A Literature Review of Trust Repair in HRI

Connor Esterwood[1] & Lionel P. Robert[2]

***Abstract*— Trust is vital for effective human–robot teams. Trust is unstable, however, and it changes over time, with decreases in trust occurring when robots make mistakes. In such cases, certain strategies identified in the human–human literature can be deployed to repair trust, including apologies, denials, explanations, and promises. Whether these strategies work in the human–robot domain, however, remains largely unknown. This is primarily because of the fragmented and dispersed state of the current literature on trust repair in HRI. As a result, this paper brings together studies on trust repair in HRI and presents a more cohesive view of when apologies, denials, explanations, and promises have been seen to repair trust. In doing so, this paper also highlights possible gaps and proposes future work. This contributes to the literature in several ways but primarily provides a starting point for future research and recommendations for studies seeking to determine how trust can be repaired in HRI.**

I. INTRODUCTION

Human–robot teams hold great potential to extend the capabilities of human flexibility, adaptability, and creativity by coupling these with a robot's capacity for accuracy, speed, and consistency [1], [2], [3], [4]. As a result, human–robot teams are beginning to emerge in a variety of work environments [5], [6]. For example, a range of fast-food chains and restaurants have added robots to their teams [7], [8], [9] and retailers have deployed shelf-scanning robots [10], [11]. Trust, however, is required for the success of these human–robot teams [12], [13], [14], [6], [15]. Trust, or accepting the vulnerability associated with relying on others, is vital for human–robot teams but is often decreased by trust violations [16], [17], [18]. Trust is dynamic and changes over time, with significant decreases in trust occurring when robots inevitably make mistakes (i.e. trust violations) [19], [20], [21]. Researchers have studied various approaches to mitigating the negative impacts of trust violations via various trust repair strategies such as apologies, denials, explanations, or promises [22].

In recent years, a number of HRI trust repair studies have emerged, yet it is not clear whether or when a particular trust repair strategy is effective. As a result, designers and researchers are unable to reliably determine whether and when certain trust repairs are effective. In response, we complemented and built upon prior discussions of trust repair in the HRI literature [19] by conducting an up-to-date and systematic review of the literature on this topic and highlighting where future work is needed.

[1] PhD student at the School of Information. The University of Michigan, 105 S State St, Ann Arbor, MI 48109, United States of America cte@umich.edu

[2] Associate professor at the School of Information and a core faculty member at the Michigan Robotics Institute. The University of Michigan, 105 S State St, Ann Arbor, MI 48109, United States of America lprobert@umich.edu

II. BACKGROUND

To contextualize the results of the HRI literature we now introduce and discuss the concepts of trust violations and repair. To this end, we first provide a summarization of trust violations and how they have been categorized. Second, we define trust repairs and provide a description of the various strategies used to repair trust.

A. Trust Violations

Trust violations are events that reduce a trustor's perceptions of trustworthiness and trust in a trustee [23], [24]. This reduction in trust has been shown to negatively impact a team's collaborative potential, regardless of whether that team is made of humans [25] or includes robots [19], [20]. In HRI, three categories of trust violations have been defined: violations of ability, integrity, and benevolence [26].

Ability-based trust violations occur when a robot violates a human's expectations of the robot's performance. Examples of this type of violation in the HRI literature are instances where a robot makes a mistake unintentionally [27]. Integrity-based trust violations occur when a robot violates a human's expectations of the robot's honesty and ethical consistency. Examples of integrity-based violations in the HRI literature take the form of a robot acting against a teammate during collaborative tasks, even when they promised not to [27]. Benevolence-based trust violations occur when a robot fails to meet a human's expectations of the robot's purpose (i.e. who the robot is designed for). Benevolence-based violations differ from integrity-based violations in that benevolence-based violations indicate a degree of malice or ill will, whereas integrity-based violations do not. The HRI literature has not explicitly examined violations of benevolence. Possible forms of this type of violation, however, might include a robot undermining a human teammate in service of another, a robot rejecting or ignoring feedback, or a robot acting in direct conflict with a human's desires. Ultimately, all of these types of violations lead to reductions in trust.

B. Trust Repairs

Trust repairs can be defined as efforts undertaken to restore trust following an actual or perceived trust violation [28], [29], [30]. Trust repair strategies typically take one of four forms: apologies, denials, explanations, or promises [22], [31], [32]. Overall each of these strategies has been seen as generally effective, but various theoretical justifications

Esterwood, C. and Robert, L. P. (2022). **A Literature Review of Trust Repair in HRI,** *Proceedings of 31th IEEE International Conference on Robot and Human Interactive Communication* (**RO-MAN 2022**), August 29 - September 2, 2022. Naples, Italy.

for how these strategies repair trust are less discussed and subject to ongoing debate [23].

Apologies are attempts to re-frame the trustee after a violation of trust has occurred [33], [34], [35]. Specifically, apologies are a type of verbal trust repair strategy that seeks to express remorse for a relational or social transgression coupled with an explicit or implicit admission of guilt [36], [22], [35]. For example, the phrase "I'm sorry I did that" is an apology. On the other hand, *denials* are rejections of culpability coupled with one or more external reasons as to why a violation of trust was committed [19, Pg. 30:16]. The goal of a denial is to shift blame from the trustee to somewhere else [37], [35]. For example, "I didn't mess up ... something else must have happened" is a denial.

Explanations are explicit verbal statements made with the goal of providing the reasons why an action has occurred [38]. An example of an explanation might be, "I see, my sensors were not calibrated so I missed the object." Explanations provide transparency, which helps others understand the inner workings or logic behind why an event happened [39], [37], [40], [22], [41]. *Promises* are assertions by a trustee designed to convey positive intentions about future acts [31]. An example of a promise is the statement, "I promise I'll do this correctly next time." A full literature review on trust repair in HRI has not been conducted. To rectify this, we transition to a discussion of our methodology for identifying and reviewing appropriate works on this topic.

III. METHOD

A. Search Process

We identified studies by conducting repeated searches across three search engines. Specifically, we focused our search using Google Scholar, Scopus, and the ACM Digital Library. To define our search terms we first conducted a series of naive searches and through iteration settled on a fixed set of search terms and Boolean logic. In particular, our terms were: *(Trust OR Trustworthiness) AND (Repair OR Recovery) AND (Robot OR "HRI" OR "Human Robot Interaction")*. We then progressively paged through the results from these terms until no relevant results were present. Relevant results were results that contained the aforementioned search terms in their titles or abstracts. After conducting this search and removing duplicates, we identified 566 results. These results were then subject to a multi-stage screening process resulting in the identification of 22 total studies.

B. Study Screening

Studies in this review underwent a three-stage screening procedure during which we applied progressively stricter inclusion criteria. Studies were screened first based on title, second on abstract, and third on their full-text content. Title screening required that studies be classified as academic works (peer-reviewed publications, theses, dissertations, etc.), written in English, with their titles or abstracts containing one or more of our search terms. Studies were then subject to a second screening focused on their abstracts. Studies at this stage of screening were required to meet all previous criteria as well as being empirical, focusing on embodied physical action robots, and including interactions between at least one human and at least one robot. Finally, we screened studies based on their full-text content. This required studies to meet all prior screening criteria and explicitly report the impact of one or more trust repairs.

In addition to all of these screening criteria, we implemented a set of exclusion criteria. These exclusion criteria were applied at all points in the screening process. Specifically, we excluded studies if they focused on embodied virtual action (EVA) agents (e.g., chat-bots) or telepresence robots, or did not examine how trust repairs impact human subjects. Studies of EVA agents and telepresence robots were excluded as a result of this paper's focus on robots as opposed to virtual agents and artificial intelligence in general. After conducting all screening, we identified 22 studies and incorporated them into this review.

IV. RESULTS OF LITERATURE REVIEW

A. Repair Strategies

Across the HRI trust repair literature, a range of trust repair strategies were examined. Across the 22 studies included in our review, 13 examined more than one trust repair strategy and 7 combined multiple strategies together. Of these strategies, the most popular was apologies (k=12), followed by explanations (k=5), denials (k=5), and promises (k=3). Aside from these, replacement [42], assurance of competency, gas-lighting [43], offering compensation [44], and requesting help [45] were examined by one study each. Regarding different combinations of repair strategies, apologies were combined with promises in three cases [46], [27], [47], and with explanations in four cases [48], [49], [50], [51]. Meanwhile, one study combined explanations with promises [52]. A breakdown of the repairs in the HRI literature is available in table I.

B. Outcomes

The current state of the trust repair literature in HRI encompasses a wide range of outcomes. These outcomes can be grouped into four broad categories: trust, acceptance, anthropomorphism, and performance. Our review found that the most common of these categories was trust. Outcomes of this type measured trust and trustworthiness and/or focused on trust-relevant behaviors such as compliance. In all, 19 studies examined the impact of trust repair strategies on trust. The second most common outcome examined was acceptance. Outcomes falling into this group were those that either measured acceptance overall or focused on one or more sub-components of acceptance such as attitude, perceived usefulness, or perceived ease of use (see: [53] and [54]) and accounted for 7 studies.

Aside from the trust and acceptance outcomes, the impact of different trust repair strategies on performance as well as humans' perceptions of a robot's anthropomorphism were also investigated. Studies categorized as examining performance were those that measured task completion time,

Study	Task	Repair(s)	Outcome(s)
[46]	Target Identification.	Apology,Apology & Promise	Acceptance,Trust
[55]	Player in Competitive Game	Other	Trust
[56]	Guide Robot	Apology, Apology & Explanation, Explanation	Acceptance,Trust
[57]	Host Tangram Game	Explanation	Trust
[45]	Player in Collaboration Game	Apology,	Acceptance, Anthropomorphism
[49]	Move objects	Apology & Explanation	Acceptance, Trust
[43]	Driving	Apology, Denial, Explanation	Trust
[58]	Driving	Apology, Denial	Trust
[50]	Provide recommendations	Apology, Apology & Explanation, Explanation	Trust
[44]	Identify and retrieve snacks.	Apology, Compensation	Acceptance, Trust
[59]	Monitoring and notification.	Other	Trust, Performance
[60]	Manufacturing Move Objects Task	Other	Trust
[61]	Driving	Other	Trust, Performance, Acceptance
[62]	Help solve math problems	Apology, Explanation	Trust
[63]	Lead evacuation	Apology	Trust
[42]	Move packages	Promise	Anthropomorphism, Attitudes,Trust
[64]	Lead evacuation	Apology, Promise	Trust
[51]	Player in Collaborative Game	Apology & Explanation	Anthropomorphism
[27]	Player in Competitive Game	Apology & Promise, Denial	Trust
[52]	Military Reconnaissance	Explanation & Promise, Promise	Acceptance,Trust
[65]	Answer healthcare questions.	Apology, Denial	Trust
[66]	Taxi dispatching recommendations	Apology, Denial	Trust

TABLE I

SUMMARY OF TASKS, REPAIRS, AND OUTCOMES EXAMINED IN THE HRI LITERATURE.

reliability, and individual or team performance. Two studies across the literature examined performance outcomes. Studies with outcomes falling into the category of anthropomorphism measured how animate, warm, intelligent, intentional, or purposeful robots appeared. In total, 3 studies examined outcomes of this kind. Table I summarizes these outcomes.

C. Findings

The majority of outcomes associated with trust repairs in HRI focused on trust. Given the relative dominance of this outcome and its relevance to assessing the efficacy of various trust repair strategies, we therefore focused on trust as our outcome of interest. In particular, we examined the existing HRI trust repair literature with the goal of determining how trust repairs have impacted trust in HRI. Furthermore, we focused on the repair strategies of apologies, denials, explanations, and promises because these strategies overlap with the human–human trust repair literature and make up the majority of repairs used in HRI. Given this focus, the following subsections summarize how apologies, denials, explanations, and promises have been found to impact trust. A summary is given in table II.

1) Apologies: The impact of apologies on trust in HRI has been largely mixed, with three studies finding that apologies repaired trust [46], [62], [58], two finding that apologies did not repair trust [44], [43], and one finding that apologies damaged trust [56]. One possible explanation for these mixed results could relate to different moderating factors, or moderators. In particular, timing has been examined and found to be influential. In such cases, apologies given closer to the event when trust was violated were found to be more effective than apologies withheld and given after time had passed [63], [64], [50]. Notably, no examination of apologies given after multiple trust violations and repairs has been conducted to date.

2) Denials: The impact of denials on trust repair appears mixed, with one study showing that denials were effective [58] and one showing that denials were ineffective [43]. Once again, moderators might explain these mixed results. To date, however, only one moderator has been explored, namely, the type of trust violation that occurred. To this end, [65] and [27] examined this potential moderator and found no significant differences in trust between denials given after ability-based versus integrity-based trust violations in the case of [27] and denials given after violations of logic, semantics, or syntax in the case of [65]. This conflicts with findings from the human-human literature that the type of violation is central to when a denial is likely to be effective [24], [22]. While this might indicate a fundamental difference between humans and robots in terms of denials and trust repair, these results could also be influenced by the relatively clear cause of the trust violations examined. This could be the case as given that the cause of the trust violations used in most HRI studies is clearly presented to subjects reducing the validity of denials. This is because denials build on a subject's doubt to provide an alternative interpretation of events [24]. Given that no studies in the HRI literature to date have explicitly accounted for a user's doubt of events or employed purposefully obscured causes for trust violations, much remains unknown as to the true suitability of denials as a trust repair strategy.

3) Explanations: The impacts of explanations on trust have been mixed, with one study indicating that explanations repair trust [62] and five studies showing that explanations do not [50], [43], [44], [56], [67]. Furthermore, two moderators appear to impact the efficacy of explanations, specifically, the severity of a trust violation [57] and the timing of an explanation [64], [50]. For severity, results were fairly straightforward in that the more severe the violation, the less effective explanations were [57]. However, for timing of explanations, studies produced conflicting findings; specifically, one study found that timing was an important variable [64] and one found that it was not [50]. A closer examination of these studies shows differences in samples sizes that might explain this disagreement. In particular, the study finding significant results [64] for timing employed sizably more subjects (n=39) than the study that found non-significant results (n=18) [50]. Regardless, this disagreement warrants further examination, and more studies on the effects of timing are needed.

4) Promises: Promises in the HRI trust repair literature have mostly been examined in combination with apologies

	Repairs Trust	Does Not Repair Trust	Damages Trust	Depends on Moderator(s)
Apology	[46], [62], [58]	[43], [44]	[56]	Timing [64], [50]
Denials	[58]	[43]	–	Violation Type [65], [27]
Explanation	[62]	[50], [43], [56], [44], [67]	–	Severity [57] Timing [64]
Promise	[42]	[52]	–	Timing [64]

TABLE II

SUMMARY OF HRI TRUST REPAIR STUDIES FINDINGS BY STRATEGY.

or explanations [27], [46], [52], [47]. Where promises have been examined independently, results have been mixed, with one study showing that promises effectively repair trust [42] and one showing that they do not [52]. One reason for these findings might once again stem from moderators. For example, promises have been argued to be influenced by timing, where promises given promptly after a violation have been found to be more effective than those given after a delay [64]. Ultimately, promises have potential as a repair strategy but there is a lack of robust examination of this repair strategy. In particular, more studies focusing on promises independent from apologies are needed. Such studies should account for not only timing but also violation type.

V. DISCUSSION & RESEARCH OPPORTUNITIES

The current literature on trust repair in HRI paints a complex and disjointed picture of whether and how robots can repair trust given that findings on the impacts of apologies, denials, explanations, and promises are largely mixed. Additionally, while moderating factors might help to explain these mixed results, the literature examining these moderators is often mixed, incomplete, or absent. As a result, it is hard to know what repair strategies are effective and when. Furthermore, studies have not thoroughly examined *how* trust repairs function (i.e. the mechanisms through which they act). As a result, future studies are warranted. In particular, studies are needed that examine repair strategies independently; compare those strategies to one another; consider theoretical mechanisms (i.e. mediators); and also account for timing, violation type, and possibly other influential moderators.

To this point, we now discuss four pressing gaps across the HRI trust repair literature in the hope of spurring future research. First, studies should incorporate trust violation type as a possible moderator. This is because trust violation type's influence on trust repairs is supported in multiple places in the human–human literature [24], [22] as well as in the HRI literature [27], [65]. In the HRI literature, however, studies have only looked at denials and have not considered violations of benevolence alongside ability and integrity. Therefore, future work should consider the full range of possible violation types and how these might impact other repairs such as apologies, explanations, and promises.

Second, only one study examined the impact of perceptive violation severity as an influential moderator of trust repairs on trust [57]. Given that this study found a direct link between perceptions of violation severity and the efficacy of explanations, it is also possible that other repairs are similarly impacted. In particular, studies finding that trust repairs were ineffective might be using more severe violations of trust, while studies finding that repairs were effective might be using less severe violations of trust. As a result, it could be that researchers seeking to determine the individual impacts of trust repairs are, in fact, only seeing the impact of different trust violations. Future researchers should therefore investigate the relative severity of violations and seek to control for this variable when designing studies on trust and trust repair in HRI.

Third, the degree of doubt a trustor has regarding the cause of a trust violation has been argued to be a prime determinant of when denials are likely to be effective methods of trust repair [24]. This could have influenced results associated with denials leading to the observed mixed results. For example, the study finding that denials were effective at repairing trust used audio vignettes [58], whereas the study finding that denials were ineffective used videos [43]. It is possible that the use of audio produced more vague circumstances and left more to the imagination than the use of clear violations captured on video. As a result, the impacts of denials might have been obscured. Aside from denials, trust repair stratagies of apologies, explanations, and promises were often associated with relatively clear trust violations. For example, Robinette et al. [64] had a robot violate trust by leading an individual to the wrong location, whereas Reig et al. [42] had users observe the robot physically make a mistake by dropping or misplacing a package. In both cases, the relative simplicity of the tasks and the clear association of the violation with the robot could have influenced trust repair. Therefore, future researchers might wish to account for the degree of confidence humans have in their interpretation of events prior to the violation and/or manipulate the relative clarity of events to determine whether this is influential, especially in the case of denials.

Fourth, studies examining denials have not considered timing. Given that apologies, explanations, and promises each have been shown to be influenced by the timing, it could be that timing is important to denials as well. In particular, it is possible that denials become more effective after time has passed between a violation and a repair strategy. For example, denials given well after the initial trust violation could be more effective because details regarding the trust violation have been forgotten or conflated with other events. This could create doubt and thus afford denials more material to work with in crafting an alternative narrative of events, as suggested in the human–human literature [22], [24]. Regardless, this moderator seems under-examined in the case of denials, and given its prominence related to apologies,

promises, and explanations future researchers might wish to examine how timing impacts denials as well.

Finally, there is an overall lack of theoretical discussion across the HRI trust repair literature, with few studies suggesting or examining various theoretical mechanisms that might be at play. While there is one exception to this trend [65], more work is clearly needed. In particular, studies could reproduce the results of [65] by examining the attribution mechanisms of trust repair [68], [35] in light of moderators. Alternatively, studies might benefit from considering other mechanisms such as the relational and structural mechanisms of trust repair proposed in the human–human literature [23]. Presently the literature is mostly devoid of an overarching trust repair model or theoretical contributions stemming from existing work. This limits the overall generalizability and explanatory strength of what findings do exist in the HRI trust repair domain, and future work is needed to build on these findings and to construct models, frameworks, or theories relevant to the field.

VI. Conclusion

Trust is vital for effective human–robot collaborations. Trust changes over time and can be reduced after robot failures. This trust, however, can be restored through trust repair strategies. In particular, apologies, denials, explanations, and promises have been shown to restore trust in the human–human literature [22], but much remains unknown regarding trust repair in HRI. To this end, our paper contributes by first summarizing what has been studied in HRI in relation to trust repair and, second, by providing a list of future research considerations based on gaps identified across this literature. In particular, studies should continue to examine trust repairs but incorporate moderators such as violation type, severity, doubt, and timing. Furthermore, there is an apparent need for more theoretical work incorporating these moderators into prospective theoretical models.

References

[1] S. You, J.-H. Kim, S. Lee, V. Kamat, and L. P. Robert Jr, "Enhancing perceived safety in human–robot collaborative construction using immersive virtual environments," *Automation in Construction*, vol. 96, pp. 161–170, 2018.

[2] L. P. Robert Jr, A. R. Dennis, and M. K. Ahuja, "Differences are different: Examining the effects of communication media on the impacts of racial and gender diversity in decision-making teams," *Information Systems Research*, vol. 29, no. 3, pp. 525–545, 2018.

[3] C. Esterwood and L. P. Robert, "Human robot team design," in *Proceedings of the 8th International Conference on Human-Agent Interaction*, 2020, pp. 251–253.

[4] S. You, L. Robert, *et al.*, "Subgroup formation in human-robot teams: A multi-study mixed method approach with implications for theory and practice," *Journal of the Association for Information Science and Technology*, 2022.

[5] J. K. Barfield, "Self-disclosure of personal information, robot appearance, and robot trustworthiness," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 67–72.

[6] J. Xu and A. Howard, "Would you take advice from a robot? Developing a framework for inferring human-robot trust in time-sensitive scenarios," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 814–820.

[7] J. Jargon and E. Morath, "Short of workers, fast-food restaurants turn to robots," *The Wall Street Journal*, 06 2018. [Online]. Available: https://www.wsj.com/articles/short-of-workers-fast-food-restaurants-turn-to-robots-1529868693

[8] C. Mims, "Amid the Labor Shortage, Robots Step in to Make the French Fries," *The Wall Street Journal*, 08 2021. [Online]. Available: https://www.wsj.com/articles/restaurant-robots-kitchen-labor-shortage-11628290623

[9] J. Morrissey, "Desperate for Workers, Restaurants Turn to Robots," *The New York Times*, 10 2021. [Online]. Available: https://www.nytimes.com/2021/10/19/business/restaurants-robots-workers.html

[10] PYMNTS.com, "Uk labor shortage compromises christmas dinner," 08 2021. [Online]. Available: https://www.pymnts.com/news/retail/2021/grocery-roundup-uk-labor-shortage-compromises-christmas-dinner/

[11] G. Marks, "Employees are not showing up to work — employers are replacing them with robots," 10 2021. [Online]. Available: https://thehill.com/opinion/technology/578484-employees-are-not-showing-up-to-work-employers-are-\replacing-them-with/

[12] S. You and L. Robert, "Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction," in *Proceedings of the 52th Hawaii International Conference on System Sciences, Jan*, 2018, pp. 8–11.

[13] M. Kirtay, E. Oztop, M. Asada, and V. V. Hafner, "Trust me! I am a robot: An affective computational account of scaffolding in robot-robot interaction," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021.

[14] V. Surendran, K. Mokhtari, and A. R. Wagner, "Your robot is watching 2: Using emotion features to predict the intent to deceive," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 447–453.

[15] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "How social robots influence people's trust in critical situations," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 1020–1025.

[16] 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)*. IEEE, 2021, pp. 183–188.

[17] ——, "Having the right attitude: How attitude impacts trust repair in human-robot interaction," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2022)*. ACM/IEEE, 2022.

[18] G. Hannibal, A. Weiss, and V. Charisi, "The robot may not notice my discomfort"–Examining the Experience of Vulnerability for Trust in Human-Robot Interaction," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 704–711.

[19] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, "Toward an understanding of trust repair in human-robot interaction: Current research and future directions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 4, pp. 1–30, 2018.

[20] M. Lewis, K. Sycara, and P. Walker, "The role of trust in human-robot interaction," in *Foundations of Trusted Autonomy*. Springer, Cham, 2018, pp. 135–159.

[21] H. Azevedo-Sa, X. J. Yang, L. P. Robert, and D. M. Tilbury, "A unified bi-directional model for natural and artificial trust in human–robot collaboration," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5913–5920, 2021.

[22] R. J. Lewicki and C. Brinsfield, "Trust repair," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 4, pp. 287–313, 2017.

[23] N. Gillespie, S. Lockey, M. Hornsey, and T. Okimoto, "Trust repair: A multilevel framework," in *Understanding Trust in Organizations*. Routledge, 2021, pp. 143–176.

[24] P. Kim, D. Ferrin, C. Cooper, and K. Dirks, "Removing the shadow of suspicion: The effects of apology versus denial for repairing competence-versus integrity-based trust violations." *Journal of Applied Psychology*, vol. 89, no. 1, p. 104, 2004.

[25] R. J. Lewicki, B. B. Bunker, *et al.*, "Developing and maintaining trust in work relationships," *Trust in organizations: Frontiers of Theory and Research*, vol. 114, p. 139, 1996.

[26] S. L. Grover, M. C. Hasel, C. Manville, and C. Serrano-Archimi, "Follower reactions to leader trust violations: A grounded theory

of violation types, likelihood of recovery, and recovery process," *European Management Journal*, vol. 32, no. 5, p. 689–702, 2014.

[27] 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)*. IEEE, 2019, pp. 57–65.

[28] A. Costa, D. Ferrin, and C. Fulmer, "Trust at work," *The Sage Handbook of Industrial, Work & Organizational Psychology*, pp. 435–467, 2018.

[29] K. T. Dirks and D. P. Skarlicki, "The relationship between being perceived as trustworthy by coworkers and individual performance," *Journal of Management*, vol. 35, no. 1, pp. 136–157, 2009.

[30] R. M. Kramer and R. J. Lewicki, "Repairing and enhancing trust: Approaches to reducing organizational trust deficits," *Academy of Management annals*, vol. 4, no. 1, pp. 245–277, 2010.

[31] M. E. Schweitzer, J. C. Hershey, and E. T. Bradlow, "Promises and lies: Restoring violated trust," *Organizational behavior and human decision processes*, vol. 101, no. 1, pp. 1–19, 2006.

[32] L. Dai and Y. Wu, "Trust maintenance and trust repair," *Psychology*, vol. 06, no. 06, p. 767–772, 2015.

[33] R. J. Bies, "The predicament of injustice: The management of moral outrage." *Research in Organizational Behavior*, 1987.

[34] M. J. Cody and M. L. McLaughlin, "Interpersonal accounting," *Handbook of language and social psychology*, pp. 227–255, 1990.

[35] E. C. Tomlinson and R. C. Mayer, "The role of causal attribution dimensions in trust repair," *Academy of Management Review*, vol. 34, no. 1, pp. 85–104, 2009.

[36] V. R. Waldron, "Encyclopedia of human relationships," in *Apologies*, 1st ed., ser. 1, H. T. Reis and S. Sprecher, Eds. Thousand Oaks, CA: Sage Publishing Inc., 2009, vol. 3, ch. Apologies, pp. 98–100.

[37] R. J. Bies and D. L. Shapiro, "Interactional fairness judgments: The influence of causal accounts," *Social Justice Research*, vol. 1, no. 2, pp. 199–218, 1987.

[38] N. Du, J. Haspiel, Q. Zhang, D. Tilbury, A. K. Pradhan, X. J. Yang, and L. P. Robert Jr, "Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload," *Transportation Research Part C: Emerging Technologies*, vol. 104, pp. 428–442, 2019.

[39] N. Isaeva, K. Gruenewald, and M. N. Saunders, "Trust theory and customer services research: theoretical review and synthesis," *The Service Industries Journal*, vol. 40, no. 15-16, pp. 1031–1063, 2020.

[40] B. Bozic, "Consumer trust repair: A critical literature review," *European Management Journal*, vol. 35, no. 4, pp. 538–547, 2017.

[41] A. S. Mattila, "How to handle PR disasters? An examination of the impact of communication response type and failure attributions on consumer perceptions," *Journal of Services Marketing*, vol. 23, pp. 211–218, 2009.

[42] S. Reig, E. J. Carter, T. Fong, J. Forlizzi, and A. Steinfeld, *Flailing, hailing, prevailing*. ACM/IEEE, 2021.

[43] S. C. Kohn, D. Quinn, R. Pak, E. J. de Visser, and T. H. Shaw, "Trust repair strategies with self-driving vehicles: An exploratory study," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 62, no. 1. Sage Publications Sage CA: Los Angeles, CA, 2018, pp. 1108–1112.

[44] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 203–210.

[45] S. Engelhardt and E. Hansson, "A comparison of three robot recovery strategies to minimize the negative impact of failure in social hri," Thesis, KTH Royal Institute of Technology, 2017.

[46] Y. Albayram, T. Jensen, M. M. H. Khan, M. A. A. Fahim, R. Buck, and E. Coman, "Investigating the effects of (empty) promises on human-automation interaction and trust repair," in *Proceedings of the 8th International Conference on Human-Agent Interaction*, 2020, pp. 6–14.

[47] R. Perkins, Z. R. Khavas, and P. Robinette, "Trust calibration and trust respect: A method for building team cohesion in human robot teams," *arXiv preprint arXiv:2110.06809*, 2021.

[48] 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.

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

[50] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. De Vries, "Trust repair in human-agent teams: the effectiveness of explanations and expressing regret," *Autonomous Agents and Multi-Agent Systems*, vol. 35, no. 2, 2021.

[51] S. S. Sebo, M. Traeger, M. Jung, and B. Scassellati, "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 178–186.

[52] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, *Is It My Looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams*. Springer International Publishing, 2018, pp. 56–69.

[53] M. Heerink, B. Krose, V. Evers, and B. Wielinga, "Measuring acceptance of an assistive social robot: A suggested toolkit," in *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2009, pp. 528–533.

[54] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MISQ*, pp. 425–478, 2003.

[55] G. M. Alarcon, A. M. Gibson, and S. A. Jessup, "Trust repair in performance, process, and purpose factors of human-robot trust," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, 2020, pp. 1–6.

[56] D. Cameron, S. de Saille, E. C. Collins, J. M. Aitken, H. Cheung, A. Chua, E. J. Loh, and J. Law, "The effect of social-cognitive recovery strategies on likability, capability and trust in social robots," *Computers in Human Behavior*, vol. 114, pp. 106–561, 2021.

[57] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, "Exploring the impact of fault justification in human-robot trust," in *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, 2018, pp. 507–513.

[58] S. C. Kohn, A. Momen, E. Wiese, Y.-C. Lee, and T. H. Shaw, "The consequences of purposefulness and human-likeness on trust repair attempts made by self-driving vehicles," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 222–226, 2019.

[59] R. Liu, Z. Cai, M. Lewis, J. Lyons, and K. Sycara, "Trust repair in human-swarm teams+," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–6.

[60] R. Luo, C. Huang, Y. Peng, B. Song, and R. Liu, "Repairing human trust by promptly correcting robot mistakes with an attention transfer model," *arXiv preprint arXiv:2103.08025*, 2021.

[61] S. Mishler, "Whose drive is it anyway? using multiple sequential drives to establish patterns of learned trust, error cost, and non-active trust repair while considering daytime and nighttime differences as a proxy for difficulty," Ph.D. dissertation, Old Dominion University, 2019.

[62] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 33–42.

[63] M. Nayyar and A. R. Wagner, *When Should a Robot Apologize? Understanding How Timing Affects Human-Robot Trust Repair*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2018, book section Chapter 26, pp. 265–274.

[64] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in *International conference on social robotics*. Springer, 2015, pp. 574–583.

[65] X. Zhang, ""sorry, it was my fault": Repairing trust in human-robot interactions," Thesis, University of Oklahoma, 2021.

[66] D. B. Quinn, "Exploring the efficacy of social trust repair in human-automation interactions," Thesis, Clemson University, 2018.

[67] K. Hald, K. Weitz, E. André, and M. Rehm, ""an error occurred!"-trust repair with virtual robot using levels of mistake explanation," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, 2021, pp. 218–226.

[68] E. C. Tomlinson, B. R. Dineen, and R. J. Lewicki, "The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise," *Journal of management*, vol. 30, no. 2, pp. 165–187, 2004.