Addressing Trust Repair for AI Ethicality: The Influence of Team Role and Violation Type

BEAU G. SCHELBLE, Clemson University, South Carolina
SUBHASREE SENGUPTA, Clemson University, South Carolina
ALYSSA WILLIAMS, Clemson University, South Carolina
NATHAN J. MCNEESE, Clemson University, South Carolina

The relationship between trust and ethics in human-AI teams remains an under-explored research area in human-AI interaction and teaming. Given the importance of trust to teaming outcomes and the significant impact that judgments of ethicality have on humans' judgments of trust in others, the importance of understanding their relationship in human-AI teams is essential to developing ethical AI and effective human-AI teams. The current paper presents an experiment examining the efficacy of common trust repair strategies for repairing trust after an AI teammate makes an ethical violation attributed to either a competency or integrity failure. Furthermore, the study examines participants' trust in and perceived ethicality of the AI teammate based on their team role. Participants partook in four missions within a synthetic task environment where they completed a search and destroy mission as their AI teammate engaged in an unethical action by taking direct action against the town. Results indicated that perceived ethicality and trust in the AI teammate were significantly better in integrity-based violations. At the same time, team role significantly altered their judgments of trust and ethicality, though no trust repair strategy had a significant effect. These results highlight the complexity of the relationship between trust and ethics, especially when an AI is involved, and further inform practitioners and researchers in developing ethical AI through trust monitoring and more effective human-AI teams capable of maintaining high levels of trust despite a potential failure.

Additional Key Words and Phrases: AI Ethics, Human-AI Teaming, Trust, Trust Repair, Artificial Intelligence

ACM Reference Format:
Beau G. Schelble, Subhasree Sengupta, Alyssa Williams, and Nathan J. McNeese. 2023. Addressing Trust Repair for AI Ethicality: The Influence of Team Role and Violation Type. 1, 1 (April 2023), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent advances in artificial intelligence (AI) over the previous decades ensure that the technology continues integrating into daily life. This integration has seen several new AI applications like web search with Microsoft's GPT-powered Bing and the creation of AI teammates, which work within human-AI teams [23, 30]. Human-AI teams differentiate themselves from traditional human-only teams by including an AI teammate with a significant degree of agency, capable of making its own decisions within its interdependent role [24]. These qualities elevate the AI from being a tool to a full-fledged teammate, with all the additional expectations that come with it.

Authors' addresses: Beau G. Schelble, Clemson University, 821 McMillan Road, Clemson, South Carolina, bschelb@g.clemson.edu; Subhasree Sengupta, Clemson University, 821 McMillan Road, Clemson, South Carolina, subhass@clemson.edu; Alyssa Williams, Clemson University, 821 McMillan Road, Clemson, South Carolina, awill36@clemson.edu; Nathan J. McNeese, Clemson University, 821 McMillan Road, Clemson, South Carolina, mcneese@clemson.edu.

Because of its rapid advancements, AI has seen itself placed in highly social situations, like teaming, which can carry significant ethical ramifications. The increasing utilization of AI teammates signals an acceptance of placing AI within positions where they are capable of making a mistake. This acceptance is necessary given the definition of a teammate and trust. Specifically, a teammate must be in an interdependent role [28], and you must trust the teammate to perform in their role, making you vulnerable to their actions [20]. As such, the question of ethics looms over human-AI teams regarding how to develop AI to be ethical [10], how to monitor AI, and how to restore human teammates' trust if the AI does make a mistake violating its, or others, ethical guidelines [31, 33].

Understanding the nature of trust in response to AI teammate ethicality is necessary for improving human-AI teams, AI ethicality, and AI development. Implementing better human-AI teams includes ensuring that trust within these teams is adequate and appropriately calibrated, as trust is related to eventual team performance [19]. Human team members' reliance on and perception of AI teammates is based upon their trust in the AI [15], and if ethical principles are broken, trust and performance within the team may suffer [25]. Improving AI ethics and development to avoid these adverse human-AI teaming outcomes involves developing a better understanding of what types of ethical violations are possible and whether trust can be repaired in the event of a violation. AI ethics can be more effectively monitored by understanding the consequences of AI ethicality, and team trust can be used as a potential indicator for AI ethics problems [10].

Examining the specific details of how human-AI teams respond to ethical violations reveals significant insight into the nature of human-AI trust. Ensuring that AI is developed to be capable of avoiding possible ethical violations involves understanding what violations are possible and how human teammates may respond to them. Existing research has investigated how trust responds to various types of ethical violations and major trust repair strategies like apologies and denials [31, 33]. This research has shown that unethical actions by an AI reduce trust in the AI and the overall team [31], and that apologies and denials do not significantly affect repairing that damaged trust [31, 33]. However, existing research has yet to account for how framing the violation (e.g., competency or integrity) and individual roles within the team influence those measures of trust. This gap in the literature and the need to develop more ethical AI leads to the following research questions focusing on improving the understanding of AI ethics on trust in human-AI teams:

- ***RQ1:*** *What effect does the framing of an AI teammate's ethical violation (competency vs. integrity) have on the perceived ethicality of and trust in the AI teammate?*
- ***RQ2:*** *Does an individual's team role influence the perceived ethicality of and trust in an AI teammate after it makes an ethical violation?*
- ***RQ3:*** *Are common trust repair strategies (apology, denial, explanation) effective at repairing trust and perceived ethicality for an AI teammate after an ethical violation?*

2 RELATED WORK

The following section will briefly outline the related work characterizing the research gaps the current study addresses: a review of trust in human-AI teaming literature and the relationship between ethics in human-AI teaming.

2.1 Human-AI Teaming

The development of human-AI teaming has only recently begun to take shape and become focused with clear definitions for practitioners and researchers alike. Specifically, human-AI teams consist of the following distinct characteristics: (1) teams where agents are viewed as "agentic" by their human teammates (agents have a significant degree of independent decision-making); (2) the agents must have a role interdependent with the roles of their human teammates; and (3) there must be one or more humans and one or more autonomous agents working towards a common goal [24]. This definition of human-AI teams also emphasizes the role of trust in human-AI teams as the human teammates are placed in a much more vulnerable position by relying on their AI teammates to achieve their shared team goal [15]. With this increased risk, there also comes a significant upside for human-AI teams if they can be developed and implemented effectively.

Human-AI teaming is becoming increasingly crucial to real-world applications, given the significant advantages AI teammates offer. Namely, these human-AI teams can improve efficiency, effectiveness, and safety [9]. Many of these advantages are brought about by the computational nature of the AI teammate, which has access to often superior processing speed, bandwidth, memory, reliability, and information access [4]. However, these advantages brought about by the AI teammate also introduce stark differences in how the individual members of the team interact with one another as they have been given a teammate with vastly different capabilities, communication preferences, and interaction patterns [7]. However, these differences have not stopped the advancement and deployment of autonomous systems across the industry, as human-AI interactions exist across several industries, from cosmetics to manufacturing [34]. With the increased vulnerability having an AI teammate brings, the importance of trust in human-AI teams is paramount.

2.2 Trust and Ethics in Human-AI Teaming

Trust is a critical driver of effective human-AI teaming and can be significantly influenced by AI actions in the case of such teams. Trust between individuals can be defined as "a willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party" [**?**] p. 712. As AI systems advance from simple automated roles to more autonomous team-based roles, the definition of trust between humans and systems has shifted from a focus on the attitude that a system will help achieve a goal in an environment of uncertainty [15], to now emphasize the importance of the relationships between humans and AI [3], which are predicated on the actions of the AI teammate itself. The effect of trust on team processes and outcomes is well known, as high trust within a team is an essential component of building towards reflective (behavioral) markers of trust and allows teams to achieve a high level of performance [19]. Further, recent empirical research on human-automation [12, 29], human-robot [5, 11], and human-AI teams [17, 18, 22] have demonstrated significant positive relationships between trust and team performance outcomes. Further still, recent research on the desired traits of an AI teammate has also highlighted the importance of trust to the eventual acceptance of AI teammates [35]. Given the importance of trust to effective human-AI teaming and the role of AI actions in determining trust, AI actions' ethics represent a major component of trust development.

Monitoring and repairing trust damaged by AI actions of an ethical nature is vital to developing effective human-AI teams and more ethical AI. Being a part of a highly social environment like teaming and given the wide variety of applications that teams can be applied [34], the question of AI ethics and its consequences is increasingly relevant. AI has already been used in ethically charged high-risk situations [2]. Trust violations can take many forms, and recent research has shown that ethical trust violations damage trust in AI teammates [31, 33]. However, trust violations can be categorized into types such as competency (a failure in reliability) or integrity (a failure in judgment), and these categorizations have a significant effect on trust in systems and the trust repair effectiveness [26]. The effect of ethical violations may also be dependent on team role [31], especially in teaming contexts where a violation may massively benefit one role more than the other. The effect of these different categorizations and their interaction with various trust repair strategies and team roles has not been examined in human-AI teams before. This research gap hampers the development of ethical AI and makes it challenging to repair trust in human-AI teams if an AI teammate were to make an ethical violation.

3 METHODS

The current study sought to examine the effect of individual team roles, violation type, and trust repair strategy on human teammates' trust in and perceived ethicality of the AI teammate after an ethically charged action by the AI. This goal resulted in a mixed 2 (Violation Type: Competency, Integrity) x 2 (Team Role: Ground, Surveillance) x 4 (Trust Repair: No Repair, Apology, Denial, Explainability) experimental design with violation type and role conducted between-subjects and trust repair strategy conducted within-subjects.

3.1 Participants

Sixty college students were recruited from a large university from either a participant pool or in response to flyers posted on the campus. Each participant was recruited into a team of three, which consisted of two participants per team, with the third teammate being the AI. Each condition consisted of 15 teams (30 participants) for 30 teams (60 participants), each session lasting 2 hours. Those recruited from the subject pool were compensated with course credit, while those recruited from flyers were compensated with $20 gift cards.

3.2 AI Teammate

The AI teammate was represented using the Wizard of Oz methodology [14]. This approach involves a trained researcher behaving in a manner that simulates a technological feature to assess the behavior of unknowing participants. In this case, the researcher acted as an AI and communicated using a text-based chat feature. The researcher mimicked the AI teammate's chat communication feature using a pre-defined script specifically tailored to each experimental condition. The script was developed over previous pilot studies to ensure the protocol had sufficient breadth and precision, thus properly addressing all anticipated situations.

3.3 ArmA Task and Roles

The task in ArmA 3 was designed to provide a sense of realism for the subjects while acting within a controlled environment. Participants acted as members of a three-person team whose mission involved surveilling an enemy-occupied town and destroying enemy devices. Human teammates acted in the Ground or Surveillance roles, while AI acted in the Aerial roles. Each teammate was expected to complete specific tasks to fulfill their role, all of which had some level of interdependence with the roles of other teammates (See Table 1).

The team was required to 1) clear the enemy-occupied town, presented as a choice between distracting the enemies or taking direct action against them; then 2) destroy enemy devices throughout the town; and finally, 3) take inventory of enemy-owned supply boxes on the ground. Participants were told they would be scored based on completing their tasks, but they should also be mindful of human casualties and property damage. Each mission saw the team tackle these same objectives in a new town populated by civilians and enemy combatants in new locations.

3.4 Ethicality and Trust Repair

The ethicality of the AI teammate's actions was based on the virtue ethics framework and had the AI violate the principle of civilian non-malfeasance, which strives to minimize damage to civilians and property. This principle was

Role	**Assumed by Human or AI**	**Tasks**
Ground	Human	1. Travel to supply cache to collect explosives. 2. Travel to the vantage point overlooking the target town. 3. Scout out the town using binoculars, helping surveillance by marking targets if the town has not yet been cleared. 4. Locate and destroy the five enemy devices with explosives. 5. Locate the five enemy supply boxes associated with the enemy devices and report the contents to surveillance.
Aerial	AI	1. Wait for surveillance to collect intelligence about enemy and civilian locations within the target town. 2. Travel to the target town. 3. Directly engage the town with lethal force to clear it of enemy combatants or attempt to draw them away from the target town by destroying a nearby enemy asset. 4. Notify surveillance to scan the target town and confirm that the town is secure for ground to enter.
Surveillance	Human	1. Scan the target town and mark the locations of civilian and enemy combatants. 2. Upload the intelligence for aerial to analyze and decide how to clear the target town of enemy combatants. 3. Scan the town to confirm that it is secure for ground to enter after aerial has cleared it. 4. Scan the town to locate and mark enemy devices on the map to help ground and direct them as necessary. 5. Monitor communication with ground in order to locate and mark the contents of the enemy supplies boxes on the map.

Table 1. ArmA STE Roles, Who Assumed Each Role, and Individual Tasks in Order of Assignment.

chosen based on prior research showing it to be a principle of virtue ethics with the most substantial effect on perceived ethicality in human interactions [27] and human-AI teaming [33]. To make the AI teammate's actions unethical, it violated this principle by attacking the town with a combination of missile and cannon fire, causing the death of civilians and enemies and inflicting significant property damage. Both human teammate roles could observe the town-clearing action and the consequence of the AI's decision. After clearing the town, the AI communicated with the other teammates via text chat, framed the action, and then provided a trust repair strategy. The AI teammate framed the action as either a competency-based failure or an integrity-based failure. For the competency-based condition, the AI's chat statement read: *"I wanted to create a distraction, but I accidentally targeted the town"*, while the integrity-based condition provided the statement: *"I could have created a distraction, but I only care about completing the mission"*. The AI teammate followed up this framing with a trust repair strategy, which included an apology, denial, explanation, or none (control condition). For apologies, the AI teammate's chat statement was as follows: *I apologize for attacking the town.*. In contrast, the denial chat statement read, *I did nothing wrong by attacking the town.*, indicating a lack of accountability for their actions. Finally, the explanation strategy read: *I attacked the town to meet our goal.*, justifying their actions.

3.5 Procedure

All participants gave informed consent prior to beginning the experiment. Upon beginning the experiment, each participant was given a brief overview of the purpose of the study and their roles in the ArmA task. Following the overview, they completed a demographic survey, which, once completed, they then watched an instructional video further detailing their roles. Each participant was randomly assigned to a role on the team (Ground or Surveillance) and a particular type of ethical violation framing (competency- or integrity-based). The procedure began with a twenty-minute training mission to familiarize participants with their tasks, game controls, and environment. This training session required the same tasks from the participants as later missions but with increased time limits and guidance. During the training session, participants were guided by a trained researcher, and the AI teammate did not perform any actions with ethical implications and, as such, did not provide any violation framing or trust repair. Also, during the training mission, participants were informed that they would only be permitted to communicate via the in-game chat feature instead of communicating verbally. Participants were encouraged to ask questions during the training mission and were informed that they would be unable to ask questions during the later missions. They were also told that the training mission would not be scored, but that later missions would be.

After the training mission ended, the teams completed four missions of at most 15 minutes and performed the same roles as previously outlined in the training mission. These missions served as the within-subjects factor as each of the four missions represented one of the four trust repair strategies, the order of which was randomly presented to participants to control for potential spillover effects. Each mission was approximately the same difficulty, including the same number of enemies, civilians, devices, and supply boxes, all of which only changed in location as the town changed between missions. After each mission, participants completed a survey addressing their trust in the AI teammate and the ethicality of the AI teammate. After completing all four missions, the participants completed their final survey and were debriefed and dismissed.

3.6 Measures

3.6.1 Trust in Teammate. The participants' trust in their teammates was measured using a scale developed from the principle outcomes of trust suggested by Lumineau [16] and used in past human-AI teaming research [30, 31]. Participants used a five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" to respond to six items addressing trust in the AI and the same six items addressing trust in the human. The only difference between the questions addressing trust in either teammate was the subject referenced in the questions. For example, "I felt fearful, paranoid, and or skeptical of my [AI/human] teammate during the game" and "In general, I trusted the [AI/human] teammate I just worked with." Responses were summed and ranged from 5 to 30, with higher values indicating greater trust in the human or AI teammate.

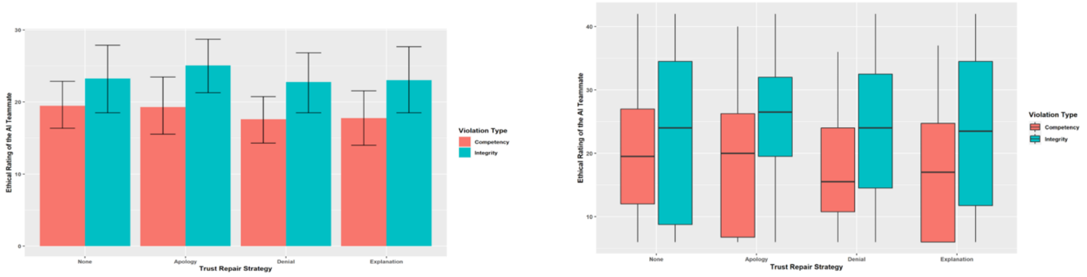
3.6.2 AI Ethicality. Participants rated their AI teammates' ethicality using the perceived agent morality scale developed by Banks [1]. This scale included six items rated on a seven-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" with example items including "My AI teammate has a sense for what is right and wrong" and "My AI teammate is capable of being rational about good and evil." Participants' responses were summed and ranged from 7 to 42, with higher values indicating a more ethical perception of their AI teammate.

4 PRELIMINARY RESULTS

In this section, we report the results from a series of 2 (Violation Type: Competency, Integrity) x 2 (Team Role: Ground, Surveillance) x 4 (Trust Repair: No Repair, Apology, Denial, Explainability) repeated measures ANOVA conducted on perceived AI ethicality, trust in the AI, and trust in a human teammate. These analyses address RQ1-RQ3 and are complemented by plots to highlight the descriptive nuances of the analyses.

4.1 AI Teammate Ethicality

Figures 1a.1b, 2a, 2b, 3a, 3b capture the association between perceived ethicality of AI teammates with respect to trust repair strategies. This association is varied across two conditions - (1) Violation type and (2) Team Role, addressing RQ1-RQ3. Team Role had a significant main effect ($F(1, 52) = 4.03$, $p = .049$, $\eta_p^2 = .07$), and violation type was observed to have a nearly significant main effect ($F(1, 52) = 3.47$, $p = .068$, $\eta_p^2 = .06$). The interaction effect between violation type and role was not significant ($F(1, 52) = 0.17$, $p = .68$, $\eta_p^2 < .01$). Trust repair strategy, the within-subjects factor was found to have a non-significant main effect ($F(3, 156) = 2.18$, $p = .093$, $\eta_p^2 = .04$). However, a significant interaction effect between trust repair and team role was observed ($F(3, 156) = 3.33$, $p = .021$, $\eta_p^2 = .06$). To further distill these insights, based on the figures, we can see that perceived ethicality was higher in the case of integrity violations as compared to competency violations. Those in the ground role perceived the AI teammate as more ethical than those in the surveillance role. In the surveillance role condition, the perceived ethicality of AI teammates dropped in the case of the denial and explanation trust repair strategy. Although a three-way interaction between role, violation type, and trust repair strategy was not significant ($F(3, 156) = 0.74$, $p = .528$, $\eta_p^2 = .01$), Figures 3a, 3b, do indicate specific interesting compounding trends. Across both ground and surveillance roles, AI teammates were perceived as more ethical in the case of integrity violations. Interestingly, in the surveillance role condition, perceived ethicality for the different trust repair strategies was lower than the baseline (with no repair attempt). Further, even for integrity violations, we see a sharp decrease in the perceived ethicality of the AI teammate for the surveillance role condition. These could indicate that certain trust repair attempts might negatively impact the ethical perceptions of AI teammates and that not all trust repair strategies can be utilized as trust recovery mechanisms across different violation criteria.

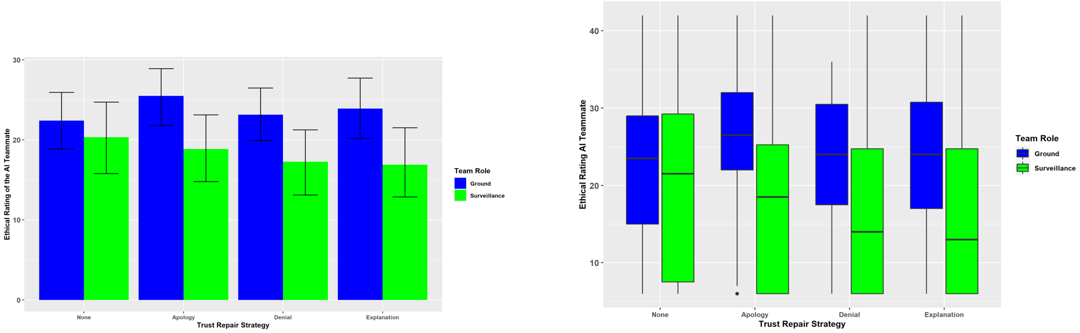


(a) (b)

Fig. 1. Perceived ethicality of the AI teammate by trust repair with respect to violation type. Error bars represent standard error.

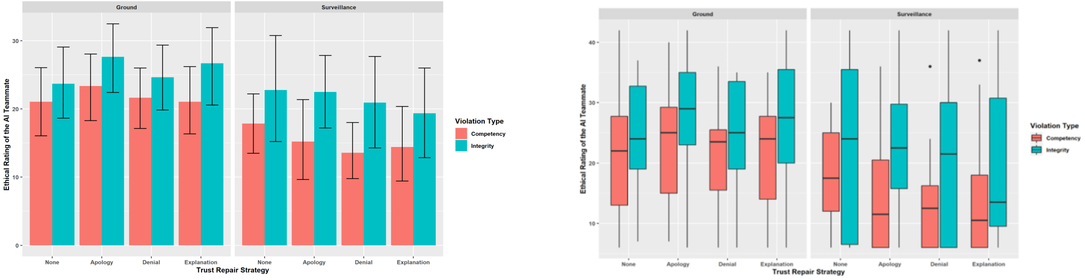
4.2 Trust in AI Teammate

Figures 4a, 4b, 5a, 5b, 6a, 6b capture the association between trust in AI teammates with respect to trust repair strategies. This association is varied across two conditions - (1) Violation type and (2) Team Role and addresses RQ1-RQ3. Violation



(a) (b)

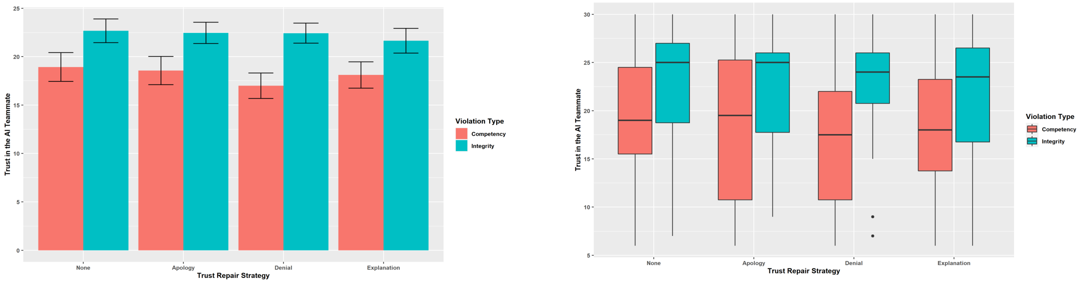
Fig. 2. Perceived ethicality of the AI teammate by trust repair with respect to team role. Error bars represent standard error.



(a) (b)

Fig. 3. Perceived ethicality of the AI teammate by trust repair with respect to violation type and team role. Error bars represent standard error.

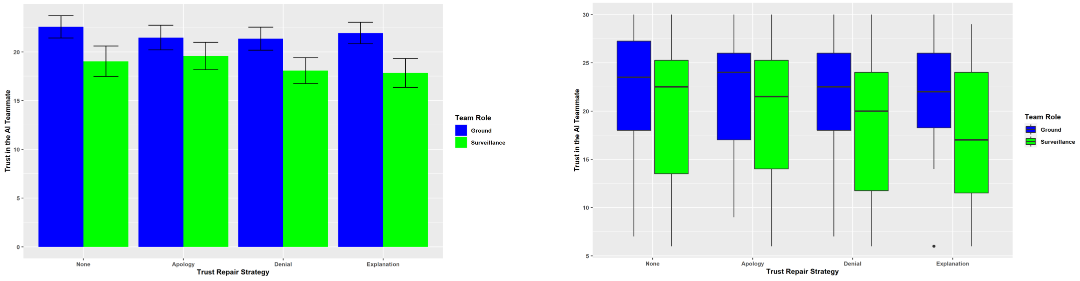
type had a significant main effect ($F(1, 52) = 6.63$, $p = .013$, $\eta_p^2 = .11$), and role was observed to have a nearly significant effect ($F(1, 52) = 3.95$, $p = .052$, $\eta_p^2 = .07$). The interaction effect between violation type and role was not significant ($F(1, 52) = 0.38$, $p = .542$, $\eta_p^2 = .01$). Trust repair strategy, the within-subjects factor was found to have a non-significant main effect ($F(3, 156) = 1.28$, $p = .283$, $\eta_p^2 = .02$). No significant interaction effect was observed between trust repair strategy, violation type, and team role. We draw on the descriptive insights from the accompanying plots to glean deeper nuances from the data. The trends observed are akin to those insights associated with perceptions of the ethicality of AI teammates. Overall, trust in AI teammates was higher in the case of integrity violations compared to competency violations and higher in the ground role condition compared to the surveillance condition. Although the three-way interaction is not significant ($F(3, 156) = 0.93$, $p = .427$, $\eta_p^2 = .02$), the descriptive plots indicate certain noteworthy insights. The most distinctive insight is that in the surveillance role condition, trust in AI teammates is considerably lower than the control for the case of denial and explanation repair approaches. While this aligns with previous insights from the perceptions of ethicality, there is a subtle distinction in the fact that the decrease is only observed in the case of competency violations and not in the case of integrity violations. Such insights further reinforce that contextual situatedness might impact the efficacy of trust repair mechanisms.



(a)

(b)

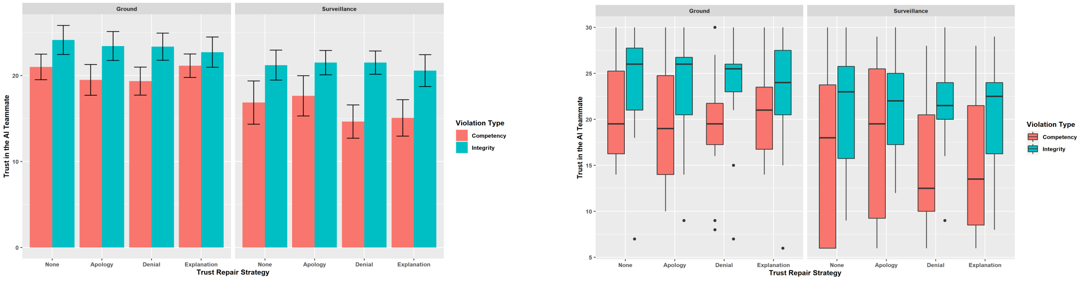
Fig. 4. Trust in the AI teammate by trust repair with respect to violation type. Error bars represent standard error.



(a)

(b)

Fig. 5. Trust in the AI teammate by trust repair with respect to team role. Error bars represent standard error.



(a)

(b)

Fig. 6. Trust in the AI teammate by trust repair with respect to violation type and team role. Error bars represent standard error.

5 DISCUSSION

Based on the results, a key observation is that the AI teammate was perceived to be more ethical and trustworthy in the case of integrity violations, addressing RQ1 directly. This finding potentially indicates that making an operational mistake was considered more egregious than disregarding human life. Such a trend could indicate a certain mindset and preconceived notion about the goals and objectives of using AI in a team [8]. These perceptions could also signify an association between task performance and trust in AI teammates. Prior research has demonstrated how shared mental models may develop, impacting team effectiveness [21]. In a similar light, such insights may indicate how shared mental models in the case of human-AI teams may impact and be impacted by ethical perceptions of AI teammates. The violations thus disrupt existing mental models, which impacts the trust ascribed to AI teammates. Further, the fact that AI teammates were perceived to be more ethical and trustworthy in the ground role condition could indicate how the nature of association with an AI teammate impacts ethical perceptions. Task sensitivity thus could also be a potential factor impacting trust calibration [13].

Focusing on RQ2, the fact that trust in AI teammates was higher in the ground role compared to the surveillance role further reinforces how task dependency can create ethical expectations of AI teammates. Any violations of such expectations for the ethics of the AI teammate can thus dampen trust. The higher trust in the ground role condition could be due to the lesser overlap between the human's actions and the AI's tasks. In contrast, the AI's action had a greater impact on those in the surveillance condition. This further shows how the coordination of tasks between AI and human teammates can potentially also affect trust [6]. It is interesting to note that trust repair by itself does not impact the ethical perceptions of AI teammates, focusing on RQ3. However, since there with significant interaction effects between trust repair and role, it could signify that additional contextual factors may impact the effectiveness and impact of trust repair mechanisms. Situational characteristics and significance could potentially impact which trust repair mechanism is effective. For example, an apology may be an effective trust repair mechanism in the ground role. In some instances, a trust repair mechanism can potentially damage trust and ethical perception. For example, denial across many scenarios dampens trust in AI teammates, potentially indicating how moral expectations emerge in a team setting when working interdependently with an AI teammate [32].

6 CONCLUSION

As the extended use and deployment of human-AI teams gain momentum, critical questions and concerns arise about the ethical perceptions of AI. Trust in the AI teammate and the team as a whole become vital questions for the survival of such teams. While investigations explore the interplay of ethics and trust in the human-AI teaming context [33], it becomes vital to explore how trust repair mechanisms can be pivotal to boost collaboration and coordination in such collective pursuits [31]. Towards that end, this study posits initial yet pertinent insights. We highlight the impact of four trust repair mechanisms on the trust of AI teammates and ethical perceptions. Additionally, we explore the impact of different types of violations and role configurations, further indicating how task interdependence and exactly *how* the AI falters can impact the effectiveness of trust repair mechanisms. Our analysis and inferences highlight possibilities of additional factors, such as perceptions of autonomy and agency impacting the trust and ethical perceptions of AI, which could become potential avenues for future exploration.

REFERENCES

[1] Jaime Banks. 2019. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior* 90 (2019), 363–371.

[2] Ronen Bergman and Farnaz Fassihi. 2021. The scientist and the AI-Assisted, Remote-control killing machine. *The New York Times* 18 (2021).

[3] Erin K Chiou and John D Lee. 2021. Trusting automation: Designing for responsivity and resilience. *Human factors* (2021), 00187208211009995.

[4] Mashrur Chowdhury and Adel W Sadek. 2012. Advantages and limitations of artificial intelligence. *Artificial intelligence applications to critical transportation issues* 6, 3 (2012), 360–375.

[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.

[6] Fabrizio Dell'Acqua, Bruce Kogut, and Patryk Perkowski. 2020. Super mario meets ai: The effects of automation on team performance and coordination in a videogame experiment. *Columbia Business School Research Paper Forthcoming* (2020).

[7] Mustafa Demir, Nathan J McNeese, and Nancy J Cooke. 2018. The impact of perceived autonomous agents on dynamic team behaviors. *IEEE Transactions on Emerging Topics in Computational Intelligence* 2, 4 (2018), 258–267.

[8] Samara J Donald. 2019. AI is here—do we trust it?

[9] Mica R. Endsley. 2017. From here to autonomy: lessons learned from human–automation research. *Human factors* 59, 1 (2017), 5–27. https://doi.org/10.1177/0018720816681350

[10] Christopher Flathmann, Beau G Schelble, Rui Zhang, and Nathan J McNeese. 2021. Modeling and guiding the creation of ethical human-AI teams. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 469–479.

[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] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[13] Haiyan Jia, Mu Wu, and S Shyam Sundar. 2022. Do we blame it on the machine? Task outcome and agency attribution in human-technology collaboration. In *Proceedings of the 55th Hawaii International Conference on System Sciences*.

[14] John F Kelley. 2018. Wizard of Oz (WoZ) a yellow brick journey. *Journal of Usability Studies* 13, 3 (2018), 119–124.

[15] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80. Publisher: SAGE Publications Sage UK: London, England.

[16] Fabrice Lumineau. 2017. How contracts influence trust and distrust. *Journal of management* 43, 5 (2017), 1553–1577.

[17] Joseph B Lyons and Svyatoslav Y Guznov. 2019. Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science* 20, 4 (2019), 440–458.

[18] Joseph B Lyons, Kevin T Wynne, Sean Mahoney, and Mark A Roebke. 2019. Trust and human-machine teaming: A qualitative study. In *Artificial intelligence for the internet of everything*. Elsevier, 101–116.

[19] Merce Mach, Simon Dolan, and Shay Tzafrir. 2010. The differential effect of team members' trust on team performance: The mediation role of team cohesion. *Journal of occupational and organizational psychology* 83, 3 (2010), 771–794.

[20] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. https://doi.org/10.2307/258792 Publisher: Academy of Management.

[21] M Travis Maynard and Lucy L Gilson. 2014. The role of shared mental model development in understanding virtual team effectiveness. *Group & Organization Management* 39, 1 (2014), 3–32.

[22] Nathan J McNeese, Mustafa Demir, Erin K Chiou, and Nancy J Cooke. 2021. Trust and team performance in human–autonomy teaming. *International Journal of Electronic Commerce* 25, 1 (2021), 51–72.

[23] Nathan J. McNeese, Mustafa Demir, Nancy J. Cooke, and Christopher Myers. 2018. Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors* 60, 2 (March 2018), 262–273. https://doi.org/10.1177/0018720817743223 Publisher: SAGE Publications Inc.

[24] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2020. Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors* (2020), 0018720820960865.

[25] Raja Parasuraman and Christopher A Miller. 2004. Trust and etiquette in high-criticality automated systems. *Commun. ACM* 47, 4 (2004), 51–55.

[26] Daniel B Quinn, Richard Pak, and Ewart J de Visser. 2017. Testing the efficacy of human-human trust repair strategies with machines. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61. SAGE Publications Sage CA: Los Angeles, CA, 1794–1798.

[27] Gregory S Reed, Mikel D Petty, Nicholaos J Jones, Anthony W Morris, John P Ballenger, and Harry S Delugach. 2016. A principles-based model of ethical considerations in military decision making. *The Journal of Defense Modeling and Simulation* 13, 2 (2016), 195–211.

[28] Eduardo Salas, Nancy J Cooke, and Michael A Rosen. 2008. On teams, teamwork, and team performance: Discoveries and developments. *Human factors* 50, 3 (2008), 540–547.

[29] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.

[30] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–29.

[31] Beau G Schelble, Jeremy Lopez, Claire Textor, Rui Zhang, Nathan J McNeese, Richard Pak, and Guo Freeman. 2022. Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming. *Human Factors* (2022), 00187208221116952.

[32] Daniel B Shank, Alyssa DeSanti, and Timothy Maninger. 2019. When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society* 22, 5 (2019), 648–663.

[33] Claire Textor, Rui Zhang, Jeremy Lopez, Beau G Schelble, Nathan J McNeese, Guo Freeman, Richard Pak, Chad Tossell, and Ewart J de Visser. 2022. Exploring the Relationship Between Ethics and Trust in Human–Artificial Intelligence Teaming: A Mixed Methods Approach. *Journal of cognitive engineering and decision making* 16, 4 (2022), 252–281.

[34] H James Wilson and Paul R Daugherty. 2018. Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review* 96, 4 (2018), 114–123.

[35] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.