: robots. Studies show that robot errors decrease user trust after a robot failure, negatively impacting performance. The question is whether it is possible to restore the lost trust. Based on the literature, we identify two steps that can support trust recovery: (1) communicating failures with the help of XAI and (2) fixing robot failures. Further, we illustrate a setup for a planned study that investigates both aspects and end the paper with a discussion of open questions regarding human-centered and robot-centered issues that should be considered in failure mitigation.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **User centered design**; **User studies**.

Additional Key Words and Phrases: industrial robots, failure mitigation, trust recovery, user-study, human-centered AI

ACM Reference Format:

Anonymous Author(s). 2023. Are Explanations All We Need? Investigating the Impact of XAI on Robot Failures and Trust Recovery in Human-Robot Cooperation Tasks. In *ACM Conference on Human Factors in Computing Systems, April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

Manuscript submitted to ACM

1 INTRODUCTION

In the course of our lives, we make failures. As children, we learn how to deal with them by understanding failure and its effects, looking for solutions and trying to avoid them in the future. This task is often challenging and accompanies us throughout our lives. In addition to our failures, there are also the ones made by fellow human beings that we must learn to deal with. Failures occur when people interact and work together on a task. Communication is one crucial point that helps us correct and avoids these failures in the future. We get in touch with others, report to them when an error has occurred, and our counterpart can explain how it happened or how the problem could be fixed or avoided. Especially in the work context, investigating failures is vital to ensure safety [17]. Nowadays, more and more companies use industrial robots in addition to human workers. These can also cause failures that humans have to deal with. The investigation of robot failures in human perception and behaviour is the subject of numerous studies. In particular, the subject of research is the question of how a robot failure, which is often accompanied by a loss of trust on the part of the human, can be returned to trusting cooperation [24]. One possibility is the use of explanations. The research area of Explainable AI (XAI) deals with explaining decisions or internal states of AI systems. Therefore, XAI supports a human-centered AI design meaning that AI systems take human needs into account [18]. Whether explanations help restore trust in human-robot cooperation was the subject of investigation in the work of Hald et al. [10]. The authors investigated the influence of explanations on trust recovery after a robot failure in a VR setting. The authors concluded that explanations alone could not restore trust in an industrial robot after a robot failure.

If an explanation is insufficient to restore trust, the logical next step is realising the solution given in the explanation and, in the process, fixing the failure. This workshop paper provides an overview of research related to the influence of fixing the robot's failure on the recovery of trust. Based on existing literature, we present explanations in the context of robot failures and introduce the combination of explanations with user error correction as a possible helpful form of interaction after a robot failure. Furthermore, we reason about the implications for a planned follow-up study extending the results of Hald et al. [10]. Finally, our paper concludes with a discussion of open questions regarding human-centered and robot-centered issues in this context. Our work contributes towards investigating XAI in real-life use cases where humans and robots must successfully interact to solve tasks.

2 BACKGROUND

2.1 Robot Failures

When robots show not-intended behaviour, terms like failure, fault, or error describe this situation [12]. A failure describes a "behavior or service being performed by the system to deviate from the ideal, normal, or correct functionality" [2, p. 9]. Failures are caused by errors in the system (e.g., mechanical errors) [12]. The causes of errors are faults (e.g., a loose screw) [12]. Honig and Oron-Gilad [12] underline that in human-robot interaction, robot failures are common due to complex and often unstructured situations.

2.2 Trust

Authors such as de Visser et al. [5], Hancock et al. [11], Lee and See [13] point out the importance of an appropriate level of trust towards a robot, as too much trust can lead to misuse of the robot and thus dangerous situations. On the other hand, too little trust can lead to the robot not being used optimally.

Therefore, studying factors influencing trust between humans and machines has been an important research subject for decades. Different definitions of trust can be found in human-robot interaction. One of the most commonly used

definitions of interaction is that of Lee and See [13]. They define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability."[13, p. 54].

Hancock et al. [11] systematically summarised possible factors influencing trust. They point out three areas influencing people's trust in robots: robot-related, environment-related, and user-related factors. Robot-related factors include the robot's personality and behaviour and the reliability and predictability of the robot's behaviour. Environment-related factors include the nature and complexity of the task, as well as aspects such as culture and communication. Finally, user-related factors include, for example, the expertise and competence of users.

When investigating the reliability of robots, Salem et al. [20] found that a robot is perceived as less trustworthy after a failure. However, despite the decrease in trust, participants followed the robot's instructions.

3 DEALING WITH ROBOT FAILURES

When an error occurs in human-robot interaction, users often cannot understand how the failure happened, how it can be fixed and how it can be avoided in the future. Lack of understanding leads to a degradation of performance and mistrust of the user. Therefore, it is vital to address and solve the lack of understanding preventing loss of performance and trust.

Honig and Oron-Gilad [12] present their Robot Failure Human Information Processing Model. Their model addresses the mitigation of robot failure from a user-centred perspective by taking into account the user's cognitive abilities [12, p. 8]:

- **Communicating Failures**: This is about how a robot failure is communicated. Communication can be done with visual indicators (e.g. lights), screens with additional information or audio and speech (e.g. sounds, up to verbal communication).
- **Perception & Comprehension of Failures**: This aspect deals with the user's perception of the failure. Only when the user perceives the failure can countermeasures be initiated. Here, the user must understand and react to the robot's failure. The user's mental model also influences how they deal with a robot's failure. An incorrect model of the robot's behaviour can make it difficult to solve the problem after a failure.
- **Solving Failures**: User motivation is required to resolve a robot failure. In addition, users must decide how to fix the failure and translate this into action.

On this basis, strategies for dealing with robot failures can be developed for the robot or the user. Honig and Oron-Gilad [12] describes three different strategies: (1) set expectations (i.e., the user can correctly assess the failure potential of the robot), (2) communicate correctly (i.e., how and what is communicated in a failure situation) and (3) ask for help (i.e., the robot asks the user for help).

4 COMMUNICATING & FIXING ROBOT FAILURES

4.1 Communicating Robot Failure

In addition to harm reduction strategies, as stated before, providing explanations of the robot's behaviour and decisions could also help to increase user trust. The field of XAI has the goal to help users to "understand, appropriately trust, and effectively manage'[9, p.44] AI partners. de Visser et al. [5] provide an overview of trust in human-robot teams. They emphasise the importance of trust for successful human-robot tasks. They mention two essential directions for (re)building human-robot relationships through trust calibration: (1) predictive and preventive trust to minimise the risk of trust violation, and (2) reactive and reparative trust when trust is violated.

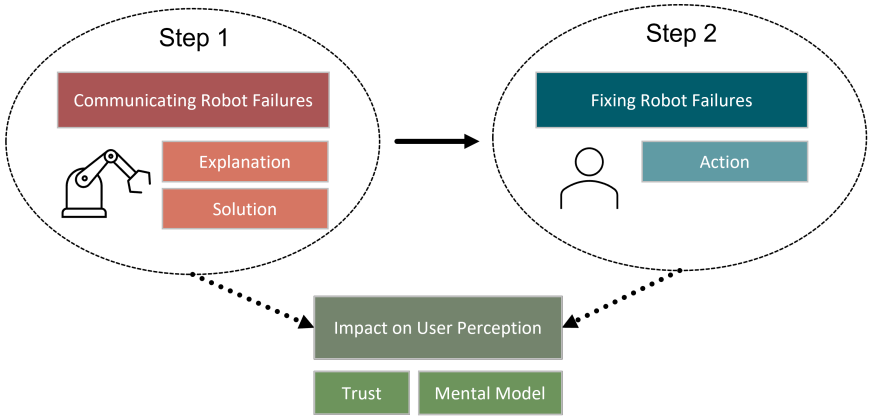

Fig. 2. Based on the literature, we conclude two steps for mitigating robot failures: first, a robot failure is communicated to the user with the help of an explanation and a possible solution. Second, the user comes into action by actually fixing the error. The steps can, but do not need to rely on each other and have an impact on users' perception of the robot.

One method for the predictive and reparative trust that de Visser et al. [5] suggests is using explanations. They suggest that explanations help calibrate trust by providing users with information about the robot. XAI approaches aim to gain insights into a robot's behaviour and goals. For example, Sheh [22] recommends using decision trees in explanatory human-robot dialogues. They argue that decision trees have two significant advantages: (1) the intrinsic explainability of if-then-else rules and (2) attribute-centric explanations that do not refer to a specific example but to attributes relevant to different actions. Therefore, they aim to use decision trees in a human-robot dialogue and enrich it with visual explanation modalities such as histograms. The work of Das et al. [4] uses handwritten explanations to study three different types of errors (i.e. navigation, arm motion planning and object recognition) of an industrial robot arm. They distinguish four different contents of explanations [4, p. 354]:

- **Action Based**: e.g., "Robot could not find the object."
- **Context Based**: e.g., "Robot could not find the object because the object is hidden from view."
- **Action Based History**: e.g., "The robot finished scanning objects at its current location, but could not find the desired object."
- **Context Based History**: e.g., "The robot finished scanning objects at its current location, but could not find the desired object because the desired object is hidden from view."

In an online user study, videos were shown where the robot performed a task. After three videos where the robot solved the task, participants saw six failure simulations without explanations. Afterwards, participants were asked to identify the cause of the failure and possible solutions. Finally, another 12 failure videos with explanations were shown. Again, participants had to identify the cause of the failure and possible solutions. The results indicate that context-based explanations are best suited for end users to identify the reason for a robot failure and find a helpful solution.

In the work of Hald et al. [10], the influence of explanations on the trust of the robot after a robot error is investigated. For this purpose, Hald et al. [10] used an interactive scenario where study participants had to sort bottles by shape with a robot. The participant had the task to sort the red bottles according to their shape, such that all red rectangular bottles will be within the white area on the left side and all round-shaped, red bottles on the right side on the white basis (see upper part in Figure 3 on page 6). The task for the cobot is identical, except its task comprises the blue bottles

instead of the red ones. The setting was realised in virtual reality (VR). In total three different conditions were tested for in the study:

- **No Explanation**: After the robot error occurred, participants only received information that an error had occurred and that the task had not been successfully completed.
- **Explanation**: Participants received an explanation of why the robot error had occurred. The explanation was projected onto the virtual table in front of them, and stated:"A computer vision error occurred. The system did not successfully distinguish the shapes in the current lighting conditions." (see Figure 3 on the next page)
- **Explanation and Solution**: In addition to the explanation, participants received information on how to avoid the error in the future.

As already found in numerous studies by other authors, Hald et al. [10] also found a loss of trust after the robot made a mistake. However, the three different conditions did not differ, i.e., no influence of explanations on a trust recovery after a robot error could be found. However, at the end of the study, the participants rated the explanations as helpful to trust or distrust the robot.

4.2 Fixing Robot Failure

So far, most of the research regarding trust in (faulty) robots focused on communication strategies, e.g., denying, apologizing for or explaining the failure [7], type of failure, temporal occurrence of failure [19], the severity of the failure [19] or the personal relevance of failures [8]. Also, characteristics of the robot, e.g., its apparel, have been the object of research on its influence on the trust. To the best of our knowledge, investigating trust repair strategies that involve resolving the failure after its occurrence has not been the main focus of any research activity. Morales et al. [16] investigated whether one is willing to assist a robot in case it not being able to grab an object. They combined the occurrence of the failure with other interactions that included the robot behaving in a harmful way, e.g., throwing a (foam) potato at the participant or crushing a bag of potato chips. The order of the different types of interaction had an influence on the willingness to help the robot. Specifically, the willingness decreased significantly when the risky behaviour, independent of the participant or an object potentially being harmed, occurred before the robot failed at grabbing a box [16]. Morales et al. [16] show that successfully resolving a robot's failure is affected by its behaviour.

Explaining the reason for a failure without fixing the failure mitigates the loss of trust after the failure [10]. Adding the proposal of a possible solution, without the possibility of correcting the failure, did not gain any further benefit in mitigating the trust loss. In Hald et al. [10] resolving the failure itself was not tested for. However, the reason for the failure of the robot as well as the solution resolving the failure was given in form of a text-based explanation (see Figure Figure 3 on the following page).

The logical follow-up question is: What influence has fixing the failure upon the explanation and/or solution given to one? Thus, we will use the experimental setup by Hald et al. [10] as a basis and adapt it to the needs for answering our research question (see figure 4 on page 7. For this, we made three changes.

Firstly, we will change the VR setting to real life as Hald et al. [10] already outlined in their look in their outlook.

One might argue that there are significant differences between VR and reality complicating the comparison of results between those different settings. While this seems to be true for the perception of proxemics [6] and safety [15]. Stressing participants with the Trier Social Stress Test elicited similar subjective ratings, EEG and EDA results when comparing the VR and real setting. Only the cortisol level differed between the conditions [23]. Also, the perception of eeriness and likeability did not differ between the presentation of a humanoid robot in real life or in VR [14]. Ünal et al.

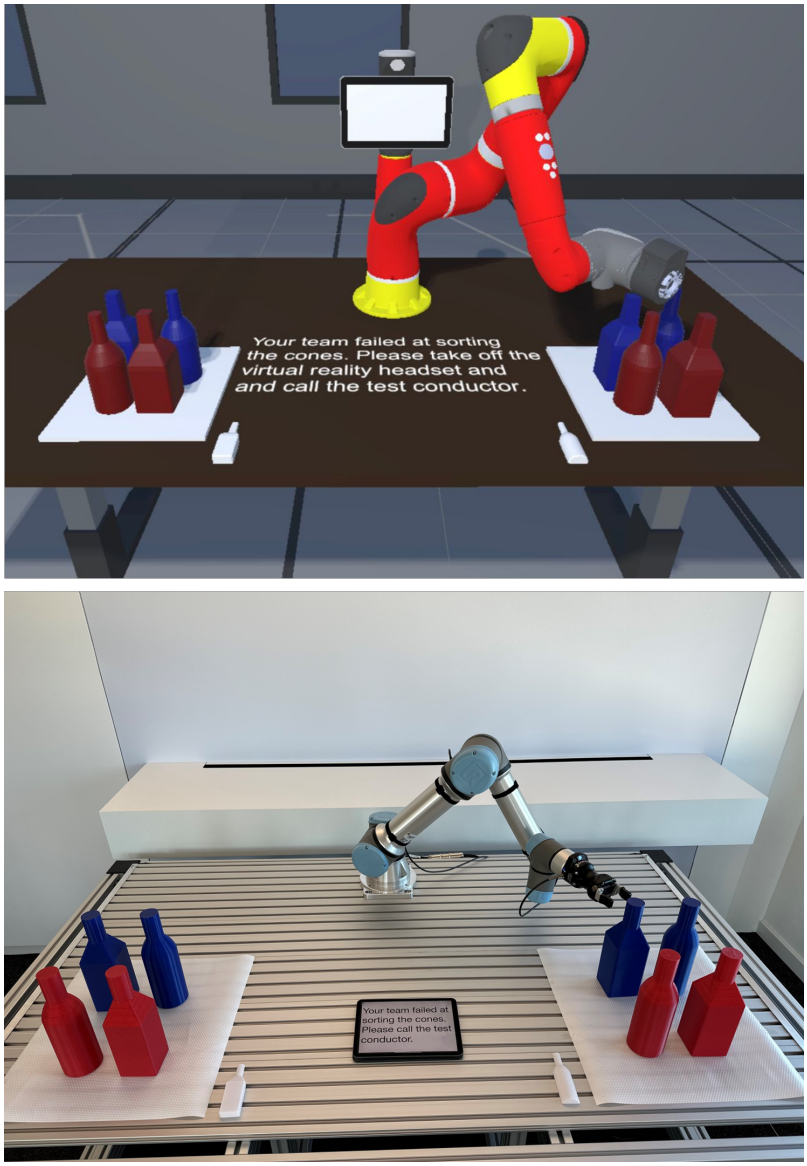

Fig. 3. The virtual setup (top) presented in Hald et al. [10] serves as a basis for the investigation of the helpfulness of explanations in solving robot failures in a real-life setup (bottom).

[26] found no difference when comparing the restorative effect of being in a real-life natural environment or a similar VR environment. We cannot rule out that our experimental setup in VR or in real life might have an influence on the results of explanations and/or solving the failure on trusting the robot or not, thus, we shall be very cautious when comparing results.

Secondly, we will extend the number of conditions. The additional fourth condition entails no explanation, but a proposal of a solution. This allows examining both factors (no) explanation and (no) solutions separately as well as in interaction.

Thirdly and most important change, the failure will be fixed in all four conditions. The **explanation and solution** condition of the previous study entailed the information that *"better lightning conditions will help with successful sorting"* [10, p. 223]. Consequently, a reasonable fix for the failure is turning on the light. For the sake of maintaining the possibility of comparing all four conditions, we must have a coherent way of acting upon the proposed solution to the failure. It would be presumptuous to expect participants in the condition without any explanation or solution to initiate turning on the light. While, it would be of interest to leave the action to the participants, as it might give them a feeling of greater involvement with the robot, the deviation between conditions would be too great to allow any comparison. Thus, standardizing across conditions by leaving the action upon the study conductor to turn on the light allows a sound comparison.

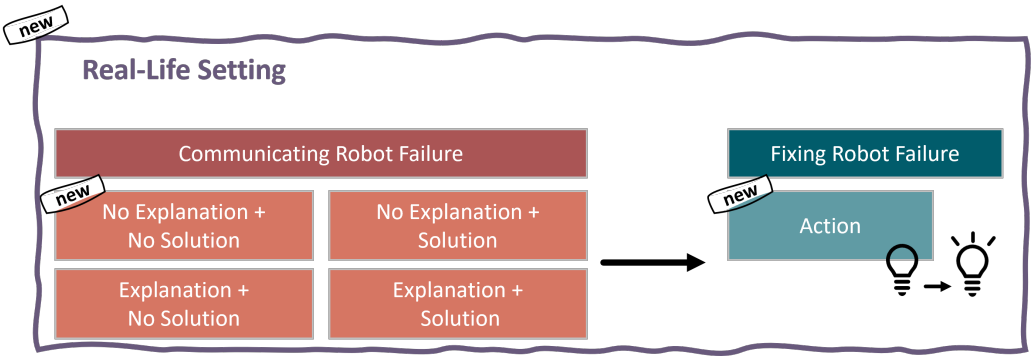

Fig. 4. The extended study design compared to the work of Hald et al. [10]. The real-life setting, an "explanation and no solution" condition, and the action of the study conductor to fix the robot failure are new components of the planned study.

5 DISCUSSION

With the perspective of the great increase of robots advancing into several areas of our life, more and more people, especially those without great knowledge of technical systems, will interact with robots and observe robots fail and possibly lose trust. Thus, the communication of failures, but also the prevention of as well as the intervention after the occurrence of failures must be further researched to allow robots to be a further supporting component in facilitating our lives. A list of open questions related to our intended experiment, but also beyond the scope it is being touched upon.

Who is responsible for robot failure mitigation? When dealing with failure mitigation, the question inevitably arises: Who is responsible for fixing the failure? The person (or robot) who caused the failure? Or the person who has an idea for a solution? Or the person (or robot) who can implement the solution? Often, humans have to correct a machine's failures because people are good at recognising and solving failures, or at least not making them worse, based on the situation's circumstances and the available options. In communication situations, it becomes apparent that people are primarily responsible for resolving miscommunication, as simple dialogue systems are often incapable of doing so [25].

In general, however, the question arises as to whether this distribution of roles makes sense. Sauppé and Mutlu [21] work shows that users want robots to help them correct failures.

For which kind of failure is communication & fixing appropriate? In addition to the responsibility for fixing the failure, the question arises: For which types of failure and in which situations are communication and fixing of failures beneficial? Especially in the case of repeated failures known to users, excessive communication by means of explanations could be counterproductive and lead to frustration among users. Moreover, not all failures can be fixed by users. As Honig and Oron-Gilad [12] highlights, whether a user can fix a failure depends on their skills, if the failure is perceived as such and if the user is motivated to fix it. A survey asking employees about the use of AI in their companies shows that the development of AI skills in employees is a challenge for companies to successfully use AI technologies [1]. The failure can be successfully resolved only when a user has the knowledge and skills about the robot and how it works. Even if there is a step-by-step solution to fix the failure, the user must understand and be able to implement the individual steps.

How should the failure be communicated? A failure can only be corrected if it is recognised. Therefore, the question arises: How should a failure be communicated so that users perceive it and are motivated to fix it? Research provides several possible design approaches and studies. Cha et al. [3], for example, pointed out that sound can transport the need for help and that light can help to communicate the urgency. The challenge in designing failure communication is that it must be appropriate to the application and situation. For example, a loud, continuous sound from a malfunctioning service robot in a meeting would be inappropriate, making it very difficult to continue the meeting. However, such a loud sound could be helpful in a production hall, as there is often a loud noise level, and the information would otherwise not be perceived. However, alternative forms of communication, e.g. notifying workers individually or using light signals, could also be helpful communication tools.

What about the robot-centered perspective? Based upon the elaboration so far, it can be assumed that we perceive the human-centered perspective to be a vital aspect of human-robot interaction. Not taking the human in the centre ought to result in technical devices with a valuable purpose and with a user-unfriendly interface making their usefulness questionable. However, the technical system's or, in particular, the robot's perspective must also be taken into account when preventing, communicating, and resolving failures. Using the example of Hald et al. [10] and our proposed study idea: How shall the robot know of the occurrence of a failure? What indicates the failure? How can the robot differentiate between successfully having sorted the bottles and not completing its task? Given it knows of the failure, how shall it know it is a computer vision error due to bad lighting conditions and not a misplaced bottle? Given it knows the cause of the failure, how shall it know that turning on the light can fix the failure and not move its own position to move its shadow of the bottles?

Only the symbiotically linking of both perspectives can create a trustful interaction between humans and robots.

REFERENCES

[1] Elisabeth André, Wilhelm Bauer, Martin Braun, Chi Tai Dang, Matthias Peissner, and Katharina Weitz. 2021. *KI-Kompetenzentwicklung bei Sach und Produktionsarbeit.* Plattform Lernende Systeme. 1–34 pages.

[2] Daniel J Brooks. 2017. *A human-centric approach to autonomous robot failures.* Ph. D. Dissertation.

[3] Elizabeth Cha, Maja Matarić, and Terrence Fong. 2016. Nonverbal signaling for non-humanoid robots during human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* 601–602. https://doi.org/10.1109/HRI.2016.7451876 ISSN: 2167-2148.

[4] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of HRI '21: ACM/IEEE International Conference on Human-Robot Interaction, Boulder, CO, USA, March*

8-11, 2021, Cindy L. Bethel, Ana Paiva, Elizabeth Broadbent, David Feil-Seifer, and Daniel Szafir (Eds.). ACM, New York, NY, USA, 351–360. https://doi.org/10.1145/3434073.3444657

[5] Ewart J de Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics* 12, 2 (2020), 459–478. https://doi.org/10.1007/s12369-019-00596-x

[6] Sahba El-Shawa, Noah Kraemer, Sara Sheikholeslami, Ross Mead, and Elizabeth A. Croft. 2017. "Is this the real life? Is this just fantasy?": Human proxemic preferences for recognizing robot gestures in physical reality and virtual reality. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 341–348. https://doi.org/10.1109/IROS.2017.8202178 ISSN: 2153-0866.

[7] Connor Esterwood and Lionel P. Robert Jr. 2023. Three Strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior* 142 (May 2023), 107658. https://doi.org/10.1016/j.chb.2023.107658

[8] Romi Gideoni, Shanee Honig, and Tal Oron-Gilad. 2022. Is It Personal? The Impact of Personally Relevant Robotic Failures (PeRFs) on Humans' Trust, Likeability, and Willingness to Use the Robot. *International Journal of Social Robotics* (Sept. 2022). https://doi.org/10.1007/s12369-022-00912-y

[9] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine* 40, 2 (2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[10] Kasper Hald, Katharina Weitz, Elisabeth André, and Matthias Rehm. 2021. "An Error Occurred!" - Trust Repair With Virtual Robot Using Levels of Mistake Explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. Association for Computing Machinery, New York, NY, USA, 218–226. https://doi.org/10.1145/3472307.3484170

[11] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[12] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology* 9 (2018), 861. https://doi.org/10.3389/fpsyg.2018.00861

[13] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[14] Martina Mara, Jan-Philipp Stein, Marc Erich Latoschik, Birgit Lugrin, Constanze Schreiner, Rafael Hostettler, and Markus Appel. 2021. User Responses to a Humanoid Robot Observed in Real Life, Virtual Reality, 3D and 2D. *Frontiers in Psychology* 12 (2021). https://www.frontiersin.org/articles/10.3389/fpsyg.2021.633178

[15] Daxton Mitchell, HeeSun Choi, and Justin M. Haney. 2020. Safety Perception and Behaviors during Human-Robot Interaction in Virtual Environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (Dec. 2020), 2087–2091. https://doi.org/10.1177/1071181320641506

[16] Cecilia G. Morales, Elizabeth J. Carter, Xiang Zhi Tan, and Aaron Steinfeld. 2019. Interaction Needs and Opportunities for Failing Robots. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. Association for Computing Machinery, New York, NY, USA, 659–670. https://doi.org/10.1145/3322276.3322345

[17] James Reason. 2000. Human error: models and management. *Bmj* 320, 7237 (2000), 768–770.

[18] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1, 1 (2019), 33–36. https://doi.org/10.1002/hbe2.117

[19] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2017. How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. In *Social Robotics (Lecture Notes in Computer Science)*, Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He (Eds.). Springer International Publishing, Cham, 42–52. https://doi.org/10.1007/978-3-319-70022-9_5

[20] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. 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 2015, Portland, OR, USA, March 2-5, 2015*, Julie A. Adams, William D. Smart, Bilge Mutlu, and Leila Takayama (Eds.). ACM, 141–148. https://doi.org/10.1145/2696454.2696497

[21] Allison Sauppé and Bilge Mutlu. 2015. 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)*. Association for Computing Machinery, New York, NY, USA, 3613–3622. https://doi.org/10.1145/2702123.2702181

[22] Raymond Ka-Man Sheh. 2017. "Why Did You Do That?" Explainable Intelligent Robots. In *The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA (AAAI Technical Report, Vol. WS-17)*. AAAI Press.

[23] Youssef Shiban, Julia Diemer, Simone Brandl, Rebecca Zack, Andreas Mühlberger, and Stefan Wüst. 2016. Trier Social Stress Test in vivo and in virtual reality: Dissociation of response domains. *International Journal of Psychophysiology* 110 (Dec. 2016), 47–55. https://doi.org/10.1016/j.ijpsycho.2016.10.008

[24] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. 2020. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 3–12. https://doi.org/10.1145/3319502.3374793

[25] Katharina Weitz, Lindsey Vanderlyn, Thang Vu Ngoc, and Elisabeth André. 2021. It's our fault!": Insights Into Users' Understanding and Interaction With an Explanatory Collaborative Dialog System. In *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, Arianna Bisazza and Omri Abend (Eds.). Association for Computational Linguistics, 1–16. https://doi.org/10.18653/v1/2021.conll-1.1

[26] A. B. Ünal, R. Pals, L. Steg, F. W. Siero, and K. I. van der Zee. 2022. Is virtual reality a valid tool for restorative environments research? *Urban Forestry & Urban Greening* 74 (Aug. 2022), 127673. https://doi.org/10.1016/j.ufug.2022.127673