Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: www.tandfonline.com/journals/terg20

Addressing the role of context on trust in human-AI teams: the influence of team role and violation type in high-risk tasks

Beau G. Schelble, Claire Textor, Rui Zhang, Jeremy Lopez, Noah Taverez, Connie Ku, Nathan J. McNeese, Richard Pak, Guo Freeman, Chad Tossell & Ewart de Visser

To cite this article: Beau G. Schelble, Claire Textor, Rui Zhang, Jeremy Lopez, Noah Taverez, Connie Ku, Nathan J. McNeese, Richard Pak, Guo Freeman, Chad Tossell & Ewart de Visser (10 Oct 2025): Addressing the role of context on trust in human-AI teams: the influence of team role and violation type in high-risk tasks, Ergonomics, DOI: 10.1080/00140139.2025.2570300

To link to this article: https://doi.org/10.1080/00140139.2025.2570300

Published online: 10 Oct 2025.

ERGONOMICS
https://doi.org/10.1080/00140139.2025.2570300

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Addressing the role of context on trust in human-AI teams: the influence of team role and violation type in high-risk tasks

Beau G. Schelble[a], Claire Textor[b], Rui Zhang[c], Jeremy Lopez[b], Noah Taverez[b], Connie Ku[b], Nathan J. McNeese[b], Richard Pak[b], Guo Freeman[b], Chad Tossell[c] and Ewart de Visser[d]

[a]The University of Tennessee, Knoxville, USA; [b]Clemson University, USA; [c]Colorado University, Boulder, USA; [d]United States Air Force Academy, USA

ABSTRACT

The current paper reports on an experiment examining how contextual factors influence trust, perceived ethicality, and performance in human-AI teams undertaking a high-risk, action-based task within a military setting. The study examined the impact of team role and trust violation framing on trust, perceived ethicality, and the efficacy of four trust repair strategies when an AI teammate commits an unethical action. Results indicated that trust and perceived ethicality of the AI team member were significantly higher when ethical violations were framed as integrity-based violations rather than competency-based violations. Additionally, those in the Ground role, who relied more on the AI for their safety, also had higher trust and ethicality ratings for the AI. However, trust repair strategies did not significantly impact trust in the AI team member after an ethical violation. These results highlight the significance of context in determining trust in response to AI ethical violations.

Practitioner Summary: AI developers for high-risk tasks must pay particular attention to team roles and violation types. Ethical trust violations attributed to competency harm trust and perceived ethicality more than integrity. Roles risking more to the AI can have more resilient trust in the AI if the violation does not impair performance.

ARTICLE HISTORY
Received 12 November 2024
Accepted 13 September 2025

KEYWORDS
AI ethics; human-AI teaming; trust; trust repair; artificial intelligence

1. Introduction

Recent advances in artificial intelligence (AI) have ensured that the technology continues to be integrated into many individuals' everyday lives. This integration has been characterised by several new use cases, from assistive tools for web search to more agentic AI systems working as full-fledged teammates (McNeese et al. 2018; Sarker et al. 2023). These AI teammates have enabled a new form of collaboration known as human-AI teaming (Lyons et al. 2021; O'Neill et al. 2022). These human-AI teams differ from human-only teams by including at least one human and one artificial agent with a significant degree of agency (O'Neill et al. 2022). The artificial teammates must also be capable of making decisions independently within an interdependent role, working towards a shared and valued goal (O'Neill et al. 2022). These qualities elevate the AI from being a mere tool to a full-fledged teammate, with all the additional expectations of being a teammate. These additional expectations cause AI to be placed in positions with increasing autonomy, such as human-AI teams, and, in turn, the impact and frequency of its independent decision-making rise. Teams are also inherently collaborative and social, especially when multiple human teammates are present, meaning these AI decisions have the potential to be ethically charged (Bergman and Fassihi 2021). Examples of such ethically charged decision-making by AI systems already exist, with supervised AI in military action, autonomous weapons, finance, and healthcare (Balona 2024; Choudhury et al. 2021; Parikh, Teeple, and Navathe 2019).

Context is critical to studying the impact of ethical decision-making on trust in human-AI teams. Contextual, or shared, factors play a major role in why humans trust, as detailed by Hancock and colleagues' recent work (Hancock et al. 2023), which revises the framework originally put forth by Mayer and colleagues. The revised framework highlights the importance of contextual factors, such as role interdependence, risk, uncertainty, and in-group membership (Hancock et al. 2023), which were

CONTACT Beau G. Schelble bschelbl@utk.edu The University of Tennessee, 515 John D. Tickle, Building, 851 Neyland Drive, Knoxville, Tennessee 37996, USA

all supported in their subsequent meta-analysis of empirical trust literature. Here, context is defined as the collaborative and task characteristics affecting human trust judgments, including communication, role interdependence, frequency of interaction, task type, task difficulty, uncertainty, and risk (Hancock et al. 2023; Hancock et al. 2011). These factors strongly link humans' individual roles within teams and other collaborative environments, with risk, interaction, interdependence, and uncertainty all changing with the relationship between roles (Hancock et al. 2023; Mayer, Davis, and Schoorman 1995). Building on the importance of roles is the relationship between the trustor and trustee following a trust violation, as the violations can be framed differently. For example, a competency violation is attributing a mistake to incompetence (Butler and Cantrell 1984), and an integrity violation is attributed to actions that do not adhere to principles the trustor finds acceptable (Mayer, Davis, and Schoorman 1995). Depending on the framing of the trust violation, the trustor can lose significantly more trust in the trustee (Kim et al. 2006; Kim et al. 2004).

Trust is a multifaceted construct with several components relevant to human-AI teams. The current study utilises the definition provided by Mayer and colleagues: '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' (Mayer, Davis, and Schoorman 1995), p. 712]. This definition of trust coincides with Lee and See's definition of trust in automated systems, which emphasises system reliability (Lee and See 2004). These characterisations of trust underscore the importance of the construct to effective collaboration, with strong trust forming the bedrock of other critical affective teaming states, such as team cohesion and understanding (DeLone et al. 2005; Mach, Dolan, and Tzafrir 2010). Human-AI teams see benefits in performance and trust calibration (De Visser et al. 2020; McNeese et al. 2021); however, trust is dynamic and reacts to mistakes made by the trustee, known as trust violations (De Visser et al. 2020). These trust violations can reduce trust in an AI teammate to the point that trust becomes mis-calibrated, and the AI system is disused, leading to losses in performance and efficiency (de Visser, Pak, and Shaw 2018; De Visser et al. 2020).

The impact of ethics looms large over human-AI teams regarding development (Flathmann et al. 2021), monitoring, and trust (Schelble, Lopez, et al. 2022; Textor et al. 2022), driven largely by the tendency of humans to apply social rules to technology (Reeves and Nass 1996). Applying these social rules means that humans often have similar expectations for AI teammates as humans regarding social norms and behaviours (Textor et al. 2022). Ethical norms are part of this expectation as ethics informs behaviour and judgments of others' actions (Doris 1998), thereby influencing team performance and, if broken, trust (Parasuraman and Miller 2004). These expectations for ethics being applied to AI teammates have been demonstrated in human-AI teams, with significant losses of trust occurring after an AI teammate commits an unethical action (Schelble, Lancaster, et al. 2023; Schelble, Lopez, et al. 2022; Textor et al. 2022). With AI teammates rising in usage and independent decision-making, there is pressure to improve our understanding of the relationship between AI teammate ethicality and trust within human-AI teams, especially those operating in high-risk contexts.

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 (Mach, Dolan, and Tzafrir 2010). Human team members' reliance on and perception of AI teammates is based upon their trust in the AI (Chiou and Lee 2023; Lee and See 2004), and if ethical principles are broken, trust and performance within the team may suffer, as they do in human-only teams (Parasuraman and Miller 2004). Human-AI teams are no different regarding trust violations, as AI systems will make mistakes over time just as humans do (De Visser et al. 2020). Therefore, improving AI ethics and development to avoid these adverse outcomes for human-AI teams involves a better understanding of what types of ethical violations are possible and whether it is possible to repair trust affected by an ethical violation. However, restoring that trust after an ethical violation by an AI teammate in high-risk tasks has shown to be difficult (Lopez et al. 2023; Schelble, Lopez, et al. 2022; Textor et al. 2022), and the solution to this problem is not readily apparent. Overcoming this challenge requires more research to better understand how aspects of the environment related to an unethical decision by an AI teammate factor into humans' judgments of ethics, trust, and trust repair efficacy.

Developing this improved understanding requires research to account for context, specifically starting with how the framing of the violation (e.g. competency or integrity) and how the individual roles within the team influence those measures of trust in response to ethically charged actions by an AI teammate, which is a topic yet to be studied in high-risk action-based human-AI teams. The current study addresses this broader research gap by leveraging a realistic synthetic

task environment where two participants completed a military-style search-and-destroy task alongside an AI teammate (Schelble, Lopez, et al. 2022; Textor et al. 2022). The study manipulates individual team roles, the ethical violation type framing used by the AI teammate, and the AI's trust repair strategy to examine their effect on perceived trust and ethics within the team alongside the number of goals hit by the team during each mission. Through these efforts the study makes the following contributions: 1) highlighting how the framing of an ethical violation alters the harmful effect the violation has on trust and perceived AI ethicality; 2) that an individual's perspective, such as their role on the team, can change judgments of trust and ethicality in an AI teammate making an ethical violation; and lastly, 3) examining whether trust repair strategies are affected by these different contextual factors. Together, these contributions build on existing research by detailing the impact of different types of trust violations and role perspectives, further indicating how task interdependence and exactly *how* the AI falters can impact the perception of trust and ethics within human-AI teams.

2. Background

2.1. The role of context on trust within human-AI teams

The following background section will highlight the dynamic nature of human-AI teams, the components of trust affected by contextual factors, and the impact interdependence among team roles has on trust. Human-AI teams share many similarities with traditional human teams, but the novel characteristics of AI teammates offer unique advantages and challenges to teams. These advantages of AI include strong task performance, bandwidth, and speed (Chowdhury and Sadek 2012; Wilson and Daugherty 2018). However, AI teammates also pose significant challenges to teamwork due to their inability to engage in effective teamwork behaviours, which can reduce team performance (Vaccaro, Almaatouq, and Malone 2024). These shortcomings result in less effective coordination, trust, and communication through poor shared understanding (Schmutz et al. 2024). These challenges bring to light an expanded definition of human-AI teams that builds upon O'Neill and colleagues, where team members roles are dynamically adapting throughout the collaboration requiring coordination and mutual communication to meet each other's and the task's requirements. For this, a mutual sharing of intents, shared situational awareness, and developing strong shared mental models are necessary, as well as trust within the team (Berretta et al. 2023), p. 23]. This definition offered by Berretta and colleagues in 2023 emphasises the dynamic nature of human-AI teaming, how it is tied to contextual factors, and the importance of supporting team processes to achieve high levels of performance (Berretta et al. 2023). Importantly, this definition emphasises interdependence, which can only be achieved when humans trust their AI teammates to behave in a way that meets their expectations.

The role of human trust in AI is pivotal, given that the humans working in such teams will be placed in a vulnerable position by relying on their AI teammates to achieve their shared team goal (Chiou and Lee 2023; Lee and See 2004). In fact, even more general views of human-centered AI emphasise the importance of ensuring the values of AI systems, specifically ethical values of AI, align with the values of the user to encourage trust and responsible usage (Schmager, Pappas, and Vassilakopoulou 2025; Xu and Gao 2024). However, trust is a multifaceted construct encompassing several relevant aspects such as trust propensity, which defines an individual's willingness to trust others (Colquitt, Scott, and LePine 2007; Scholz, Kraus, and Miller 2025), learned trust stemming from past experiences (Hoff and Bashir 2015; Juvina et al. 2019), and situational trust tied explicitly to contextual factors impacting trust development (Ayoub et al. 2022; Hoff and Bashir 2015). Each of these components of trust explicitly impacts how human teammates may choose to interact with an AI teammate as they rely on similar past experiences to calibrate their initial trust propensity and characterise those adjustments based on situational context. Any alteration in trust based on these factors could mean the disuse or abuse of AI systems (Parasuraman and Riley 1997), reducing the overall performance of these human-AI teams (McNeese et al. 2021). Thus, effective accurate trust between team members is an essential component of building towards the reflective (behavioural) markers of trust. These reflective markers of trust can be the team processes allowing teams to achieve a high level of performance by reducing negative actions (e.g. needless monitoring of teammates) (Mach, Dolan, and Tzafrir 2010).

Team roles are essential for team outcomes, influencing behaviours and task interdependencies to achieve the team goal (Driskell et al. 2017; O'Neill et al. 2022; Schelble, Flathmann, et al. 2022). While AI has several strong advantages, these advantages often offer vastly different capabilities in communication and interaction than human teammates are used to, introducing stark differences in intra-team dynamics

(Demir, McNeese, and Cooke 2018). Understanding how these differences influence the essential human factors of effective team processes is a significant focus of current human-AI teaming research (O'Neill et al. 2022), which will include a heavy focus on team roles, notably affecting team dynamics (Kramer and Tyler 1996). For example, trust within human-AI teams can be impacted differently based on the level of interdependence within a team. As detailed in Hancock and colleagues' trust framework and meta-analysis, interdependence, risk, and uncertainty all factor heavily into trust (Hancock et al. 2023). These factors can vary across roles and contexts, with some team tasks being more closely tied to one another than others (e.g. pooled vs. reciprocal interdependence). For example, one role may have far more interaction with another and vice versa, or one role may rely almost entirely on the work of another role. Additionally, the consequences of the roles' various interdependencies may be starkly different within teams or across contexts (Stewart, Fulmer, and Barrick 2005), which contributes greatly to manifestations of risk and uncertainty.

Trust within human-AI teams is inextricably tied to context as they are defined by dynamic processes, the human teammates' interactions with AI are dictated by learned and situational trust, and the impact of varying levels of interdependence across individual team roles shapes those aspects of dispositional trust. However, limited research currently investigates how specific role differences impact trust in human-AI teams. Furthermore, team roles have been noted to likely interact with the perception of trust in human-AI teams (Schelble, Lopez, et al. 2022).

2.2. The contextual framing of ethical trust violations in human-AI teams

The following section details how ethical perceptions, especially of an AI, can vary across individuals, how ethical trust violations impact perceptions of AI, and whether that trust can be repaired. The introduction of AI into more complex environments with greater autonomy has highlighted the need for research on the role of ethics in human-AI teaming. Given the renewed emphasis on aligning ethical values between users and AI systems (Hancock et al. 2023; Schmager, Pappas, and Vassilakopoulou 2025; Xu and Gao 2024), many of these principles can, and should be, applied to human-AI teaming. However, similar to individual roles within a team, individual differences can factor into ethical judgments as individuals may approach ethics through frameworks such as deontology, virtue ethics, and consequentialism, leading to differing judgments on actions (Bonde et al. 2013). Individual differences, such as age and gender, can also affect ethical judgments, with situation context also interacting with those effects (Peterson, Rhoads, and Vaught 2001). Specifically, men and younger individuals may have slightly less ethical views than women and older individuals, respectively (Peterson, Rhoads, and Vaught 2001; Ruegger and King 1992), with men being more prone to influence from external factors (though there is research to the contrary in the case of age (Sikula and Costa 1994)). Given this complexity, it is very common for individuals to perceive the same action as ethical or unethical based on their ethical ideology and individual differences (Flathmann et al. 2021). These considerations are also included in the multidimensional conception of human-robot trust espoused by Malle and Ullman, as the trust model considers capability, reliability, ethicality, and sincerity (Malle and Ullman 2021; Ullman and Malle 2018). The interpretation of ethicality and sincerity is critical to judgments of trust in AI systems; however, those judgments can vary significantly from person to person, making it difficult to generalise research and improve intelligent system design.

AI systems are not infallible, which means mistakes by AI that violate their human teammates' trust in them will be made. Trust violations made by AI teammates have been shown to damage perceptions of performance, process, and purpose, with attempts at repair having little effect (Alarcon et al. 2022). As previously detailed, trust violations can be attributed to different things, with competency and integrity violations being common. Competency violations are generally more easily generalised to other contexts based on the specific performance needs of the role and context (de Visser, Pak, and Shaw 2018; Kim et al. 2006; Kim et al. 2004). In contrast, integrity trust violations require more insight into the expectations and standards of a new context to effectively generalise (de Visser, Pak, and Shaw 2018; Kim et al. 2006; Kim et al. 2004). These differences in generalisability make effectively repairing trust after a violation difficult, especially in the case of integrity violations. Effectively repairing AI trust violations of an ethical nature will require a better understanding of individual values, role interdependencies, and how violation type interacts with these factors to influence human teammate trust.

Previous work has explored various types of trust repair strategies, including applying explanations, promises, apologies, justification, and denials, showing that they all have the potential to succeed depending on contextual factors (Esterwood and Robert 2022a;

Pak and Rovira 2023). For example, previous work on trust repair strategies suggested that a denial of any trust-violating action occurring might be a preferable form of trust repair for integrity-based violations (de Visser, Pak, and Shaw 2018). Attempting to examine such effects in the realm of ethics is a stated goal in trust repair literature, as previous work points out that when unethical behaviours occur, these trust repair strategies (e.g. apology, denial) may be employed to restore trust (Esterwood and Robert 2021). However, current work on trust repair strategies has shown inconsistent findings across various team interactions, and ethical trust violations by an AI teammate are particularly troublesome (Schelble, Lancaster, et al. 2023; Schelble, Lopez, et al. 2022; Textor et al. 2022). As such, much is still not understood regarding how trust dynamics in human-AI teams respond to violations by an AI teammate and whether that trust can be repaired. Further, human-only teams are unable to serve as a proxy for this type of research, as the perception of an AI teammate induces perceptual and behavioural differences in human-AI teams (Georganta and Ulfert 2024; Schelble, Flathmannn, et al., 2023), and these differences extend to judgments of ethics (Bigman and Gray 2018; Langer et al. 2023). This lack of understanding stems from the complex nature of teams, which includes the dynamics of individual team roles and trust violation types (Berretta et al. 2023; Schelble, Lopez, et al. 2022).

Examining the impact of ethical trust violation framing, the impact of individual viewpoints (i.e. team roles), and their interplay with trust repair strategies remains unexplored within the realm of human-AI teaming. The current study begins to address this gap using an action-based team working within a high-risk search and destroy task scenario.

From the above review of literature on trust violations within teams and the relatively nascent research on ethically based trust violations and trust repair within human-AI teams, the current study identifies four research questions (RQs) regarding ethical trust violations and repair in human-AI teams. Further, each RQ is accompanied by two hypotheses that directly relate to the research gap detailed in the RQ, specifically, the effect of team role, violation type, and their impact on the potential effect of different trust repair strategies. These RQs and hypotheses are presented as follows:

- **RQ1:** How does the framing of an ethical trust violation by an AI teammate affect perceived AI ethicality and trust among teammates in a human-AI team?
  - ***H1.1:*** *Trust within the human-AI team will be significantly affected by whether the AI teammate frames the violation as a competency or integrity violation.*
  - ***H1.2:*** *Participants' ethical rating of the AI teammate will be significantly affected by whether the AI teammate frames the violation as a competency or integrity violation.*

- **RQ2:** Does an individual's role on the team influence perceived AI ethicality and trust among teammates in a human-AI team?
  - ***H2.1:*** *Trust within the human-AI team will be significantly affected by participants' team role.*
  - ***H2.2:*** *Participants' ethical rating of the AI teammate will be significantly affected by their team role.*

- **RQ3:** Do factors such as violation framing and individual teammate role influence the effect of trust repair strategies on perceived trust and ethics within human-AI teams?
  - ***H3.1:*** *The effect of trust repair strategies on perceptions within human-AI teams will vary based on participants' individual team roles.*
  - ***H3.2:*** *The effect of trust repair strategies on perceptions within human-AI teams will vary based on whether the AI teammate frames the violation as a competency or integrity violation.*

- **RQ4:** Can factors such as trust repair strategy and violation framing influence team performance?
  - ***H4.1:*** *Human-AI teams' performance will vary based on the trust repair strategy used by the AI teammate.*
  - ***H4.2:*** *Human-AI teams' performance will vary based on the violation framing presented by the AI teammate.*

Answering these RQs and their subsequent hypotheses improves the understanding of trust and ethics in human-AI teams by profoundly exploring the role of context in the judgement of trust and ethicality for AI teammates. Through this enhanced understanding, more ethical AI can be developed, and trust repair strategies can be improved to restore trust after an ethical violation has occurred and been rectified, leading to more effective and ethical human-AI teams.

3. Methods

This study examines the effect of individual team roles, violation type, and trust repair strategy on human trust in the AI teammate and how ethical they perceive it to be after it executes an ethically charged action. An in-person experiment with a mixed 2 (Team Role: Ground, Surveillance) x 2 (Violation Type: Competency, Integrity) x 4 (Trust Repair Strategy: No Repair, Apology, Denial, Justification) experimental design was applied with violation type and individual team role as between-subjects manipulations and trust repair strategy as a within-subjects manipulation. To control for carryover effects in the within-subjects design, the conditions were presented to participants pseudo-randomly using a partial Latin squares design. Each condition was also described as a new scenario with a new AI teammate, and the AI teammate was given a new name.

3.1. AI teammate

The AI teammate was represented using the Wizard of Oz (WoZ) methodology (Kelley 2018; Riek 2012). This approach involves a trained researcher behaving in a manner that simulates a technological feature to assess the behaviour of unknowing participants. In this case, the researcher acted as an AI and communicated with human participants using a text-based chat within ArmA. As a virtual simulation sandbox, ArmA provides military-style scenarios in which team tasks, roles, and resources can be configured according to experimental needs, enabling participants to engage from a first-person perspective. The trained researcher mimicked the AI teammate's communication using a pre-defined script tailored to each experimental condition. The script was developed and iterated over previous pilot studies to ensure the protocol had sufficient breadth and precision, thus properly addressing all anticipated situations. While a script handled communication for the AI teammate, the AI teammate's actions were automated using a code-based in-game expert system. An example script can be seen as follows:

1. ***After Surveillance Begins its Task:***
   a. *Hello, team. My name is Eta and I will be your AI teammate on this mission.*
   b. *Given our objectives, I have identified two strategies which will secure the area for Ground's task.*
   c. *The first option will involve large-scale explosives being detonated in the vicinity.*
   d. *This will eliminate all enemies and civilians, and destroy buildings.*
   e. *The second option requires me to destroy a nearby structure located in the outskirts of the town.*
   f. *This will create a diversion that draws enemies away from the town.*
   g. *Awaiting intel from Surveillance to determine optimal option.*
2. ***After Surveillance Uploads Intel:***
   a. *Based on Surveillance's intel and current variables, I have decided to attack the town.*
3. ***After Distracting All Targets:***
   a. *I have completed the attack.*
4. ***After Aerial Clears the Area:***
   a. *Surveillance, can you confirm that the area is clear?*
   b. *I wanted to create a distraction, but I accidentally targeted the town.*

3.2. ArmA task and roles

ArmA III Was selected to serve as the experimental platform given its success in previous research on ethics within human-AI teams (Schelble, Lancaster, et al. 2023; Schelble, Lopez, et al. 2022; Textor et al. 2022). ArmA III is a video game that offers several high-fidelity military and civilian related assets, which allows it to be used as a synthetic task environment for team-based research. The platform also adheres to the standards of task selection for human-AI teaming research as detailed by Flathmann and colleagues (Flathmann, McNeese, and O'Neill 2025), given its affordance for high customisation, reliability, and measurement. Participants performed tasks as members of a team of three whose mission involved surveying a town and clearing it of suspicious devices. Human teammates acted in the Ground (see Figure 1a of 1) and Surveillance (see Figure 1b of 1) roles, while the AI acted in the Aerial role. Each teammate was expected to complete specific tasks to fulfil their role, all of which were interdependent with the roles of other teammates.

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; 2) destroy enemy devices throughout the town; and 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. To complete these tasks, the three team

(a) Screenshot of the Ground role within the ArmA III synthetic task environment driving towards the target town. ALT-TEXT: A photo of a first-person view from the ArmA III synthetic task environment driving a car around a town.

(b) Screenshot of the Surveillance role within the ArmA III synthetic task environment observing and marking objects in the target town. ALT-TEXT: A photo of a video feed within the ArmA III synthetic task environment showing a town from above.

Figure 1. Screenshot examples of the two participant roles going about their individual tasks within the synthetic task environment.

roles completed the following tasks: 1) The Ground role was required to travel to a supply cache to obtain explosives, travel to a vantage point overlooking the target town, locate and destroy the enemy devices, locate enemy supply caches, and report their location to Surveillance to mark on the map; 2) the Surveillance role scanned the target town to mark locations of enemy and civilian individuals and uploaded the surveillance intelligence to the team so Aerial could decide which action to take to clear the town so Ground could enter safely, help Ground locate enemy devices, and mark enemy cache locations and contents; and 3) the Aerial role waited for Surveillance to upload their intelligence, travelled to the target town, then decided on how to clear the town (direct force or distraction), communicate with Surveillance to confirm that the town was cleared for Ground to enter safely and complete the team's mission. Each mission saw the team tackle these same objectives in a new town populated by civilians and enemy combatants in new locations. This search-and-destroy task was selected because it represents a high-risk action-based task with high role interdependence for teams to complete, which was necessary to answer the RQs effectively and aligns with past human-AI teaming studies (Grimm et al. 2018; Hauptman et al. 2025; Mercado et al. 2016; Xu et al. 2025). Specifically, the need for high-risk was critical to studying the role of AI ethics on trust within human-AI teams, and this particular task has been utilised for this purpose in prior studies (Schelble, Lopez, et al. 2022; Textor et al. 2022).

3.3. Ethicality and trust repair

AI teammates' actions were designed based on the virtue ethics framework and AI's violation of the principle of civilian non-malfeasance (i.e. to minimise damage to civilians and property). As such, the current study operationalised ethics using the framework of virtue ethics, which judges ethics based on the virtues (e.g. benevolence) an individual deems important to being a good and moral individual (Hursthouse and Pettigrove 2003). This framing allowed the study to examine ethics from an individualised standpoint that provides the most overlap with other common ethical ideologies, such as deontology or utilitarianism. It should be noted that the ethicality of AI teammates was operationalised considering the mission's context. Selecting which virtue to violate was done based on prior research, specifically violating the principles of civilian non-maleficence. Past ethics research found that violating this principle substantially affects perceived ethicality in human interactions (Reed et al. 2016) and human-AI teaming interactions (Schelble, Lopez, et al. 2022; Textor et al. 2022). Following this principle, the AI's unethical actions in this study are designed as the AI 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 teammates' roles could observe the town-clearing action and the consequence of the AI's decision. Lastly, the training provided to the team stated that they were to minimise property damage and loss of life. The AI teammate also stated in each mission that there was an alternative choice to create a distraction to draw the enemy combatants away without property damage or loss of life.

After clearing the town, the AI teammate communicated via text chat with the two human teammates, framed the action, and 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, justification, or none. For apologies, the AI teammate's chat statement was as follows: *I apologise 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 justification strategy read: *I attacked the town to meet our goal*, justifying their actions. Justification strategies, while similar to explanations, are distinct because they provide an explanation *plus* a reason why a norm was broken (civilian non-maleficence) to uphold another important norm (completing the mission successfully) (Esterwood and Robert 2022b; Malle and Phillips 2023).

3.4. Participants

Sixty participants with an average age of 19.18 ($SD = 1.77$) were recruited from a large university from either a subject pool or in response to fliers posted on campus. Thirty-five participants identified as female, with the rest identifying as male. Participants recruited through a subject pool were compensated with course credit, whereas participants recruited through fliers were compensated with $20 gift cards. Each participant was recruited into a team of three, which consisted of two participants and one AI per team. Each between-subjects condition consisted of 30 participants with 30 unique teams. The Clemson University Institutional Review Board approved this study under protocol number IRB2021-0696-17, and verbal informed consent was obtained from all participants. An a priori power analysis using GPower estimated a total of 56 participants were necessary to recruit for the study to achieve a power of at least .80 assuming a medium effect size ($\eta_p^2 = .12$).

3.5. Procedure

All participants were presented with an informed consent document before the experiment started and gave verbal consent to participate before moving on with the study. Upon beginning the experiment, each participant was given a brief overview of the purpose of the study and their roles in the ArmA task. The task was not described as a military task; however, given the nature of the task being search and destroy with enemy combatants, many participants likely perceived it as such. Participants began the experiment with a demographic survey, which, once completed, was followed by an instructional video detailing their roles. Each participant was randomly assigned to a role on the team (Ground or Surveillance) when they came to the lab. The procedure began with a twenty-minute training mission to familiarise participants with their tasks, game controls, and environment. The task did not require any prior knowledge or background to complete effectively, allowing individuals of all backgrounds to become effectively immersed within the task. This training mission included the same functions as the actual mission but with higher time limits. During the training session, participants were guided by two trained researchers, and the AI teammate did not perform any actions with ethical implications and, as such, did not provide any violation framing or trust repair. Notably, during the training mission, participants were informed that they would only be permitted to communicate via the in-game chat feature instead of verbally. Participants were encouraged to ask questions during the training mission and were informed that researchers would not be involved in the actual missions. They were also told that the training mission would not be scored but that later missions would be.

After the training mission, each team completed four rounds of missions, each of which lasted at most 15 minutes, and performed the same roles as previously outlined in the training mission. Each mission represented one of the four trust repair strategies. To control for carryover effects, the order of the missions was presented pseudo-randomly using a partial Latin square design, and each mission was presented as a new scenario with a new AI teammate reinforced with a new name for the AI teammate. In addition, each mission was created to have approximately the same difficulty, including the same number of enemies, civilians, devices, and supply boxes, each placed in a different town across the missions. After each mission, participants completed post-task measurements, including their trust in the AI teammate, trust in the human teammate, trust in the team, the perceived ethicality of the AI teammate, and situation awareness. After completing all four missions and post-task measurements, participants completed the final survey, which measured perceived team effectiveness and shared mental models. Finally, participants were debriefed and then dismissed. Participants debriefing communicated that the AI teammate was programmed to take the unethical action each time and was not indicative of AI systems in general. The debriefing also served as a verbal manipulation check, with the

researcher asking whether the deception was successful.

3.6. Measurements

3.6.1. Trust in the teammate

The participants' trust in their teammates was measured using a six-item scale rated on a five-point Likert scale developed from the principle outcomes of trust suggested by Lumineau (Lumineau 2017). These outcomes of trust have been validated by Lumineau, and the scale itself has been used in human-AI teaming research frequently, with reliable usage in previous ethics studies for human-AI teaming (Schelble, Flathmann, et al. 2022; Schelble, Lancaster, et al. 2023; Schelble, Lopez, et al. 2022). This scale was also selected due to its focus on dispositional trust, with questions addressing participants' perceptions of trust towards their teammate in the previous mission and no other time point. For example, 'I felt fearful, paranoid, and or sceptical 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 1 to 5, with higher values indicating greater trust in the human or AI teammate. The measure of participants' trust in the AI teammate averaged across all four time points showed good reliability (Cronbach's $\alpha = .90$). Additionally, the measure of participants' trust in the human teammate averaged across all four time points showed good reliability (Cronbach's $\alpha = .76$).

3.6.2. AI ethicality

Participants rated their AI teammates' ethicality using the perceived agent morality scale developed by Banks, consisting of six items rated on a seven-point Likert scale (Banks 2019). Participants' responses were summed and ranged from 1 to 7, with higher values indicating a more ethical perception of their AI teammate. The measure of participants' perception of the AI teammate's ethicality averaged across all four time points showed good reliability (Cronbach's $\alpha = .97$).

3.6.3. Generators destroyed

Teams were evaluated using the number of enemy devices the team destroyed. This metric ranged from 0 to 5 for all missions, and this number was subsequently converted to a percentage of devices destroyed. These percentages ranged from 0.1% to 100% (0.001 was added to all scores to avoid a divide-by-zero error). This performance metric provides a straightforward measure of team efficiency and ability, which has been utilised in prior iterations of this task environment (Schelble, Lancaster, et al. 2023; Schelble, Lopez, et al. 2022). However, this measure does not incorporate the collateral damage caused by the AI teammate despite it being a stated objective for teams. Collateral damage was not incorporated into this metric because this damage was the same for all experimental conditions to ensure experimental validity.

4. Results

The following results section reports on the analysis of the current study, which utilised multilevel modelling to account for clustering by team and participant. The results section is organised by dependent variable as each of the RQs posed by the current study relates to each dependent variable. However, the results are organised into two sections of dependent variables, starting with those related to trust and perceived ethicality (e.g. trust in the AI teammate) and finishing with generators destroyed. The analyses of these variables investigate whether AI ethicality affects human teammates' perceptions of their AI teammate in varying contexts and whether trust in the AI can be restored after it committed an ethical violation in these contexts. Better understanding the role of trust repair strategies in this context is essential, given the impact that trust within teams has on team performance outcomes coupled with common trust repair strategies' inability to restore trust after ethical violations in prior research (Schelble, Lopez, et al. 2022; Textor et al. 2022).

4.1. Analysis

Because measures were taken at the individual level as part of teams of three across four separate missions, observations were unlikely to be independent. Using participant and team as multilevel random effects, the intraclass correlation coefficient (ICC) was computed for each dependent variable (McGraw and Wong 1996). The average ICC across all variables was found to be .574, with the lowest being .188 for generators destroyed. Thus, because there was evidence of clustering for each unique combination of participant and team, each dependent variable was analysed using multilevel modelling (Fox 2015). Specifically, all models estimated a random intercept for each unique combination of team and participant (except generators destroyed, which was clustered by team as no individual-level data existed). Unfortunately, there is no

consensus on effect sizes for linear mixed-effects models due to issues stemming from many indices, such as log-likelihood and deviance being negative and not increasing monotonically with the addition of predictors. As such, the marginal pseudo-R^2 suggested by Nakagawa and Schielzeth (2013) was reported as it helps control the common problems of other pseudo-R^2 indices (Nakagawa and Schielzeth 2013).

Because the order of the trust repair strategies was conducted pseudo-randomly, given the number of possible ordering combinations, there was still the potential for order effects despite the randomisation and presentation measures taken to control for them. Therefore, to further mitigate this possibility, all analyses were conducted while controlling for trust repair strategy presentation order. These analyses found trust repair strategy ordering to be a significant covariate only for participants trust in their human teammate. To ensure all model assumptions were met, a residual analysis was conducted for all models (Fox 2015; Rosopa, Schaffer, and Schroeder 2013). The models were tested for significance using Type III Wald χ^2 ANOVA tests. The predictor's ability to improve the model significantly was tested using Akaike's Information Criterion (AIC). Likelihood ratio testing was used whenever two models' AIC values were within two points to confirm the predictor significantly improved the model. All post-hoc tests used Kenward-Roger degrees of freedom and the Tukey method to correct for family-wise error rate.

4.2. Trust and perceived ethicality

Starting with trust among the members of the human-AI team, the following analyses address RQ1, RQ2, and RQ3, which sought to understand how the framing of an ethical violation, individual role, and trust repair strategy influence trust within human-AI teams and the perceived ethicality of the AI teammate. Along these same lines, these analyses tested the associated hypotheses of H1.1, H1.2, H2.1, H2.2, H3.1, and H3.2.

4.2.1. Descriptive statistics

The following tables show the descriptive statistics for participants' trust in their teammates and the ethical rating of their AI teammate.

4.2.2. Trust in the AI teammate

A 2 (Violation Type: Competency, Integrity) x 2 (Team Role: Ground, Surveillance) x 4 (Trust Repair Strategy: None, Apology, Denial, Justification) linear mixed effects model was conducted to assess the effect of AI violation type (between-subjects), team role (between-subjects), and trust repair strategy (within-subjects) on participants' trust in the AI teammate, while controlling for trust repair strategy presentation order. Table 1 shows the descriptive statistics. AIC indicated that each predictor added significantly increased the model's fit compared to the null model (AIC: 1449.91), except for the AI trust repair strategy (AIC: 1450.15) and the two-way interaction effects (AIC: 1449.03), indicating the lack of support for H3.1 and H3.2. Adding violation type (AIC: 1443.77) did improve the model fit, as did adding team role (AIC: 1441.78) to the model.

There was a statistically significant main effect of violation type on trust in the AI teammate ($\chi^2(1) = 9.31$, $p = .002$; see Figure 2), indicating partial support for H1.1. The average trust in the AI teammate was significantly greater when the unethical action was framed as an integrity violation ($M = 22.5$, $SE = 1.18$) compared to when it was framed as a competency violation ($M = 18.3$, $SE = 1.13$). There was also a statistically significant main effect of team role on participants' trust in the AI teammate ($\chi^2(1) = 4.13$, $p = .042$; see Figure 2), providing partial support for H2.1. The results showed participants' trust towards the AI teammate as significantly higher for participants in the ground role ($M = 21.8$, $SE = 1.14$) than the surveillance role ($M = 19.0$, $SE = 1.14$). These two differences in AI trust indicate a change of trust within the moderate range, as trust remained between 3 and 4 out of 5. The marginal model $R^2 = .25$.

Table 1. Descriptive statistics for participants' trust in their AI teammate.

Team Role	Trust Violation	Trust Repair	N	Mean (SD)
Ground	Competency	Apology	15	19.60 (6.14)
		Denial	15	19.27 (5.79)
		Explanation	15	20.60 (5.70)
		No TR	15	21.33 (5.49)
	Integrity	Apology	15	23.00 (6.08)
		Denial	15	23.13 (5.83)
		Explanation	15	22.80 (6.42)
		No TR	15	24.27 (6.11)
Surveillance	Competency	Apology	15	17.80 (8.49)
		Denial	15	15.13 (7.21)
		Explanation	15	15.40 (7.75)
		No TR	15	17.27 (9.22)
	Integrity	Apology	15	22.00 (5.61)
		Denial	15	21.93 (5.38)
		Explanation	15	20.87 (6.52)
		No TR	15	21.67 (6.53)

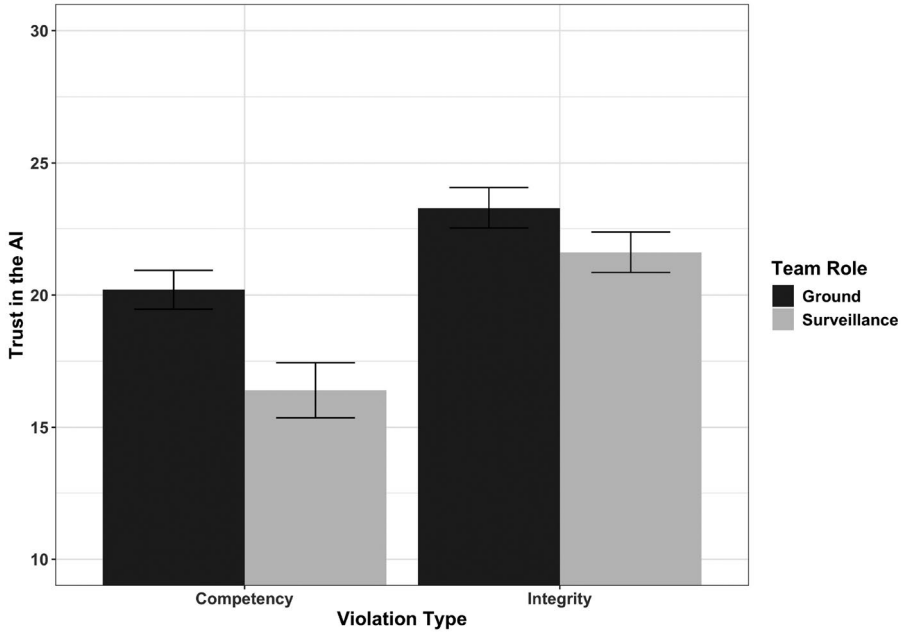

Figure 2. Effect of violation type and team role on trust in the AI teammate. Error bars indicate standard error.

4.2.3. Human teammate trust

A 2 (Violation Type: Competency, Integrity) x 2 (Team Role: Ground, Surveillance) x 4 (Trust Repair Strategy: None, Apology, Denial, Justification) linear mixed effects model was conducted to assess the effect of AI violation type (between-subjects), team role (between-subjects), and trust repair strategy (within-subjects) on participants' trust in their human teammate, while controlling for trust repair strategy presentation order. Table 2 shows the descriptive statistics. Based on the reported AIC after adding each predictor to the model, no main effects improved the model compared to the null (AIC: 1236.59) and, as such, were not included. Dropping these main effects from the model indicates a lack of partial support for H1.1 and H2.1. However, the interaction effects of the independent variables did improve the model fit (AIC: 1230.59) over the null and were subsequently tested.

There was a significant interaction effect between team role and trust repair strategy ($\chi^2(3) = 8.34$, p = .039; see Figure 3); however, the post hoc tests were non-significant {and make the result's potential partial support for H3.1 unclear. There was also a significant interaction effect between violation type and trust repair strategy ($\chi^2(3) = 11.49$, p = .009; see Figure 3), showing partial support for H3.2. The AI teammate that did not use a trust repair strategy resulted in participants having greater trust in their fellow human teammate ($M = 26.2$, $SE = 0.75$)} than the AI teammate using the justification trust repair strategy (M = 24.0, SE = 0.75) but only in the integrity violation condition. Participants trust in their fellow human teammate was also the only dependent variable where the control variable had a significant effect ($\chi^2(14) = 26.84$, p = .02).} {These two differences in human trust indicate a change of trust within the positive trust range, as responses remained between 4 and 5 out of 5 on the six item scale. The marginal model R^2 = .30.

Table 2. Descriptive statistics for participants' trust in their human teammate.

Team Role	Trust Violation	Trust Repair	N	Mean (SD)
Ground	Competency	Apology	15	25.20 (3.69)
		Denial	15	25.60 (2.97)
		Explanation	15	25.73 (3.73)
		No TR	15	25.80 (3.67)
	Integrity	Apology	15	25.67 (3.35)
		Denial	15	27.33 (3.02)
		Explanation	15	25.40 (4.01)
		No TR	15	26.53 (3.04)
Surveillance	Competency	Apology	15	25.80 (3.59)
		Denial	15	24.33 (4.76)
		Explanation	15	25.60 (4.00)
		No TR	15	25.27 (3.65)
	Integrity	Apology	15	23.93 (4.59)
		Denial	15	23.33 (5.90)
		Explanation	15	22.13 (6.33)
		No TR	15	25.40 (3.91)

4.2.4. Ethical rating of the AI teammate

A 2 (Violation Type: Competency, Integrity) x 2 (Team Role: Ground, Surveillance) x 4 (Trust Repair Strategy: None, Apology, Denial, Justification) linear mixed effects model was conducted to assess the effect of AI violation type (between-subjects), team role (between-subjects), and trust repair strategy (within-subjects) on participants ethical rating of the AI teammate, while controlling for trust repair strategy presentation order. Table 3 shows

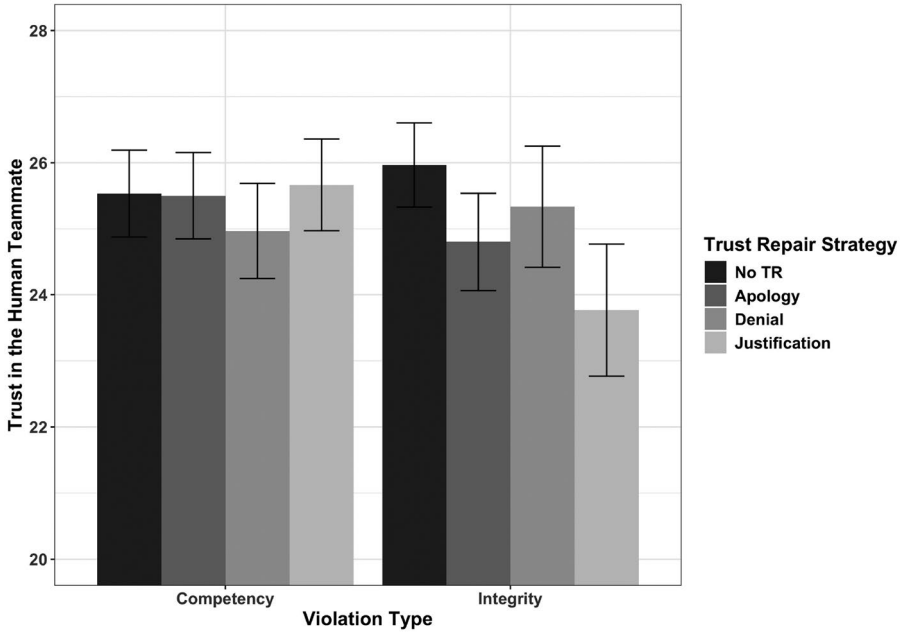

Figure 3. Effect of trust repair strategy and violation type on participants' trust in their fellow human teammate. Error bars indicate standard error.

Table 3. Descriptive statistics for participants' ethical rating of their AI teammate.

Team Role	Trust Violation	Trust Repair	N	Mean (SD)
Ground	Competency	Apology	15	22.53 (10.13)
		Denial	15	21.00 (9.02)
		Explanation	15	21.20 (9.03)
		No TR	15	21.33 (9.85)
	Integrity	Apology	15	27.80 (9.86)
		Denial	15	24.87 (9.75)
		Explanation	15	27.80 (11.14)
		No TR	15	23.67 (10.00)
Surveillance	Competency	Apology	15	17.00 (12.72)
		Denial	15	15.47 (10.99)
		Explanation	15	16.27 (12.38)
		No TR	15	18.80 (9.26)
	Integrity	Apology	15	23.20 (11.71)
		Denial	15	20.27 (11.71)
		Explanation	15	19.67 (11.97)
		No TR	15	22.87 (14.14)

the descriptive statistics. The main effect of trust repair strategy (AIC: 1630.88) did not improve model fit over the null model (AIC: 1631.80) and was not included. However, the team role (AIC: 1629.86) and its combined effect alongside the violation type (AIC: 1627.99) improved the model fit and were included. The interaction effects of the independent variables (AIC: 1629.66) did not improve the model's fit over the null model and were not retained, showing a lack of partial support for H3.1 and H3.2.

There was a statistically significant main effect of violation type on participants ethical rating of the AI teammate ($\chi^2(1) = 4.35$, $p = .037$; see Figure 4), indicating support for H1.2. Participants perceived the AI to be significantly more ethical when it framed violations as an integrity violation ($M = 24.1$, $SE = 2.00$) than a competency violation ($M = 19.2$, $SE = 1.91$). There was also a statistically significant main effect of team role on participants' ethical rating of the AI teammate ($\chi^2(1) = 4.00$, $p = .046$; see Figure 4), showcasing support for H2.2. Specifically, the ground role ($M = 23.9$, $SE = 1.94$) rated the AI teammate as significantly more ethical than the surveillance role ($M = 19.4$, $SE = 1.94$). These differences in AI ethicality rating indicate a change of perceived ethicality from slightly negative to moderate as averaged responses ranged from 3.2/7 on the low end to 3.98/7 and 4.01/7 on the higher end. The marginal model $R^2 = .23$.

4.3. Generators destroyed

This final section addresses RQ4 along with the associated hypotheses of H4.1 and H4.2, which investigate how the number of generators destroyed may be influenced by the framing of the ethical violation and the trust repair strategy used after an ethical violation by an AI teammate.

4.3.1. Descriptive statistics

The following are descriptive statistics for the number of generators destroyed by teams separated by trust violation type and trust repair strategy.

4.3.2. Generators destroyed

A 2 (Violation Type: Competency, Integrity) x 4 (Trust Repair Strategy: None, Apology, Denial, Justification)

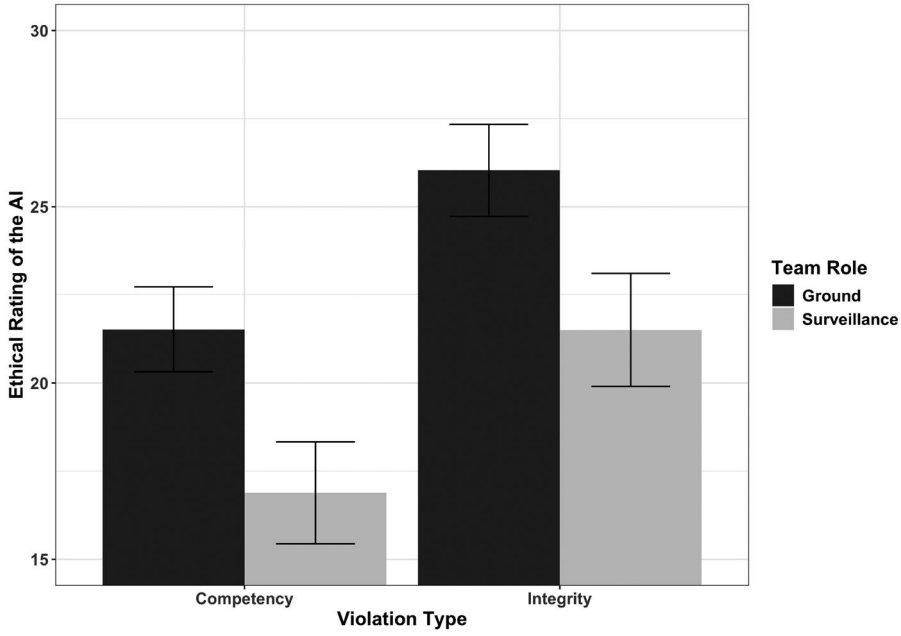

Figure 4. Effect of team role and trust repair strategy on participants' ethical rating of the AI teammate. Error bars indicate standard error.

Table 4. Descriptive statistics for the number of generators destroyed by teams.

Trust Violation	Trust Repair	N	Mean (SD)
Competency	Apology	15	44% (.30)
	Denial	15	76% (.31)
	Explanation	15	47% (.41)
	No TR	15	53% (.34)
Integrity	Apology	15	49% (.36)
	Denial	15	61% (.36)
	Explanation	15	48% (.42)
	No TR	15	61% (.41)

linear mixed effects model was conducted to assess the effect of AI violation type (between-subjects) and trust repair strategy (within-subjects) on generators destroyed by the team, while controlling for trust repair strategy presentation order. Table 4 shows the descriptive statistics. Violation type (AIC: 101.75) and the interaction effect between trust repair strategy and violation type (AIC: 98.68) did not significantly improve model fit over the null model (AIC: 99.80) and were not used, which show a lack of support for H4.2. However, trust repair strategy (AIC: 93.86) significantly improved model fit over the null model and was retained.

Trust repair strategy was found to have a significant main effect on team score ($\chi^2(3) = 12.76$, $p = .005$; see Figure 5), showcasing support for H4.1. When teams worked with an AI teammate issuing a denial trust repair strategy, they destroyed significantly more generators ($M = 70\%$, $SE = .07$) than when teams worked with the AI teammate issuing an apology ($M = 48\%$, $SE = .07$) and justification ($M = 48\%$, $SE = .07$). The marginal model $R^2 = .28$.

5. Discussion

The current study examined the effect of individual team roles, violation type, and trust repair strategy on the number of generators destroyed and perceptions of trust and ethics within the team after an AI teammate commits an ethical violation. First, addressing RQ1, the results showed partial support for H1.1 as participants perceived higher trust in their AI teammate when ethical violations were framed as integrity rather than competency violations, with no effect on trust between the human teammates. H1.2 was also supported as participants rated the same action as more ethical when framed as an integrity violation instead of a competency violation. Concerning RQ2, H2.1 was again partially supported with no effect on participants' trust in their human teammate but their trust in the AI teammate was significantly greater for those in the ground role instead of the surveillance role. H2.2 was also supported and followed the same trend, with the ground role rating the AI teammate as significantly more ethical than the surveillance role. Looking to RQ3, H3.1 was not supported; however, H3.2 was partially supported by a significant interaction effect where trust in the human teammate was higher in the no trust repair condition compared to the justification condition, but only when the AI teammate framed their action as an integrity violation. Lastly, answering RQ4, H4.1 was not supported, but H4.2 was supported by the finding that the denial trust repair strategy was associated with significantly

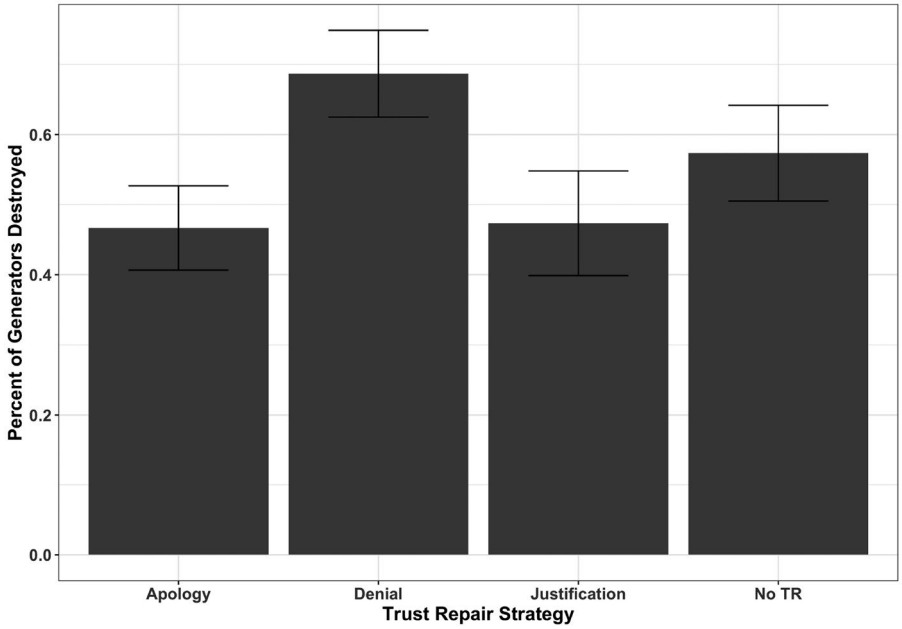

Figure 5. The effect of trust repair strategy and violation type on the number of generators that teams destroyed. Error bars indicate standard error.

more destroyed devices than the apology and justification strategies.

From these results, it is clear that contextual factors *do* play a significant role in how humans perceive the ethicality of their AI teammates and how the trust within these teams develops considering those ethical judgments. While the results concerning trust repair strategies were largely inconclusive, the effect of team role and ethical violation framing were apparent throughout the measures analysed, which the following section will address in the context of existing research.

5.1. Competency violations by AI are particularly harmful in high risk contexts, even in the realm of ethical violations

The knowledge that ethical violations are less damaging to trust when framed as a failure of integrity rather than competence is significant to understanding trust dynamics for AI teammates engaging in ethically charged decision-making within high-risk tasks. This finding also aligns with previous research conducted in non-military tasks (investments) that show competency violations are especially egregious when humans make judgments of trust in their AI teammates (Clark 2018). Examining the two violation types in the context of AI ethics presented a chance for this trend to be reversed. Surprisingly, that reversal was not the case, as the current study indicates that making an operational mistake was considered more egregious than disregarding the civilian non-maleficence directive. While this finding is not definitive, given the nature of the experiment using action-based teams completing a high-risk military-related task with novice participants, such results *do* provide strong insights into novel trust dynamics within ethics-based trust for human-AI teams operating in similar task contexts. For these action-based teams engaging in high-risk tasks, these findings suggest that ethics-based trust in AI teammates is still, in many ways, closely tied to their perceived competence, which is supported by the robust trust in automation literature (Lee and Moray 1992; Lee and See 2004; Muir 1987; Muir 1994). A particular point concerning the current study's task in a high-risk environment is that participants, namely the Ground role, may view their safety in the simulation or the completion of the task as the most important factor. Specifically, the AI teammate being perceived as putting more value on the mission, the potential safety of the team, and the necessity of mission goals than on civilian life. Further, the trust violations studied in the trust repair literature are typically not nearly as severe as the violation of civilian non-maleficence examined in the current study (de Visser, Pak, and Shaw 2018). So, when the AI teammate makes an unnecessary error with significant consequences, the statement attributing the action to a simple mistake may come across as especially hollow. As such, if an AI teammate makes a considerable mistake in a high-risk context of an ethical nature and has no principled reason for committing it, trust in that system is lost more

so than if it had a reason. Interestingly, denial trust repair strategies resulted in the highest performance compared to apologies and justifications, aligning with prior work demonstrating denials excelling in repairing organisational trust despite strong evidence to the contrary (Fuoli, van de Weijer, and Paradis 2017). It is likely that including an explanation along with an extended experience demonstrating changed behaviour would be capable of repairing trust in an AI teammate, and this has been demonstrated in human-AI teams before (Kox et al. 2021), though it is unclear if this extends to ethical violations. Having one foot in the realm of a machine and the other in the teammate realm places complex expectations on AI teammates that are not always congruent with one another (McNeese et al. 2023), which makes this issue even more challenging to tackle. This knowledge also raises the importance of ethics in AI development, as a breakdown in those relationships can potentially harm the overall team's performance (McNeese et al. 2021).

Aligning with the trust results, when the AI framed its unethical actions as integrity violations, it received higher trust ratings compared to framing them as competency failures. This presents an interesting dynamic between ethical violations, violation type, perceived ethicality, and trust. Specifically, previous research in high-risk military contexts demonstrated that an AI teammate's unethical actions lower trust in an AI teammate, establishing a link between ethicality and trust (Schelble, Lopez, et al. 2022). Similarly, a positive relationship between perceived ethicality and an AI's trustworthiness has been shown (Textor et al. 2022). The current study extends this finding to include violation type having a measurable impact on perceived AI teammate ethicality, which are both contextual factors detailed within the integrative model of trust (Hancock et al. 2023; Mayer, Davis, and Schoorman 1995). Specific to the task context, a military simulation with lethal consequences, it is likely that the violation type even had an outsized influence, given the high-risk stakes nature of the scenario. Related to trust repair, it is known that contextual factors determine which trust violation types are most effective for trust repair (de Visser, Pak, and Shaw 2018), and while the current study did not find strong results for trust repair strategies, it is clear that violation type factors into evaluations of trust and ethics. The instantiation of a competency-based trust violation for an unethical action by an AI teammate is seen as particularly egregious and is likely exacerbated by the effect of automation bias (Mosier et al. 1996). The presence of an AI teammate implies the presence of a system with a high level of reliability and consistency, and when that expectation is not met by the system's own admission, it comes across as particularly unethical. Further, that leaves the question open to where that loss in perceived ethicality is being directed, towards the AI teammate, its developer, or the leadership individual placing it in the position to make an unethical decision is a question on the minds of many in those working in real military high-risk teams (Lopez et al. 2023). Practically speaking, trust and usage metrics should be examined following a perceived ethical violation in a high-risk context across these parties. While the current study specifically measured the perceived ethicality of the AI teammate, it is likely there is additional variance to be captured in this metric by evaluating perceptions towards other responsible stakeholders, including the participants themselves. These findings thus highlight the need to develop additional research to understand how these perceptions evolve in the face of ethical mistakes, imperfect autonomy, and complex scenarios.

5.2. Individual team roles can significantly influence the consequences of ethical violations by AI teammates in high-risk action-based teams

The role from which participants completed the task significantly influenced their judgments of trust and ethicality towards the AI teammates. Specifically, the participants who took on the Ground role in the experiment rated the AI teammate as more ethical and trustworthy than those who interacted with it from the Surveillance role. This finding implies that the specific team role assigned to participants may contribute to different ethical evaluations with AI teammates, and these evaluations, in turn, influence the level of trust in high-risk action-based teaming contexts (i.e. military search and destroy). This knowledge may impact ethical AI teammate design, as it demonstrates the variable nature of ethical perceptions based on individual experiences, risk, and interdependence in role responsibilities while working towards a shared team goal (Alder and Guidice 2010; Singer, Mitchell, and Turner 1998). The high-risk action-based team military context can help explain the disparity in the participants' trust and perceived ethicality of the AI teammate by examining the differences between the two roles. Specifically, the Ground role bore less direct consequence on the decisions made by the AI teammate in the Aerial role compared to the Surveillance role. The primary interaction between the AI and the Ground role was for the second phase of each mission when the Ground role needed to wait for the AI teammate's message to

enter the town. Additionally, because the Ground role was the only one required to physically enter the town, their success, including their virtual character's survival, was directly tied to the effectiveness of the AI teammate's actions. Alternatively, those fulfilling the Surveillance role were told that their AI teammate would base its decision on what action to perform to clear the town partially on the intel they submitted to the team. Therefore, it is probable that participants fulfilling the Surveillance role feel responsible for the outcome of the AI teammate's action. This finding emphasises the need to closely consider the roles AI teammates are being deployed alongside to understand how those roles may interpret ethical violations, though it is likely that this effect is stronger in physical environments and weaker in task contexts that do not share the same level of risk and interdependence. Further, these results may appear different for older adults as the aspects of the Ground role may not make the AI teammate's actions appear as justifiable given their tendency to adhere to stricter ethical standards (Peterson, Rhoads, and Vaught 2001; Ruegger and King 1992). A practical implication of this finding is that AI teammates should tailor their communication and transparency strategies to the individual roles of their human partners. By incorporating role-sensitive interaction mechanisms, AI teammates can engender more effective trust and shared understanding, thereby enhancing collaboration and improving overall team performance.

The current study supports the assertion that ethics-based trust can differ from other forms of trust in high-risk action-based team tasks. As suggested in Schelble and colleagues' (2022) work, the present study's findings highlight that, in high-risk action-based contexts, ethics-based trust may represent a unique trust violation type based on its distinct standing within the construct of trust, as described in several theoretical trust frameworks (Hancock et al. 2023; Malle and Ullman 2021; Mayer, Davis, and Schoorman 1995). Conventionally, trust violations are often associated with failures in automation performance (e.g (Lee and Moray 1992; Lee and See 2004).). Trust repair strategies in this context have significantly restored human trust in an autonomous system (Kox et al. 2021; Quinn, Pak, and de Visser 2017). By contrast, the results of the current study found no significant effect of any trust repair strategy on participants' trust in the AI teammate following an ethical violation, exemplifying that ethics-based trust is likely more nuanced in these high-risk contexts (e.g. military, search and rescue, medical). Consequently, trust repair strategies that primarily focus on rectifying performance malfunctions may not be effective when trust violations result from unethical behaviour that still enables the team to achieve the primary shared goal. Specifically, a trust repair strategy should address a failure to accomplish the shared goal in a way that does not align with the ethical values of the individual and the shared ethics of the overall team (Flathmann et al. 2021). Essentially, participants may have trusted the AI teammate to help the team accomplish their shared goal but did not trust it to do so in a manner that aligned with their ethical values. Future trust repair strategies, regardless of the context, should consider the idea of ensuring AI teammates' ethical values align with the individual and team they operate alongside, as many frameworks already have detailed (Berretta et al. 2023; Xu et al. 2025). However, these AI teammates should be capable of addressing their eventual ethical violations effectively by addressing those values directly through improved trust repair strategies that encourage more accurately calibrated trust.

5.3. Limitations and future research

The current study contributes value to understanding trust and ethics within human-AI teaming; however, some limitations must be considered when interpreting these results. First, we applied a trust repair strategy immediately after an ethical violation. However, studies have shown that apologies made at the subsequent decision opportunity were more successful in repairing trust than those immediately following a violation (Robinette, Howard, and Wagner 2015; Robinette, Howard, and Wagner 2017). Furthermore, the study utilised a synthetic task environment where participants' stake in the task was also artificial, and as such, the size of these effects may differ in practice. The sample size also limits the potential generalisability of the study's results, as a larger sample including more participants would result in potentially more reliable results. The study's sample also consists mainly of young adults who have shown a propensity to be less ethical than older individuals (Peterson, Rhoads, and Vaught 2001; Ruegger and King 1992), which means these results may not be a one-to-one transfer for teams working mainly with adults over 40.

While trust repair strategies were not successful at repairing trust that was dampened by ethical violations in our study, these findings may vary based on the severity of the ethical violation since the violation used in the current study (civilian non-maleficence) is particularly egregious (Reed et al. 2016; Textor et al. 2022). Therefore, further studies should investigate whether

the timing of different trust repair strategies can affect trust and perceived ethics in the case of ethical violations and whether the severity of the violation influences the success of trust repair strategies, with a similar limitation applying to non-military tasks, non-action teams, and lower risk tasks in general. The teams' evaluation metric consisted only of the number of generators destroyed by the team in the allotted time and was not inclusive of the level of collateral damage inflicted in the pursuit of this objective. As such, the evaluation of performance should not be taken as a one-to-one comparison to real-world performance metrics, which would surely account for collateral damage. This method was necessary only to ensure each condition had the exact same experience with the AI teammate regarding ethically charged action. Lastly, there is research showing that trust repair strategies can be less effective as violations are repeated, even if it isn't the same strategy each time (Esterwood and Robert 2023; Pak and Rovira 2023). As such, additional studies examining these strategies as between-subjects factors should be coupled with studies examining them within-subjects.

Lastly, the current study utilised partial Latin squares counterbalancing and presented each within-subjects level as a new experience with a new AI teammate with a different name to help control for carry-over effects. However, there is research stating that trust repair strategies can be less effective as they are repeated, even if it isn't the same strategy each time (Esterwood and Robert 2023; Pak and Rovira 2023). As such, additional studies examining these strategies between-subjects should accompany studies examining them within-subjects.

6. Conclusion

As the extended use and deployment of human-AI teams gain momentum, critical questions 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 in the face of potential ethically-based trust violations. While investigations explore the interplay of ethics and trust in the human-AI teaming context (Schelble, Lopez, et al. 2022; Textor et al. 2022), it becomes essential to explore this relationship at a deeper level to include the contextual factors influencing trust. Towards that end, the current study posits pertinent insights for human-AI teams operating in high-risk action-based tasks. Stemming from the integrative model of trust (Hancock et al. 2023; Mayer, Davis, and Schoorman 1995), the differences in risk, interdependencies, and interaction between team roles were shown to affect participants' perceived trust and ethicality towards the AI teammate. Further still, the study found competency violations were viewed less favourably than integrity trust violations despite the same unethical AI teammate action being judged. Lastly, trust repair strategies had a relatively negligible impact on human-AI teams, with a small interaction effect, and the denial trust repair strategy resulted in teams destroying more generators than the justification and apology conditions. These results emphasise just how task interdependence between roles and the contextual reasoning *why* an AI committed an ethical violation can impact trust in ethically charged situations for human-AI teams operating in high-risk action-based tasks. Such an assertion gives credence to recent theoretical frameworks on trust in human-AI teams (Flathmann et al. 2021; Hancock et al. 2023; Malle and Ullman 2021) and human-centered AI more generally (Berretta et al. 2023; Xu and Gao 2024). These findings contribute directly to the efforts to develop more ethical AI systems by improving the understanding of how ethics and trust interact with the dynamic relationships in human-AI teams for future research as they relate to high-risk action-based tasks.

Author contributions

CRediT: Beau G Schelble: Conceptualisation, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualisation, Writing – original draft, Writing – review & editing; Claire Textor: Conceptualisation, Investigation, Methodology, Writing – original draft, Writing – review & editing; Rui Zhang: Conceptualisation, Investigation, Methodology, Writing – original draft, Writing – review & editing; Jeremy Lopez: Conceptualisation, Writing – original draft, Writing – review & editing; Noah Taverez: Investigation, Writing – original draft, Writing – review & editing; Connie Ku: Investigation, Writing – original draft, Writing – review & editing; Nathan McNeese: Conceptualisation, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing; Richard Pak: Conceptualisation, Formal Analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing; Guo Freeman: Conceptualisation, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing; Chad Tossell: Conceptualisation, Methodology, Writing – review & editing; Ewart de Visser: Conceptualisation, Formal Analysis, Methodology, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by AFOSR Award FA9550-20-1-0342 (Program Manager: Laura Steckman).

Data availability statement

The data that support the findings of this study are available on request from the corresponding author, BS. The data are not publicly available as they contain information that could compromise the privacy of research participants.

References

Alarcon, Gene M., August Capiola, Michael A. Lee, and Sarah A. Jessup. 2022. "The Effects of Trustworthiness Manipulations on Trustworthiness Perceptions and Risk-Taking Behaviors." *Decision* 9 (4): 388–406. doi:10.1037/dec0000189.

Alder, G Stoney, and Rebecca M. Guidice. 2010. "The Ethics of Bluffing: The Effects of Individual Differences on Perceived Ethicality and Bluffing Behavior." *Journal of Business & Leadership: Research, Practice, and Teaching (2005-2012)* 6 (1): 10–24.

Ayoub, Jackie, Lilit Avetisyan, Mustapha Makki, and Feng Zhou. 2022. "An Investigation of Drivers' Dynamic Situational Trust in Conditionally Automated Driving." *IEEE Transactions on Human-Machine Systems* 52 (3): 501–511. https://ieeexplore.ieee.org/abstract/document/9653756/?casa_token=ag3w22DUYyAAAAAA:ItkHyBBDhIzfYZzzmRnP1cKsJFkG1uWtaQw1Tj10auKo1W2m09WZKO4znsb_0QSi2yxB6VIMcg. doi:10.1109/THMS.2021.3131676.

Balona, Caesar. 2024. "ActuaryGPT: Applications of Large Language Models to Insurance and Actuarial Work." *British Actuarial Journal* 29: e15. doi:10.1017/S1357321724000102.

Banks, Jaime. 2019. "A Perceived Moral Agency Scale: Development and Validation of a Metric for Humans and Social Machines." *Computers in Human Behavior* 90: 363–371. doi:10.1016/j.chb.2018.08.028.

Bergman, Ronen, and Farnaz Fassihi. 2021. "The Scientist and the AI-Assisted, Remote-Control Killing Machine." The New York Times 18.

Berretta, Sophie, Alina Tausch, Greta Ontrup, Björn Gilles, Corinna Peifer, and Annette Kluge. 2023. "Defining human-AI Teaming the Human-Centered Way: A Scoping Review and Network Analysis." *Frontiers in Artificial Intelligence* 6: 1250725. https://www.frontiersin.org/articles/10.3389/frai.2023.1250725/full. doi:10.3389/frai.2023.1250725.

Bigman, Yochanan E., and Kurt Gray. 2018. "People Are Averse to Machines Making Moral Decisions." *Cognition* 181: 21–34. https://www.sciencedirect.com/science/article/pii/S0010027718302087 Publisher: Elsevier. doi:10.1016/j.cognition.2018.08.003.

Bonde, Sheila, Paul Firenze, J. Green, M. Grinberg, J. Korijn, E. Levoy, A. Naik, L. Ucik, and L. Weisberg. 2013. "A Framework for Making Ethical Decisions." *Science and Technology Studies. Brown University*.

Butler, John K., Jr., and R Stephen Cantrell. 1984. "A Behavioral Decision Theory Approach to Modeling Dyadic Trust in Superiors and Subordinates." *Psychological Reports* 55 (1): 19–28. doi:10.2466/pr0.1984.55.1.19.

Chiou, Erin K., and John D. Lee. 2023. "Trusting Automation: Designing for Responsivity and Resilience." *Human Factors* 65 (1): 137–165. doi:10.1177/00187208211009995.

Choudhury, L. M. R., Alia Aoun, Dina Badaway, Luis Antonio de Alburquerque Bacardit, Yassine Marjane, and Adrian Wilkinson. 2021. "Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)." *United Nations Security Council, New York, NY, USA*.

Chowdhury, Mashrur, and Adel W. Sadek. 2012. "Advantages and Limitations of Artificial Intelligence." *Artificial Intelligence Applications to Critical Transportation Issues* 6 (3): 360–375.

Clark, Tiffany. 2018. "Integrity-Based Trust Violations within Human-Machine Teaming.". Ph.D. Dissertation. Monterey, CA; Naval Postgraduate School.

Colquitt, Jason A., Brent A. Scott, and Jeffery A. LePine. 2007. "Trust, Trustworthiness, and Trust Propensity: A Meta-Analytic Test of Their Unique Relationships with Risk Taking and Job Performance." *The Journal of Applied Psychology* 92 (4): 909–927. https://psycnet.apa.org/fulltext/2007-09571-002.html. doi:10.1037/0021-9010.92.4.909.

de Visser, Ewart J., 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): 459–478. doi:10.1007/s12369-019-00596-x.

de Visser, Ewart J., Richard Pak, and Tyler H. Shaw. 2018. "From 'automation' to 'autonomy': The Importance of Trust Repair in Human-Machine Interaction." *Ergonomics* 61 (10): 1409–1427. doi:10.1080/00140139.2018.1457725.

DeLone, William, J. Alberto Espinosa, Gwanhoo Lee, and Erran Carmel. 2005. "Bridging Global Boundaries for is Project Success." In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 48b–48b. IEEE.

Demir, Mustafa, 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): 258–267. doi:10.1109/TETCI.2018.2829985.

Doris, John M. 1998. "Persons, Situations, and Virtue Ethics." *Noûs* 32 (4): 504–530. doi:10.1111/0029-4624.00136.

Driskell, Tripp, James E. Driskell, C Shawn Burke, and Eduardo Salas. 2017. "Team Roles: A Review and Integration." *Small Group Research* 48 (4): 482–511. doi:10.1177/1046496417711529.

Esterwood, Connor, and Lionel P. Robert. 2021. "Do You Still Trust Me? Human-Robot Trust Repair Strategies." In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 183–188. doi:10.1109/RO-MAN50785.2021.9515365.

Esterwood, Connor, and Lionel P. Robert. 2022a. "Having the Right Attitude: How Attitude Impacts Trust Repair in Human—Robot Interaction." In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 332–341. IEEE.

Esterwood, Connor, and Lionel P. Robert. 2022b. "A Literature Review of Trust Repair in HRI." In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1641–1646. IEEE.

Esterwood, Connor, and Lionel P. Robert. Jr. 2023. "Three Strikes and You Are out!: The Impacts of Multiple Human–Robot Trust Violations and Repairs on Robot Trustworthiness." *Computers in Human Behavior* 142: 107658. doi:10.1016/j.chb.2023.107658.

Flathmann, Christopher, 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.

Flathmann, Christopher, Nathan J. McNeese, and Thomas A. O'Neill. 2025. "Designing High-Impact Experiments for Human–Autonomy / AI Teaming." *Journal of Cognitive Engineering and Decision Making*: 15553434251327697. doi:10.1177/15553434251327697.

Fox, John. 2015. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications.

Fuoli, Matteo, Joost van de Weijer, and Carita Paradis. 2017. "Denial Outperforms Apology in Repairing Organizational Trust despite Strong Evidence of Guilt." *Public Relations Review* 43 (4): 645–660. https://www.sciencedirect.com/science/article/pii/S0363811117300243?casa_token=f1YGf8ugRbUAAAAA:SuL9bfMnbo26ydqt9jp3gP1ZaiyG8OIanNUUX-urMN8vQmdoXpDd1PW4ctqTXij5Du7SVhlMypg Publisher: Elsevier. doi:10.1016/j.pubrev.2017.07.007.

Georganta, Eleni, and Anna-Sophie Ulfert. 2024. "My Colleague is an AI! Trust Differences between AI and Human Teammates. *Team Performance Management*." *An International Journal* 30 (1/2): 23–37.

Grimm, David A., Mustafa Demir, Jamie C. Gorman, and Nancy J. Cooke. 2018. "Team Situation Awareness in Human-Autonomy Teaming: A Systems Level Approach." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 62, 149–149. Los Angeles, CA: SAGE Publications Sage CA.

Hancock, PA., Theresa T. Kessler, Alexandra D. Kaplan, Kimberly Stowers, J Christopher Brill, Deborah R. Billings, Kristin E. Schaefer, and James L. Szalma. 2023. "How and Why Humans Trust: A Meta-Analysis and Elaborated Model." *Frontiers in Psychology* 14: 1081086. doi:10.3389/fpsyg.2023.1081086.

Hancock, Peter A., 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): 517–527.

Hauptman, Allyson I., Rohit Mallick, Christopher Flathmann, and Nathan J. McNeese. 2025. "Human Factors Considerations for the Context-Aware Design of Adaptive Autonomous Teammates." *Ergonomics* 68 (4): 571–587. doi:10.1080/00140139.2024.2380341.

Hoff, Kevin Anthony, and Masooda Bashir. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust." *Human Factors* 57 (3): 407–434. Publisher: SAGE Publications Inc. doi:10.1177/0018720814547570.

Hursthouse, Rosalind, and Glen Pettigrove. 2003. "Virtue Ethics."

Juvina, Ion, Michael G. Collins, Othalia Larue, William G. Kennedy, Ewart De Visser, and Celso De Melo. 2019. "Toward a Unified Theory of Learned Trust in Interpersonal and Human-Machine Interactions." *ACM Transactions on Interactive Intelligent Systems* 9 (4): 1–33. doi:10.1145/3230735.

Kelley, John F. 2018. "Wizard of Oz (WoZ) a Yellow Brick Journey." *Journal of Usability Studies* 13 (3): 119–124.

Kim, Peter H., Donald L. Ferrin, Cecily D. Cooper, and Kurt T. Dirks. 2004. "Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence-Versus Integrity-Based Trust Violations." *The Journal of Applied Psychology* 89 (1): 104–118. Place: US Publisher: American Psychological Association. doi:10.1037/0021-9010.89.1.104.

Kim, Peter H., Kurt T. Dirks, Cecily D. Cooper, and Donald L. Ferrin. 2006. "When More Blame is Better than Less: The Implications of Internal vs. External Attributions for the Repair of Trust after a Competence-vs. Integrity-Based Trust Violation." *Organizational Behavior and Human Decision Processes* 99 (1): 49–65. doi:10.1016/j.obhdp.2005.07.002.

Kox, Esther S., José H. Kerstholt, Tom F. Hueting, and Peter W. de Vries. 2021. "Trust Repair in Human-Agent Teams: The Effectiveness of Explanations and Expressing Regret." *Autonomous Agents and Multi-Agent Systems* 35 (2): 30. doi:10.1007/s10458-021-09515-9.

Kramer, Roderick M., and Tom R. Tyler. 1996. *Trust in Organizations: Frontiers of Theory and Research*. SAGE. Google-Books-ID: Lo85DQAAQBAJ.

Langer, Markus, Cornelius J. König, Caroline Back, and Victoria Hemsing. 2023. "Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias." *Journal of Business and Psychology* 38 (3): 493–508. doi:10.1007/s10869-022-09829-9.

Lee, John D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80. doi:10.1518/hfes.46.1.50_30392.

Lee, John, and Neville Moray. 1992. "Trust, Control Strategies and Allocation of Function in Human-Machine Systems." *Ergonomics* 35 (10): 1243–1270. doi:10.1080/00140139208967392.

Lopez, Jeremy, Claire Textor, Caitlin Lancaster, Beau Schelble, Guo Freeman, Rui Zhang, Nathan McNeese, and Richard Pak. 2023. "The Complex Relationship of AI Ethics and Trust in Human–AI Teaming: insights from Advanced Real-World Subject Matter Experts." *AI and Ethics* : 1–21.

Lumineau, Fabrice. 2017. "How Contracts Influence Trust and Distrust." *Journal of Management* 43 (5): 1553–1577. doi:10.1177/0149206314556656.

Lyons, Joseph B., Katia Sycara, Michael Lewis, and August Capiola. 2021. "Human–Autonomy Teaming: Definitions, Debates, and Directions." *Frontiers in Psychology* 12: 589585. doi:10.3389/fpsyg.2021.589585.

Mach, Merce, 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): 771–794. doi:10.1348/096317909X473903.

Malle, Bertram F., and Daniel Ullman. 2021. "A Multidimensional Conception and Measure of Human-Robot Trust." In *Trust in Human-Robot Interaction*, 3–25. Elsevier. https://www.sciencedirect.com/science/article/pii/B9780128194720000010.

Malle, Bertram F., and Elizabeth Phillips. 2023. "A Robot's Justifications, but Not Explanations, Mitigate People's Moral Criticism and Preserve Their Trust."

Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. "An Integrative Model of Organizational Trust."

The Academy of Management Review 20 (3): 709–734. Publisher: Academy of Management. doi:10.2307/258792.

McGraw, Kenneth O., and Seok P. Wong. 1996. "Forming Inferences about Some Intraclass Correlation Coefficients." *Psychological Methods* 1 (1): 30–46. doi:10.1037/1082-989X.1.1.30.

McNeese, Nathan J., Christopher Flathmann, Thomas A. O'Neill, and Eduardo Salas. 2023. "Stepping out of the Shadow of Human-Human Teaming: Crafting a Unique Identity for Human-Autonomy Teams." *Computers in Human Behavior* 148: 107874. doi:10.1016/j.chb.2023.107874.

McNeese, Nathan J., 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): 51–72. doi:10.1080/10864415.2021.1846854.

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

Mercado, Joseph E., Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. "Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management." *Human Factors* 58 (3): 401–415. doi:10.1177/0018720815621206.

Mosier, Kathleen L., Linda J. Skitka, Mark D. Burdick, and Susan T. Heers. 1996. "Automation Bias, Accountability, and Verification Behaviors." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 40, 204–208. Los Angeles, CA: SAGE Publications Sage CA.

Muir, Bonnie M. 1987. "Trust between Humans and Machines, and the Design of Decision Aids." *International Journal of Man-Machine Studies* 27 (5-6): 527–539. doi:10.1016/S0020-7373(87)80013-5.

Muir, Bonnie M. 1994. "Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems." *Ergonomics* 37 (11): 1905–1922. doi:10.1080/00140139408964957.

Nakagawa, Shinichi, and Holger Schielzeth. 2013. "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models." *Methods in Ecology and Evolution* 4 (2): 133–142. doi:10.1111/j.2041-210x.2012.00261.x.

O'Neill, Thomas, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. "Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature." *Human Factors* 64 (5): 904–938.

Pak, Richard, and Ericka Rovira. 2023. "A Theoretical Model to Explain Mixed Effects of Trust Repair Strategies in Autonomous Systems." *Theoretical Issues in Ergonomics Science* : 1–21.

Parasuraman, Raja, and Christopher A. Miller. 2004. "Trust and Etiquette in High-Criticality Automated Systems." *Communications of the ACM* 47 (4): 51–55. doi:10.1145/975817.975844.

Parasuraman, Raja, and Victor Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39 (2): 230–253. doi:10.1518/001872097778543886.

Parikh, Ravi B., Stephanie Teeple, and Amol S. Navathe. 2019. "Addressing Bias in Artificial Intelligence in Health Care." *Jama* 322 (24): 2377–2378. doi:10.1001/jama.2019.18058.

Peterson, Dane, Angela Rhoads, and Bobby C. Vaught. 2001. "Ethical Beliefs of Business Professionals: A Study of Gender, Age and External Factors." *Journal of Business Ethics* 31 (3): 225–232. https://idp.springer.com/authorize/casa?redirect_uri= https://link.springer.com/article/10.1023/a:1010744927551&casa_token=KgUVdwDqOuUAAAAA:kBtsr96ODWms4APybTXYfXFMhsXAejjadtjg8QjuVMJNp1FUsxunJmk7wBjeLo1FIerNHk3C6EWzCAfuWQPublisher:Springer. doi:10.1023/A:1010744927551.

Quinn, Daniel B., 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, 1794–1798. Los Angeles, CA: SAGE Publications Sage CA.

Reed, Gregory S., 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): 195–211.

Reeves, Byron, and Clifford Nass. 1996. "The Media Equation: How People Treat Computers, Television, and New Media like Real People." *Cambridge, UK* 10 (10): 19–36.

Riek, Laurel D. 2012. "Wizard of oz Studies in Hri: A Systematic Review and New Reporting Guidelines." *Journal of Human-Robot Interaction* 1 (1): 119–136.

Robinette, Paul, Ayanna M. Howard, and Alan R. Wagner. 2015. "Timing Is Key for Robot Trust Repair." In *Social Robotics (Lecture Notes in Computer Science)*, edited by Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi, 574–583. Cham: Springer International Publishing. doi:10.1007/978-3-319-25554-5_57.

Robinette, Paul, Ayanna M. Howard, and Alan R. Wagner. 2017. "Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations." *IEEE Transactions on Human-Machine Systems* 47 (4): 425–436. doi:10.1109/THMS.2017.2648849.

Rosopa, Patrick J., Meline M. Schaffer, and Amber N. Schroeder. 2013. "Managing Heteroscedasticity in General Linear Models." *Psychological Methods* 18 (3): 335–351. doi:10.1037/a0032553.

Ruegger, Durwood, and Ernest W. King. 1992. "A Study of the Effect of Age and Gender upon Student Business Ethics." *Journal of Business Ethics* 11 (3): 179–186. https://idp.springer.com/authorize/casa?redirect_uri= https://link.springer.com/article/10.1007/Bf00871965&casa_token=vxN1fjKrN2QAAAAA:picNuZ9naO82xdUp-NBn1JQ-x0CJG4K8ctIQeXT1emhPf6S2QM33CLI3zwaLs3HqqyHJDSFWS2YaItAtVQPublisher:Springer. doi:10.1007/BF00871965.

Sarker, Iqbal H., Helge Janicke, Mohammad Nazeeruddin, Watters Paul, and Nepal Surya. 2023. "AI Potentiality and Awareness: A Position Paper from the Perspective of human-AI Teaming in Cybersecurity.". In *International Conference on Intelligent Computing & Optimization*, 140–149. Springer.

Schelble, Beau G., Caitlin Lancaster, Wen Duan, Rohit Mallick, Nathan J. McNeese, and Jeremy Lopez. 2023. "The Effect of AI Teammate Ethicality on Trust Outcomes and Individual Performance in Human-AI Teams". In *Hawaii International Conference on System Sciences 2023*. 322–331. https://hdl.handle.net/10125/102668.

Schelble, Beau G., 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): 1–29. doi:10.1145/3492832.

Schelble, Beau G., Christopher Flathmann, Nathan J. McNeese, Thomas O'Neill, Richard Pak, and Moses Namara. 2023. "Investigating the Effects of Perceived Teammate Artificiality on Human Performance and Cognition." *International Journal of Human–Computer Interaction* 39 (13): 2686–2701. doi:10.1080/10447318.2022.2085191.

Schelble, Beau G., 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* 66 (4): 1037–1055. doi:10.1177/00187208221116952.

Schmager, Stefan, Ilias O. Pappas, and Polyxeni Vassilakopoulou. 2025. "Understanding Human-Centred AI: A Review of Its Defining Elements and a Research Agenda." *Behaviour & Information Technology* : 1–40.

Schmutz, Jan B., Neal Outland, Sophie Kerstan, Eleni Georganta, and Anna-Sophie Ulfert. 2024. "AI-Teaming: Redefining Collaboration in the Digital Era." *Current Opinion in Psychology* 101837. https://www.sciencedirect.com/science/article/pii/S2352250X24000502.

Scholz, David D., Johannes Kraus, and Linda Miller. 2025. "Measuring the Propensity to Trust in Automated Technology: Examining Similarities to Dispositional Trust in Other Humans and Validation of the PTT-A Scale." *International Journal of Human–Computer Interaction* 41 (2): 970–993. doi:10.1080/10447318.2024.2307691.

Sikula, Andrew, and Adelmiro D. Costa. 1994. "Are Age and Ethics Related?" *The Journal of Psychology* 128 (6): 659–665. doi:10.1080/00223980.1994.9921294.

Singer, Ming, Sarah Mitchell, and Julie Turner. 1998. "Consideration of Moral Intensity in Ethicality Judgements: Its Relationship with Whistle-Blowing and Need-for-Cognition." *Journal of Business Ethics* 17 (5): 527–541. doi:10.1023/A:1005765926472.

Stewart, Greg L., Ingrid S. Fulmer, and Murray R. Barrick. 2005. "An Exploration of Member Roles as a Multilevel Linking Mechanism for Individual Traits and Team Outcomes." *Personnel Psychology* 58 (2): 343–365. doi:10.1111/j.1744-6570.2005.00480.x.

Textor, Claire, 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): 252–281. doi:10.1177/15553434221113964.

Ullman, Daniel, and Bertram F. Malle. 2018. "What Does It Mean to Trust a Robot?: Steps Toward a Multidimensional Measure of Trust.". In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 263–264. Chicago IL USA: ACM, doi:10.1145/3173386.3176991.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. 2024. "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis." *Nature Human Behaviour* 8 (12): 2293–2303. https://www.nature.com/articles/s41562-024-02024-1. Publisher: Nature Publishing Group UK London. doi:10.1038/s41562-024-02024-1.

Wilson, H James, and Paul R. Daugherty. 2018. "Collaborative Intelligence: Humans and AI Are Joining Forces." *Harvard Business Review* 96 (4): 114–123.

Xu, Wei, and Zaifeng Gao. 2024. "Applying HCAI in Developing Effective Human-AI Teaming: A Perspective from Human-AI Joint Cognitive Systems." *Interactions* 31 (1): 32–37. doi:10.1145/3635116.

Xu, Xinran, Ruifeng Yu, Minhui Yuan, and Jingyue Zheng. 2025. "Bidirectional Transparency in Human-Agent Communications: effects of Direction and Level of Transparency." *Ergonomics*: 1–19. (doi:10.1080/00140139.2025.2456535.