Do You Still Trust Me? Human-Robot Trust Repair Strategies

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

Abstract— **Trust is vital to promoting human and robot collaboration, but like human teammates, robots make mistakes that undermine trust. As a result, a human's perception of his or her robot teammate's trustworthiness can dramatically decrease [1], [2], [3], [4]. Trustworthiness consists of three distinct dimensions: ability (i.e. competency), benevolence (i.e. concern for the trustor) and integrity (i.e. honesty) [5], [6]. Taken together, decreases in trustworthiness decreases trust in the robot [7]. To address this, we conducted a 2 (high vs. low anthropomorphism) x 4 (trust repair strategies) between-subjects experiment. Preliminary results of the first 164 participants (between 19 and 24 per cell) highlight which repair strategies are effective relative to ability, integrity and benevolence and the robot's anthropomorphism. Overall, this paper contributes to the HRI trust repair literature.**

I. INTRODUCTION

Humans and robots are increasingly finding themselves in collaborative work arrangements [8], [9], [10], [11], [12]. Trust is vital to promoting human and robot collaboration, but like human teammates robots make mistakes that undermine trust. As a result, a human's perception of a robot's trustworthiness can dramatically decrease [1], [2], [3], [13], [14], [15]. Trustworthiness is the degree to which a trustee is perceived as worthy of an individual's trust. In this sense, trustworthiness precedes and largely determines the trust a trustor places in a trustee [5], [16]. Trustworthiness consists of three distinct elements: ability (i.e. competency), benevolence (i.e. concern for the trustor), and integrity (i.e. honesty) [5], [6]. Taken together, decreases in trustworthiness decreases in robot trust [7].

Fortunately, trust in HRI can be repaired via different trust repair strategies [17], [1], [18], [19], [20]. These strategies include apologies, denials, explanations and promises. Despite prior research on trust repair strategies in HRI, critical questions remain regarding the effectiveness of strategies. It is also not clear whether such repair strategies would be more or less effective at repairing specific elements of trustworthiness. Answering this question is important because each element's impacts can vary greatly in a given context [21], [22], [23]. As a result, it is important to know which strategy is most effective for repairing which trustworthiness element.

In this paper we provide the preliminary results of an ongoing study on the effectiveness of repair strategies with the goal of informing further design and deployment of trust repair strategies in HRI. To this end, we conducted a 2 (high vs. low anthropomorphism) x 4 (trust repair strategies) between-subjects experiment with 164 participants. Results highlight which repair strategies are more or less effective relative to ability, integrity and benevolence and whether this varies by the robot's anthropomorphism. Overall, this paper contributes to the HRI literature on trust and trust repair.

II. BACKGROUND

A. Trustworthiness and HRI

Trustworthiness is the degree to which a trustee is perceived as worthy of an individual's trust. Trustworthiness precedes and largely determines the trust a trustor places in a trustee [5], [24]. Trustworthiness is comprised of ability, benevolence and integrity [5], [6]. Ability is the skills or competencies that individuals have at their disposal [5]. Benevolence reflects the degree of concern the trustee has for the trustor over and above any egocentric motives [5, pp.718]. Finally, integrity is "the degree to which the trustee is honest and consistently follows a set of principles" [21, Pg.2].

Studies examining trustworthiness in HRI have identified unique relationships for each of these elements. For example, [25] examined the impacts that a robot's reliability and social intent had on ability, benevolence and integrity. This study found reliability to significantly impact ability and integrity but not benevolence, while social intent influenced integrity and benevolence but not ability. In addition, [26] examined a robot's perceived ability, benevolence and integrity between different human participant's gender identities. They found participants' gender significantly impacted their perceptions of ability and benevolence but not integrity. These effects, however, may have been non-significant because of the speed and transience of the interactions depicted in the videos shown to participants, thus making observable indications of integrity less salient. Finally, [21] also examined ability, benevolence and integrity in an HRI context. Their findings indicated that ability and integrity were significantly related to perceived robot intelligence whereas benevolence was not. In sum, recent work has found ability, benevolence and integrity to be significantly impacted in a range of unique ways. It is therefore likely that these elements will also be impacted in varying ways by different repair strategies.

B. Trust Repair and HRI

Trust repair can be defined as the efforts undertaken by the trustee to restore trust following an actual or perceived trust violation [27], [28], [29]. The human-human literature has

[1]Connor Esterwood is a 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]Lionel P. Robert is an 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

Esterwood, C. and Robert, L. P. (2021). **Do You Still Trust Me? Human-Robot Trust Repair Strategies,** *Proceedings of 30th IEEE International Conference on Robot and Human Interactive Communication* (**RO-MAN 2021**), Aug 08 - Aug 12, 2021. Virtual Conference (originally in Vancouver, BC, Canada).

classified these efforts as either apologies, denials, explanations, or promises that have each been found to be effective methods of trust repair [30], [31], [32]. Apologies are a type of verbal trust repair strategy that seeks to express remorse for a relational or social transgression [33]; denials are rejections of culpability typically accompanied by external reasons as to why a violation of trust was committed [17]; explanations are explicit verbal statements with the goal of providing the reason(s) why an action has occurred [34]; and promises are assertions by the trustee designed to convey positive intentions about future acts [35].

The HRI literature provides support for the efficacy of apologies, denials, and promises [18], [19], [36]. For example, [19] and [18] both found apologies and denials to be effective at repairing trust in HRI. However, apologies were more effective for violations of ability and denials were more effective for violations of integrity [18], [19]. Additionally, [36] examined the impact of timing on trust repair as well as the efficacy of promises when compared to apologies and denials. Results indicated that timing was influential for trust repair and that promises were largely more effective than apologies or denials.

C. Trustworthiness and Anthropomorphism

Anthropomorphism can be defined as "the degree to which an entity is perceived as human-like" [37, P.34] and has been found to significantly impact humans' trust and trustworthiness in robots [38], [21], [39], [40], [41], [42], [43]. In short, humans tend to trust robots high in anthropomorphism over those low in anthropomorphism. For instance, [38] found perceived anthropomorphism of an agent to have a positive relationship with trust in that agent. Similarly, [39] examined anthropomorphism and trust in humanoid service robots. Consistent with [38], they also found a positive relationship between anthropomorphism and trust. Given this, it is possible that anthropomorphism could also be important to determining the effectiveness of a given repair strategy. Despite this, we know little about if or how anthropomorphism might influence the efficacy of trust repair strategies.

III. METHOD

To test our hypotheses, we conducted an online experiment using a high-fidelity simulated human-robot interaction task. This experiment varied robot anthropomorphism (high/anthropomorphic vs. low/mechanoid) and the trust repair strategy (apology, denial, explanation or promise) deployed after a trust violation.

A. Participants

For this paper, given the preliminary nature of this study, we only examined 164 participants all from Amazon Mechanical Turk. However, given the preliminary nature of this study, we examined the first 164 participants. These subjects were 72% male and ranged in age between 22 years and 68 years old with an average age of 39. Participants were compensated at a rate of $15.00/hr or more for their participation in this study, which lasted approximately 15-25 min in total. This research complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at the University of Michigan. Informed consents were gathered upon the acceptance of the "HIT" by participants on M-Turk.

B. Task

Participants were required to work as a member of a heterogeneous human-robot team. The team was tasked with loading a specific set of boxes onto a conveyor belt. The team comprised one robot and one human. The human's role was that of quality assurance and the robot's was that of picker. Ten boxes were picked up by the robot and presented to the human who would review the number on the box and either approve or reject it. If the human approved the box, the robot would place it on the conveyor belt and if the human rejected it, the robot would return the box to a nearby pile of boxes. When all 10 boxes were processed (approved or rejected) the task ended. Over the 10 boxes, the robot was programmed to make three mistakes by incorrectly presenting the human with a wrong box three times. A 70% reliability rate was chosen based on [44] which found that automation only increases performance when its reliability is greater than 67%. For consistency, the robot's reliability did not vary by condition and the robot made the same mistake with the same set of boxes for each participant. In other words, the mistake type and the timing of the mistake was constant across the experiment. The experiment only varied the repair strategy and anthropomorphism.

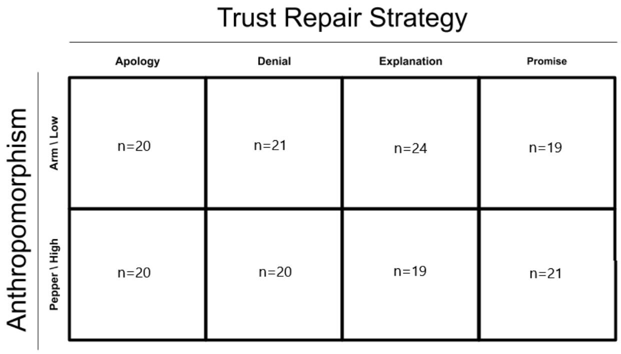
C. Apparatus

To accomplish their assigned task, participants were presented with an online based interactive scenario developed in the UnReal Engine version 4.24. This scenario represented a factory environment and the participant could look around the environment at will but was stationary in one location [see figure 2]. Two screens were placed on a tabletop and displayed the correct serial number, the time it took to approve or reject a box and participant's total score based on points gained for a correct box (+1) or lost for an incorrect box (-1). Points were neither given nor subtracted in cases where the human correctly accepted or rejected the robot's box. Points did not reflect participants pay and were present merely as a means of encouraging completion and attention.

D. Experimental Design

This study employed a between-subjects experimental design with eight conditions. These conditions varied by repair strategy and by degree of anthropomorphism. Consistent with this design, each participant encountered one of four repair strategies (apology, denial, explanation, or promise) and one of two anthropomorphic conditions (high/anthropomorphic vs. low/mechanoid). Figure 1 represents the 2X4 experimental design utilized in this study.

The trust repair strategies in this study were apologies, denials, explanations and promises. Each of these was applied three times after each trust violation and was consistent



Trust Repair Strategy

Anthropomorphism

	Apology	Denial	Explanation	Promise
Arm \ Low	n=20	n=21	n=24	n=19
Pepper \ High	n=20	n=20	n=19	n=21

Fig. 1. 2x4 Visual Representation of Experimental Conditions With n For Each Condition.

within experimental groups. In this study we wanted to examine the impact of repeated trust violations. It is unrealistic to believe that a robot would be 100% reliable, and one mistake could easily be overlooked or even missed by a human. Our assumption was that less than perfect reliability would result in multiple mistakes not just one.

In the apology condition, the robot stated, *"I'm sorry I got the wrong box that time."* In the denial condition, the robot stated, *"I picked the correct box that time so something else must have gone wrong."* In the explanation condition, the robot stated, *"I see, that was the wrong serial number."* In the promise condition, the robot stated *"I'll do better next time and get the right box."* These messages were conveyed to the participants via audio and subtitle text.

In addition to varying trust repair strategies, we varied the degree of anthropomorphism of the robot. In the high anthropomorphism condition, we utilized an anthropomorphic appearing robot (Pepper-like robot) and in the low anthropomorphism condition, we utilized a mechanoid appearing robot (generic robotic arm). In selecting suitable robots we used the anthropomorphic robot database [45]. Pepper (anthropomorphic) had an overall human likeness score of 42.17 so we expect our pepper-like robot to be similarly anthropomorphic. For the generic robot arm, a similar robotic arm (UR-3) manipulator to the one used in this study had a score of 6.08. Both robots appeared suitable for the task and environment given their capacity to act as pickers and porters of boxes. Both robot's movements were consistent and designed to be realistic with their respective form factors Images of both robots are presented in figure 2

E. Dependent Variables

The dependent variables in this study were the participant's ratings of the robot's ability, benevolence, and integrity. These were measured via a set of nine items adapted from [46], [47] and [48]. This measure provided reliabilities of $\alpha = 0.72$, $\alpha = 0.94$, and $\alpha = 0.93$ for ability, benevolence and integrity, respectively. The measure was deployed as part of the post-test questionnaire.

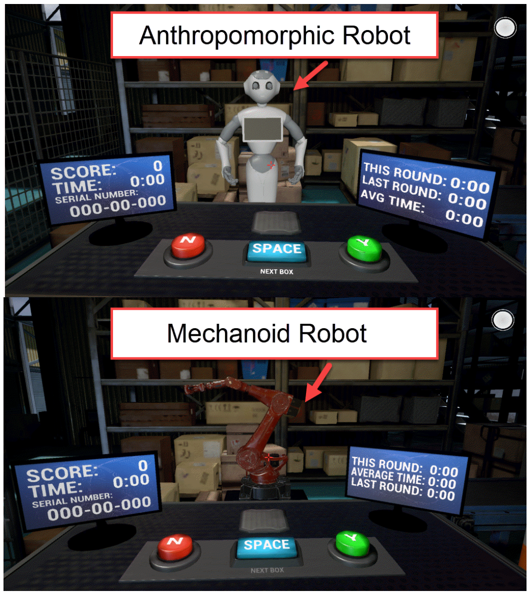


Fig. 2. Robotic Forms and Environment Used in Experiment.

F. Procedure

Upon accepting the "HIT" (task) in Amazon Mechanical Turk, participants were first directed to a training scenario. In this scenario, participants were introduced to the virtual environment and the web-based interface. They were then guided through the task with a generic mannequin. During this training scenario participants were presented with one incorrect box (mismatching serial numbers) and one correct box (matching serial numbers) and were instructed how to respond appropriately by modal dialogues.

After completing the training scenario, participants completed a pretest questionnaire measuring basic demographic information and were assigned to an experimental condition. Trust violations took the form of the robot picking the wrong box and presenting it to the participant. Trust violations occurred after every third box. Upon completion of all 10 box tasks, participants were presented with the post test questionnaire and their their payment code. WE used attention check questions to ensure data integrity and these were randomly distributed throughout our questionnaires. If at any point the participants failed one of the attention check questions, their participation was terminated, data deleted, and no payment was given. A total of 37 participants were excluded in this manner.

IV. Results

This study was part of an ongoing larger study investigating HRI trust repair. At the time of this report, the study had an average of 19 participants per experimental condition. Given the relatively small sample size, the paper's goal was to provide preliminary results on the ongoing study. As such, we are interested in highlighting statistically significant findings and identifying trends that might inform future research. Next we elaborate on the findings in detail.

A. Main Effects: Repair Strategies

As visible in table I, the type of repair strategy has a statistically significant impact on a robot's integrity and benevolence. In addition, the results of a post-hoc investigation of these significant main effects and an investigation of means (visible in figure 3) provide additional insights into these relationships. Specifically, this study found that independent of anthropomorphism and after several trust violations, explanations were more effective at repairing integrity than apologies, denials or promises. Notably, there was a statistically significant difference between explanations and denials ($p < 0.005$); see Figure 3. Further, explanations appeared less effective at repairing benevolence than promises or apologies although this difference failed to reach statistical significance.

Generally, promises led to lower integrity than explanations, but statistically higher integrity than apologies ($p = 0.025$). Promises also resulted in the highest benevolence, followed by apologies, explanations and denials. Notably, the differences in benevolence between explanations and promises ($p = 0.028$) as well as denials and promises ($p < 0.005$) even reached statistical significance. Last, apologies resulted in higher ability, benevolence, and integrity than denials but, these effects were only significant for benevolence ($p < 0.005$). Taken together, these results provide support for the idea that different trust repair strategies have different impacts on ability, benevolence and integrity.

Outcome	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Ability	3	8.2	2.71	1.80	0.14
Integrity	**3**	**31.9**	**10.62**	**4.54**	**0.03***
Benevolence	**3**	**44.5**	**14.84**	**5.47**	**0.001****

TABLE I

MAIN EFFECT OF REPAIR STRATEGY IN PREDICTING ABILITY, BENEVOLENCE, AND INTEGRITY.

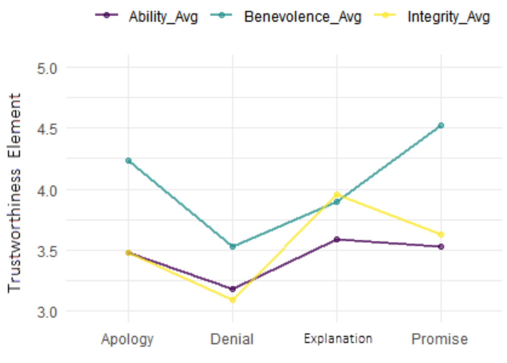


Fig. 3. Slope chart shows ability, integrity and benevolence by repair condition.

B. Interaction: Anthropomorphism and Repair Strategy

When examining interaction effects between anthropomorphism and repair strategy, there was a significant interaction effect for integrity (F(3, 156) = 3.65, p = 0.014; Eta^2 (partial) = 0.07, 90% CI [0.008, 0.13]). After a post-hoc examination, several non-significant trends emerged. As shown in figure 4, integrity was the highest for explanations, but only if those explanations were given by an anthropomorphic robot. In cases where explanations were given by a mechanoid robot, promises led to the highest integrity. Notably, when promises were given by an anthropomorphic robot they led to the lowest integrity of all the strategies. Apologies and denials in this case were fairly consistent with only slightly higher integrity when provided by a mechanoid as opposed to an anthropomorphic robot. Taken as a whole, these trends suggests that promises and explanations have different impacts based on a robot's anthropomorphism but not apologies and denials.

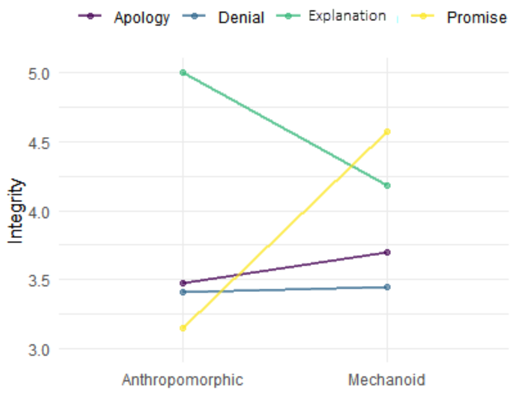


Fig. 4. Interaction plot shows integrity by anthropomorphic condition for each repair strategy.

An interaction effect between repair strategy and anthropomorphism was also observed for benevolence. This effect, however, was only marginally significant (F(3, 156) = 2.24, p = 0.086; Eta^2 (partial) = 0.04, 90% CI [0.00, 0.09]). An inspection of means uncovers several useful trends that are similar to those observed for integrity. As visible in figure 5, benevolence were highest for explanations but only if those explanations were given by an anthropomorphic robot. In cases where explanations were given by a mechanoid robot, promises led to the highest benevolence. Diverging from previous results, benevolence was higher for anthropomorphic robots in cases of apologies and denials. Taken as a whole, these trends also appear to indicate that promises, explanations and denials function differently based on a robot's anthropomorphism but not apologies.

V. DISCUSSION

The goal of this paper was to provide the preliminary results of an ongoing study on the effectiveness trust repair strategies relative to the robot's anthropomorphism. This study's results provide preliminary evidence to suggest that repair strategies can have distinct impacts on a robot's integrity and benevolence. Specifically, this study found that independent of anthropomorphism, explanations lead

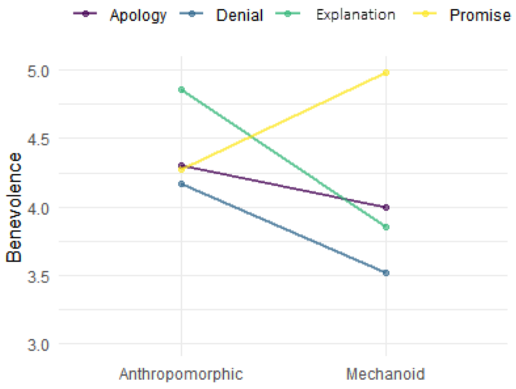


Fig. 5. Interaction plot shows benevolence by anthropomorphism condition per repair strategy.

to higher perceptions of a robot's integrity than apologies, denials, or promises. However, the impacts of denial and apologies were not as apparent. Contributions, implications and study limitations are detailed next.

This study goes beyond the existing literature by examining explanations as a repair strategy. Existing literature has either examined apologies and denials with regards to trustworthiness [18], [19] or promises with regards to trust [36]. Explanations as a repair strategy seemed to yield the best results when compared to apologies, denials and promises on integrity. This study unlike the previous literature exposed participants to repeated trust violations. That being said, our findings regarding apologies and denials were consistent with the existing literature. Specifically, apologies produced higher integrity and benevolence than denials [18], [19]. Similarity, our findings on promises were consistent with the literature in that promises outperformed apologies and denials with regards to benevolence and integrity [see: [36]].

Results showed that ability, benevolence and integrity are each impacted differently by a particular repair strategy. Namely, none of the repair strategies was more or less effective at promoting ability in the presence of repeated trust violations. However, results showed that repair strategies do differ in their effectiveness with regard to integrity and benevolence. This might imply that apologies, denials, explanations and promises impact trust not through repairing perceptions of ability but instead through repairing perceptions of integrity and benevolence. It is also notable that benevolence scores were higher than ability and integrity across all repair strategies except for explanations on integrity. This might suggest that repair strategies, other than explanations, work mainly via benevolence to repair trust. Ultimately, more research is needed with larger sample sizes to explore these trends and observations.

This study also contributes to the literature on anthropomorphism and robot trust by identifying the role of anthropomorphism in determining the effectiveness of a given repair strategy. Anthropomorphism has been shown to impact trust in robots, with anthropomorphic robots engendering more trust than mechanoid robots [38], [21], [39], [40], [41], [42]. Our results broadly agree with these findings for integrity when explanations are given and for benevolence when apologies, denials, and explanations are given, but our findings diverge from this in other cases. Specifically, apologies, denials and promises given by an anthropomorphic robot resulted in lower perceptions of integrity than when given by a mechanoid robot which led to lower trustworthiness. This trend is also visible for benevolence when promises are used. Together these results paint a complex picture and provide evidence to suggest that anthropomorphism's influence on the effectiveness of a given repair strategy may not be linear.

One limitation of this study is its reliance on the virtual representation of physical robots. This approach offers more flexibility with regards to changing physical attributes. However, it is still possible that these virtual representations may have weakened the impact of anthropomorphism. Future research could be done to replicate our findings with physical robots in a real world setting.

ACKNOWLEDGMENT

This study would like to thank the Emerging Technology Group at the University of Michigan's James and Anne Duderstadt Center. In particular we wish to thank Sara Eskandari and Stephanie O'Malley for their help in the development of our experimental platform.

REFERENCES

[1] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 1–8.

[2] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "Human perceptions of the severity of domestic robot errors," in *Social Robotics*, A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, Eds. Cham: Springer International Publishing, 2017, pp. 647–656.

[3] W. Mou, M. Ruocco, D. Zanatto, and A. Cangelosi, "When would you trust a robot? a study on trust and theory of mind in human-robot interactions," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 956–962.

[4] S. You and L. Robert, "Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[5] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[6] L. P. Robert, A. R. Denis, and Y.-T. C. Hung, "Individual swift trust and knowledge-based trust in face-to-face and virtual team members," *Journal of Management Information Systems*, vol. 26, no. 2, pp. 241–279, 2009.

[7] S. You and L. P. Robert Jr, "Human-robot similarity and willingness to work with a robotic co-worker," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 251–260.

[8] C. Esterwood and L. Robert, "Robots and covid-19: Re-imagining human-robot collaborative work in terms of reducing risks to essential workers," *ROBONOMICS: The Journal of the Automated Economy*, vol. 1, pp. 9–9, 2021.

[9] C. Esterwood and L. P. Robert, "Personality in healthcare human robot interaction (h-hri) a literature review and brief critique," in *Proceedings of the 8th International Conference on Human-Agent Interaction*, 2020, pp. 87–95.

[10] C. Esterwood, K. Essenmacher, H. Yang, F. Zeng, and L. P. Robert, "A meta-analysis of human personality and robot acceptance in human-robot interaction," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–18.

[11] 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.

[12] J. Xu and A. Howard, "The impact of first impressions on human-robot trust during problem-solving scenarios," in *2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2018, pp. 435–441.

[13] 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 *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[14] H. Azevedo-Sa, X. J. Yang, L. P. Robert Jr, and D. M. Tilbury, "A unified bi-directional model for natural and artificial trust in human-robot collaboration," *arXiv preprint arXiv:2106.02194*, 2021.

[15] 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 *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[16] L. Robert and S. You, "Are you satisfied yet? shared leadership, trust and individual satisfaction in virtual teams," in *Proceedings of the iConference*, 2013.

[17] 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.

[18] 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.

[19] D. B. Quinn, "Exploring the efficacy of social trust repair in human-automation interactions," Master's thesis, Clemson University, 5 2018.

[20] E. J. De Visser, R. Pak, and T. H. Shaw, "From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction," *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, 2018.

[21] W. Kim, N. Kim, J. B. Lyons, and C. S. Nam, "Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach," *Applied ergonomics*, vol. 85, p. 103056, 2020.

[22] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[23] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 109–116.

[24] E. A. Sharp, R. Thwaites, A. Curtis, and J. Millar, "Trust and trustworthiness: conceptual distinctions and their implications for natural resources management," *Journal of Environmental Planning and Management*, vol. 56, no. 8, pp. 1246–1265, 2013.

[25] J. B. Lyons, T. Vo, K. T. Wynne, S. Mahoney, C. S. Nam, and D. Gallimore, "Trusting autonomous security robots: The role of reliability and stated social intent," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, p. 001872082090162, 2020.

[26] D. Gallimore, J. B. Lyons, T. Vo, S. Mahoney, and K. T. Wynne, "Trusting robocop: Gender-based effects on trust of an autonomous robot," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://dx.doi.org/10.3389/fpsyg.2019.00482

[27] A. Costa, D. Ferrin, and C. Fulmer, "Trust at work," *The sage handbook of industrial, work & organizational psychology*, pp. 435–467, 2018.

[28] 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.

[29] 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.

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

[31] P. H. Kim, D. L. Ferrin, C. D. Cooper, and K. T. 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] T. Jensen, M. M. H. Khan, and Y. Albayram, "The role of behavioral anthropomorphism in human-automation trust calibration," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 33–53.

[38] 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.

[39] M. M. van Pinxteren, R. W. Wetzels, J. Rüger, M. Pluymaekers, and M. Wetzels, "Trust in humanoid robots: implications for services marketing," *Journal of Services Marketing*, 2019.

[40] F. Eyssel, L. de Ruiter, D. Kuchenbrandt, S. Bobinger, and F. Hegel, "'if you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2012, pp. 125–126.

[41] M. B. Mathur and D. B. Reichling, "An uncanny game of trust: social trustworthiness of robots inferred from subtle anthropomorphic facial cues," in *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2009, pp. 313–314.

[42] E. Torta, E. van Dijk, P. A. Ruijten, and R. H. Cuijpers, "The ultimatum game as measurement tool for anthropomorphism in human–robot interaction," in *International Conference on Social Robotics*. Springer, 2013, pp. 209–217.

[43] T. Matsui and S. Yamada, "Robot's impression of appearance and their trustworthy and emotion richness," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 88–93.

[44] J. R. Rein, A. J. Masalonis, J. Messina, and B. Willems, "Meta-analysis of the effect of imperfect alert automation on system performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 57, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2013, pp. 280–284.

[45] E. Phillips, X. Zhao, D. Ullman, and B. F. Malle, "What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 105–113.

[46] D. H. Mcknight, M. Carter, J. B. Thatcher, and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Transactions on management information systems (TMIS)*, vol. 2, no. 2, pp. 1–25, 2011.

[47] J. Bernotat, F. Eyssel, and J. Sachse, "Shape it–the influence of robot body shape on gender perception in robots," in *International Conference on Social Robotics*. Springer, 2017, pp. 75–84.

[48] ——, "The (fe)male robot: how robot body shape impacts first impressions and trust towards robots," *International Journal of Social Robotics*, pp. 1–13, 2019.