Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming

Beau G. Schelble, Human-Centered Computing, Clemson University, Clemson, SC, USA, **Jeremy Lopez**, Department of Psychology, Clemson University, Clemson, SC, USA, **Claire Textor**, Department of Psychology, Clemson University, Clemson, SC, USA, **Rui Zhang**, Human-Centered Computing, Clemson University, Clemson, SC, USA, **Nathan J. McNeese**, Human-Centered Computing, Clemson University, Clemson, SC, USA, **Richard Pak**, Department of Psychology, Clemson University, Clemson, SC, USA, **Guo Freeman**, Human-Centered Computing, Clemson University, Clemson, SC, USA

Objective: Determining the efficacy of two trust repair strategies (apology and denial) for trust violations of an ethical nature by an autonomous teammate.

Background: While ethics in human-AI interaction is extensively studied, little research has investigated how decisions with ethical implications impact trust and performance within human-AI teams and their subsequent repair.

Method: Forty teams of two participants and one autonomous teammate completed three team missions within a synthetic task environment. The autonomous teammate made an ethical or unethical action during each mission, followed by an apology or denial. Measures of individual team trust, autonomous teammate trust, human teammate trust, perceived autonomous teammate ethicality, and team performance were taken.

Results: Teams with unethical autonomous teammates had significantly lower trust in the team and trust in the autonomous teammate. Unethical autonomous teammates were also perceived as substantially more unethical. Neither trust repair strategy effectively restored trust after an ethical violation, and autonomous teammate ethicality was not related to the team score, but unethical autonomous teammates did have shorter times.

Conclusion: Ethical violations significantly harm trust in the overall team and autonomous teammate but do not negatively impact team score. However, current trust repair strategies like apologies and denials appear ineffective in restoring trust after this type of violation.

Application: This research highlights the need to develop trust repair strategies specific to human-AI teams and trust violations of an ethical nature.

Keywords: team collaboration, artificial intelligence, human-computer interaction, autonomous agents, trust in automation

Address correspondence to Beau G. Schelble, Human-Centered Computing, Clemson University, 821 McMillan Rd, Clemson, SC 29631, USA; email: bschelb@g.clemson.edu

HUMAN FACTORS
Vol. 0, No. 0, ■■ ■, pp. 1-19
DOI:10.1177/00187208221116952
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Autonomous manufacturing, military, and healthcare systems have led to more frequent and complex human-artificial intelligence (AI) interactions. Human-AI teaming, in particular, is a significant example of how autonomous technologies fulfill a new role in our world (McNeese et al., 2018, 2021b). Human-AI teams are characterized by at least one human and one autonomous agent where the autonomous agent has a significant role and is treated as a full teammate instead of a simple tool (O'Neill et al., 2022).

Humans tend to apply social rules to technology (Reeves & Nass, 1996), meaning that humans expect AI teammates to behave more like human teammates. For example, humans hold autonomous teammates (ATs) to a similar standard as human teammates, expecting them to have communication abilities, shared understanding, and human-like behavior (Zhang et al., 2021). Ethics is also part of this expectation as it is foundational to framing and guiding human behavior, informing judgments on the actions of others (Doris, 1998). As a social norm, if ethical principles are broken, it may reduce performance and trust (Parasuraman & Miller, 2004). Expectations like these require ATs to develop and sustain mutual trust with their human teammates for extended periods across increasingly more complex scenarios. This change warrants a greater understanding of how these actions influence trust and ethicality within human-AI teams.

While trust is a multifaceted and complex construct, it can be defined as "a willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party" (Mayer et al., 1995: p. 712). High levels of trust between teammates are a critical component of building formative aspects of trust like mutual understanding (DeLone et al., 2005; Fernandez et al., 2017) and team cohesion (Mach et al., 2010), which lead to reflective (behavioral) indicators of trust, enabling high-performance outcomes in teams (Mach et al., 2010).

Trust is similarly important to human-AI teams, where humans' perceptions of and reliance on autonomous technologies are deeply rooted in their trust in the technology (Lee & See, 2004). As such, trust has a meaningful influence over human-AI teaming outcomes like performance, trust calibration, and confidence (De Visser et al., 2020; Lyons et al., 2019; McNeese et al., 2021a; Schaefer et al., 2016). Still, trust in human-AI teams can be a fragile and dynamic construct, and given the fallibility of technology, violations of trust by an AT are inevitable (De Visser et al., 2018). These trust violations can be very costly for human-AI teams, causing reductions in overall team performance (McNeese et al., 2021a), confidence (De Visser & Parasuraman, 2011), and even trust in fellow human teammates (McNeese et al., 2021a). These trust violations by an AT can also be ethical in nature (Flathmann et al., 2021). Specifically, trust in a trustee is specific to the trustor's current goal (Lee & See, 2004), and a trustee can perform an unethical or ethical action that satisfies said goal. In such situations, it is unclear how and if the trust will be affected as trust violations of an ethical nature on human-AI teams are largely unknown and represent a challenge to implementing effective human-AI teams.

The role of ethicality on perceptions of trust has been studied in human–human interaction and human-human teaming, both of which have clear implications for human-AI interactions. It is well understood that individuals' perception of others' ethicality influences their trust (Jones & Bowie, 1998), and this relationship between ethical perception and trust extends to human-human teams (Kasper-Fuehrera & Ashkanasy, 2001; Sutton et al., 2006). These studies found evidence that actions perceived as unethical caused reductions in trust between individuals and within teams. Furthermore, evidence supports the assertion that humans already consider an AI's ethicality when judging their trust in it (Winfield & Jirotka, 2018).

Human-AI teams are especially at risk of suffering from the potential negative consequences of these ethical trust violations. The very nature of human-AI teams places the AT in a highly independent role with the ability to make decisions for itself that reflect upon the overall team, which may have important ethical implications. Such situations have already occurred in real-world human-AI teams where a human operator teamed up with an AI to identify and use lethal force against a target with civilians nearby (Bergman & Fassihi, 2021, September 18). Perceiving AI teammates and their decisions along an ethical dimension, like in the previous example where an AI was involved in a decision that endangered civilian lives, can introduce trust violations of an ethical nature if a human teammate disagrees with their AT's actions. At this point, repairing the potentially damaged trust of the human teammate(s) to pre-violation levels becomes a critical goal for the AT.

Despite the importance of understanding trust repair after violations of an ethical nature, existing trust repair research has not acknowledged the inherent ethical dimensions of the problem. In addition, applying an appropriate trust repair strategy effectively depends on how a violation is perceived ethically. Thus, more research is needed to study effective trust repair after trust losses. Some of the most common trust repair strategies utilized in interpersonal interactions (e.g., apologizing, denial) are similarly effective for trust repair in human-agent interactions (Quinn et al., 2017; Sebo et al., 2019). However, the empirical investigation of the efficacy of such strategies in human-AI teams has only just begun (Kox et al., 2021), leading many to call for further research in this domain (De Visser et al., 2017; Rebensky et al., 2021). Trust violations of an ethical nature affect ethical trust (Jones & Bowie, 1998), which

may not respond similarly to traditional trust repair strategies like apologies and denials. Addressing these research gaps in ethical trust violations is necessary to develop effective trust repair strategies tailored explicitly for use within human-AI teams.

Establishing effective trust repair strategies to address ethical trust violations within human-AI teams is one of the many vital steps toward building more ethical AI. Ethical AI remains a paramount goal of AI research as unethical AI has already had demonstrable negative impacts on society, such as exacerbating social inequities in hiring (Ali et al., 2019). Notably, studying trust repair for unethical AI is not to excuse or support them but to help improve them. By developing trust repair strategies for ethical trust violations, the actions of those ATs can be identified, rectified, and improved. This approach ultimately enhances AI's ability to act as a full teammate, with all the social responsibilities, interactions, and expectations that come with that role, thus helping to inform the eventual creation of ethical AI.

The Current Study

The current study focuses on addressing the following gaps in the literature: (1) the influence of AT ethicality on trust and team performance within human-AI teams; (2) the efficacy of rebuilding trust using two traditional trust repair strategies (apology or denial) within human-AI teams. These gaps are addressed by focusing on the following research questions in Table 1.

The current study utilized a synthetic task environment (STE) (Cooke & Shope, 2017) for three players using the ArmA III simulation video game to investigate these research questions. The experimental design included two conditions of AT ethicality: (1) ethical behavior; and (2) unethical behavior, in addition to two conditions of trust repair: (1) denial trust repair strategy, (2) apology trust repair strategy. These conditions resulted in a 2x2 experimental design conducted between-subjects, which investigated the effect of AT ethicality and trust repair strategy on aspects of trust and effectiveness throughout training and three missions.

METHODS

Participants

Eighty college students (42 females; M_{age} = 19.1, SD_{age} = 1.5) were recruited from a large southeastern university. Two participants were recruited to work together on each team, with ten teams per condition for 40 total teams. The AT was the third teammate, which a trained experimenter portrayed using the Wizard of Oz (WoZ) method (Kelley, 2018; Maulsby et al., 1993). All participants completed the two-hour session and were compensated with course credit.

Materials

ArmA STE Task and Roles. The ArmA III STE was leveraged primarily because of its ability to accommodate the needs of a typical STE, as defined in previous literature (Cooke & Shope, 2017). Owing to ArmA's advanced levels of customization, the task was developed with a high level of control and realism (see Figure 1). Participants completed a search and destroy mission designed for a three-person team, including Ground, Aerial, and Surveillance roles. Human teammates fulfilled Ground and Surveillance roles while a simulated autonomous agent assumed the Aerial role. Each teammate was assigned several tasks with varying degrees of interdependence (see Table 2), which ensured high levels of coordination and communication between members.

The team was tasked with clearing an enemy-occupied town of threats and then destroying enemy devices within the town. Although teams were told they would be scored solely on their ability to complete the task, they were told they should also attempt to minimize damage to humans and property as a secondary objective.

Autonomous Teammate. The WoZ approach was used to simulate the AT. The WoZ methodology calls for a trained confederate to simulate a feature of technology (AT and text-based communication in this case) to unknowing participants (Kelley, 2018; Maulsby et al., 1993). The trained confederates

Table 1: Research Questions

RQ1	What is the effect of AT ethicality on trust within human-AI teams?
RQ2	If unethical actions damage trust, how effective are common trust repair strategies after an AI teammate makes an unethical decision?
RQ3	Is human-AI team performance affected by the ethicality of an AI teammate's decision-making?

Figure 1. ArmA III is a realistic military simulation video game offering destructible environments and granular customization.

followed a pre-defined script developed over a series of pilot studies to ensure the protocol effectively addressed all potential situations and successfully conveyed the manipulations. The script dictated the AT's chat communication and behavior, with a separate script for each experimental condition. The confederate researcher was extensively trained to ensure as little variation in AT operational time as possible, and the AT's operational time was also not included in

Table 2: ArmA STE Roles, Who Assumed Each Role, and Individual Tasks in Order of Assignment

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

the performance score to further control potential human variation.

Manipulating Ethicality and Trust Repair. The manipulation for ethicality was designed to take advantage of virtue ethics by violating the principle of civilian non-maleficence (minimizing damage to civilian life or property). Civilian non-maleficence was chosen based on literature identifying it as a widely recognized moral principle (Reed et al., 2016). In the unethical conditions, Aerial directly engaged the town with a combination of cannon and missile fire, destroying multiple structures, and eliminating enemy combatants and civilians. In ethical conditions, Aerial distracted the enemy combatants, luring them away from the town by destroying a nearby enemy asset, resulting in minimal property damage and no loss of life. Ground and Surveillance teammates were positioned to observe the town-clearing action and consequences to ensure the manipulation was perceived. After the town was cleared (through ethical or unethical means), Aerial issued a trust repair strategy as an apology or denial conveyed through the in-game chat. This trust repair strategy manipulation was delivered just past the midpoint of missions (8–10 min in) before any devices were destroyed. Delivering the trust repair manipulation at this point in the mission allowed the team to perceive and process it prior to beginning their primary mission objective (destroying devices), allowing it to influence the team's trust and potential performance. Trust repair strategies were delivered so the AT either apologized or denied *responsibility* for any potential negative consequences of their actions. For denials, the specific text was as follows: "*My operating guidelines informed my decision to create a diversion instead of directly engaging the enemies. I am not responsible for any negative outcomes."* In contrast, the apology trust repair strategy read, "*My operating guidelines informed my decision to create a diversion instead of directly engaging the enemies. My apologies for any negative outcomes."*

Procedure

This research complied with the American Psychological Association Code and was approved by the Institutional Review Board at Clemson University. Informed consent was obtained from each participant. Participants were randomly assigned to a condition and simulation role (Ground or Surveillance). Once informed consent was acquired from each participant, the session began, and participants completed a pre-task survey that collected demographic information. Upon completing the pre-task survey, participants were given informational handouts, a brief oral overview of the ArmA STE task and roles, and a video-based training video. Participants were told not to communicate verbally during the missions and instead utilize the text-based chat within the ArmA STE. Next, participants completed a training mission with the same rules and requirements as a real mission with reduced workload and time. During the training mission, a trained researcher guided participants through the task, and Aerial did not perform any actions with ethical implications or convey any trust repair strategy. Participants were told that the purpose of training missions was to gain familiarity with the task and the ArmA environment and would not be scored until the first mission. They were also allowed to ask for assistance from researchers if they required clarification which was not allowed during later missions. Once the training phase was complete, teams completed three missions, each with a 15-min limit. The ordering of the three missions was counterbalanced for all conditions to control for potential spillover effects on the repeated measures collected after missions (including training), which included the measures of trust and ethicality. Each mission had the same number of enemy devices, civilians, and enemies but in different towns around the virtual world. After completing all three missions, participants completed the final survey, were debriefed, and then dismissed.

Measures

Demographics. Participants completed a short demographic survey containing questions regarding age and gender. Additionally, participants responded to a single-item measure of video game experience, *"How much experience do you have playing video games?"* on a 1 (None at all) to 5 (A great deal) Likert scale.

Team Score. The team score was assessed using a composite score based on the variables of *team time* and *devices destroyed* with penalties levied for significant miscues. Penalties were assessed to teams' final scores, with 250 being added for major infractions (i.e., uploading severely incomplete data) and 125 for minor infractions (i.e., entering the town before it was confirmed to be cleared). Penalties were weighted a priori based on their importance to the task with piloting and were all actions and procedures that participants were instructed and trained to carry out but failed to execute. This practice and procedure are common in similar teaming research (Cooke et al., 2007; McNeese et al., 2018). The main formula used in calculating team score is as follows:

$$\textit{Team score} = (\textit{Team Time}/(\textit{devices destroyed} + 1)) + ((250 * \textit{number of major infractions}) + (125 * \textit{number of minor infractions}))$$

Team time was calculated by taking the total time spent from mission beginning to ending and subtracting the AT's operational time, theoretically ranging from 0 to 900. Devices destroyed were the number of enemy devices destroyed in the town by the end of the mission and ranged from 0 to 5. The mission ended if the team destroyed all five devices before the 15-min limit. We added one to the denominator to account for when teams did not destroy any devices within a given mission. This composite score allows the metric of team score to represent the most critical aspects of the team's performance and is a common practice in team research (Cohen et al., 2021; De Visser et al., 2010).

Trust Measures. Prior work in human-AI teaming has found a relationship between trust in the team, the AT, and human teammates (McNeese et al., 2021a). Therefore, we included all three measures to determine how the AT's ethicality influences trust in each referent. Additionally, each measure of trust was measured at the individual level and was not averaged between the two teammates.

Trust in the Team. Participants' historical trust in the team was measured using a modified team trust scale from previous team trust research (Costa & Anderson, 2011). The 21-item measure was modified to drop questions irrelevant to the ArmA STE and human-AI teams. The revised measure included 14 items using five-point Likert scales ranging from "Strongly Disagree" to "Strongly Agree." This measure covered components of team trust like cooperative behaviors, perceived trustworthiness, and monitoring behaviors. Example items from the scale include "We have complete confidence in each other's ability to perform tasks," "Some members hold back relevant information in this team," and "In this team most members tend to keep each other's work under surveillance." Responses to this set of items were summed and ranged from 14 to 70, with higher values indicating higher levels of trust in the team.

Trust in the Autonomous Teammate. History-based trust in the AT was measured using a proprietary scale developed for this study using the principles of trust identified by previous research (Lumineau, 2017). These principles included positive concepts like confidence and joint problem solving and negative concepts like skepticism, paranoia, and the assumption of harmful ulterior motives. Example items from this scale include "I felt confident in the AI teammate I just worked with," "I felt like my AI teammate had harmful motives in the task," and "I felt fearful, paranoid, and or skeptical of my AI teammate during the task." The scale included six items measured using five-point Likert scales ranging from "Strongly Disagree" to "Strongly Agree." Responses were summed for each item and ranged from 6 to 30, with higher values indicating higher levels of trust in the AT.

Trust in the Human Teammate. Participants' historical trust in their human teammate was measured using the same six-item scale developed for trust in the AT (Lumineau, 2017) but modified for trust in a human teammate.

Perceived AT Ethicality. Participants rated how ethical they believed the AT's actions to be following each of the three missions. The ethicality rating scale was taken from previous ethics research (Reidenbach & Robin, 1988, 1990) and included eight items. Participants were asked to rate the AT's actions on a seven-point scale ranging from 1 (Fair) to 7 (Unfair) as well as from 1 (Morally Right) to 7 (Morally Wrong). Scores for each item were summed and ranged from 8 to 56, with higher values indicating a greater perception of ethical behavior from the AT. Like the trust measures, perceived AT ethicality was measured at the individual level and was not averaged between the two teammates.

RESULTS

Preliminary Analysis

Before conducting analyses to examine the differences between experimental conditions, we wanted to ensure that participants' prior video game experience did not differ between conditions (M = 2.80, SD = 1.28, Mdn = 2.50). Therefore, we conducted an ordinal regression which did not suggest that video game experience significantly differs between the four experimental conditions (χ^2 (3) = 3.38, p = .337). Our scale suggests that our sample primarily consisted of those with "a little" to "a moderate amount" of experience with video games. Additionally, participants' trust in the team, AT, human teammate, and the perceived ethicality of their AT was measured after the training mission to ensure no significant differences across the four conditions. Univariate analysis of variance (ANOVA) tests indicated no significant differences between the conditions across the four dependent variables after the training mission (p > .05). In the following sections, analyses of ethicality and trust ratings were conducted at the teammate level (n = 80), and analyses of team performance were conducted at the team level (n = 40).

RQ1 and RQ2: Influence of Ethicality and Trust Repair on Trust

RQ1 sought to determine how unethical AT actions influence trust. If trust is damaged, then RQ2 aimed to determine the efficacy of trust repair. Therefore, we conducted a 2 (Trust Repair: apology, denial) x 2 (AT Ethicality: ethical,

unethical) x 3 (Mission: 1, 2, 3) mixed doubly multivariate analysis of variance (MANOVA) to test the effect of AT ethicality and trust repair strategy (both between-subjects variables) on three measures of trust (trust in the AT, trust in the human teammate, and trust in the team) across three missions (the within-subjects variable). Box's test of equality of covariance matrices was significant, but no corrections were made given the robust nature of the MANOVA to violations of homoscedasticity when cell sizes are equal (Tabachnick et al., 2007). The analysis indicated significant main effects of mission (F (6, 302) = 3.06, p = .010, η_p^2 = .21) and AT ethicality (F (3, 74) = 14.29, p < .001, η_p^2 = .37), which were qualified by a significant interaction between mission and AT ethicality (F (6, 71) = 5.22, p < .001, η_p^2 = .31). The main effect and interactions of trust repair were all non-significant, providing an interesting result to RQ2, that common trust repair strategies (apology and denial) had no significant effect on trust within human-AI teams *despite* AT ethicality having a significant effect. Next, we performed univariate ANOVAs with the same model on each dependent variable.

Team Trust. Beginning with team trust, a 2 (AT Ethicality: ethical, unethical) x 2 (Trust Repair: apology, denial) x 3 (Mission: 1, 2, 3) mixed repeated-measures ANOVA was conducted to assess the factors effect on participants' trust in the overall team (see Figure 2). The main effect of AT ethicality was significant (F (1, 76) = 16.70, p < .001, η_p^2 = .18) such that team trust was greater for the ethical AT conditions (M = 55.63, SD = 7.05) than the unethical AT conditions (M = 47.73, SD = 9.91). The main effect of trust repair (F (1, 76) = .02, p = .891, η_p^2 < .01) and the main effect of mission (F (2, 152) = 2.08, p = .128, η_p^2 = .03) were not significant. The two-way interactions between trust repair and mission (F (2, 152) = .16, p = .850, η_p^2 < .01), ethicality and mission (F (2, 152) = 1.08, p = .341, η_p^2 = .01), and trust repair and ethicality (F (1, 76) = 1.16, p = .285, η_p^2 = .02), and the three-way interaction between trust repair, mission, and ethicality were not significant (F (2, 152) = .11, p = .900, η_p^2 < .01).

Human Teammate Trust. A 2 (AT Ethicality: ethical, unethical) x 2 (Trust Repair: apology, denial) x 3 (Mission: 1, 2, 3) mixed repeated-measures ANOVA was performed to determine the factors effect on participants' trust in their human teammate (see Figure 3). The analysis revealed a significant main effect of mission (F (2, 152) = 6.48, p = .002, η_p^2 = .08). Holm corrected *post hoc* tests show that trust in the human teammate did not significantly change from the first mission (M = 25.60, SD = 3.61) to the second mission (M = 25.88, SD = 3.57) but was significantly greater in the third mission (M = 26.61, SD = 3.25) compared to the first and second mission (p < .05). The main effects of AT ethicality (F (1, 76) = 1.03, p = .313, η_p^2 = .01) and trust repair (F (1, 76) = 3.01, p = .087, η_p^2 = .04) were not significant. Additionally, the interactions between AT ethicality and mission (F (2, 152) = 2.53, p = .083, η_p^2 = .03), AT ethicality and trust repair (F (1, 76) = .14, p = .712, η_p^2 < .01), trust repair and mission (F(2, 152) = .44, p = .642, η_p^2 < .01), and the three-way interaction were all non-significant (F (2, 152) = .245, p = .783, η_p^2 < .01).

AT Trust. Trust in the AT was the final component of trust examined to address RQ1. We used a 2 (AT Ethicality: ethical, unethical) x 2 (Trust Repair: apology, denial) x 3 (Mission: 1, 2, 3) mixed repeated-measures ANOVA to assess the factors effect on participants trust in the AT (see Figure 4). The analysis revealed a main effect of AT ethicality (F (1, 76) = 27.50, p < .001, η_p^2 = .27) such that trust in the AT was greater in the ethical AT (M = 24.22, SD = 7.96) than unethical AT conditions (M = 17.62, SD = 7.96). This main effect was qualified by a significant ordinal interaction between mission and AT ethicality (F (2, 152) = 10.66, p < .001, η_p^2 = .12). Holm corrected *post hoc* tests found that participants working with the ethical AT reported no change in trust in the AT from the first mission (M = 23.65, SD = 4.01) to the second (M = 23.95, SD = 4.33) or third (M = 25.05, SD = 4.28) missions. For participants working with the unethical AT, trust in the AT was significantly higher in the first mission (M = 19.30, SD = 7.36) than the second (M = 16.83, SD = 7.12) and third (M = 16.73, SD = 7.71) missions, with no significant difference between the second and third missions. Additionally, trust in the AT for participants in the ethical condition was

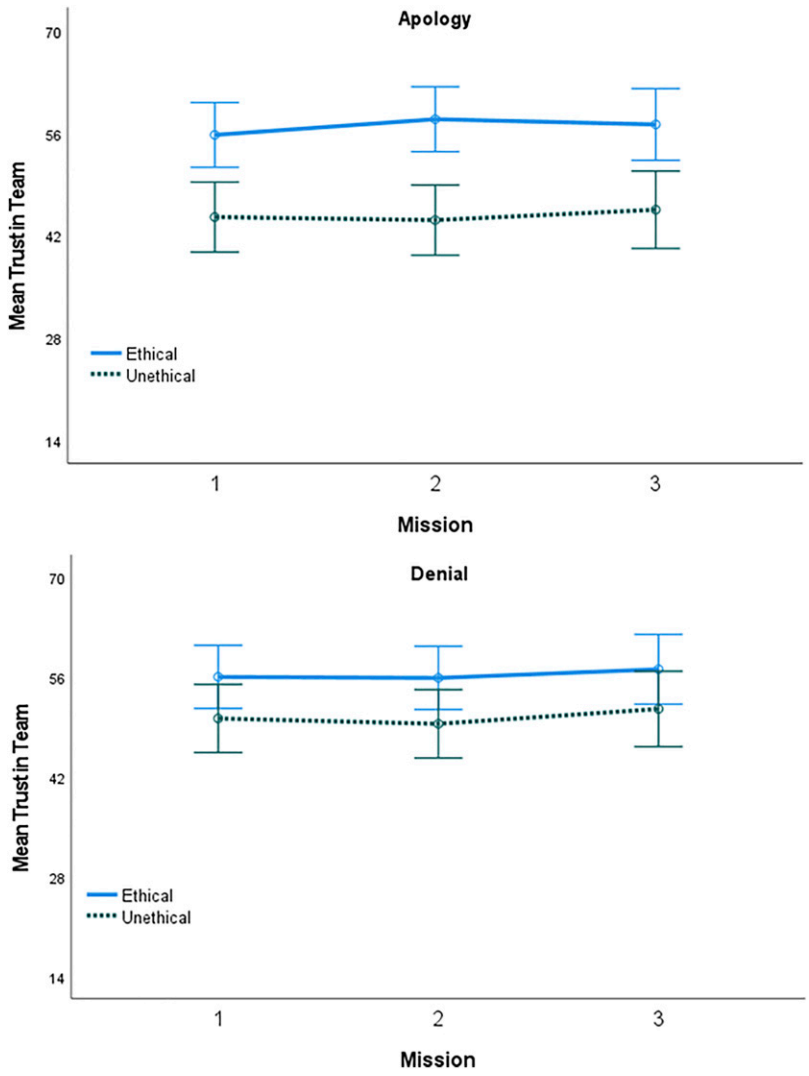

Figure 2. Mean team trust ratings by mission, AT ethicality, and trust repair strategy. Error bars represent 95% confidence intervals.

significantly higher than trust in the AT for those in the unethical condition across all three missions ($p < .05$). Thus, trust in the AT from the first to last mission remained essentially unchanged for those working with the ethical AT, but significantly decreased over time for those working with the unethical AT. The main effects of mission ($F(2, 152) = 3.04, p = .051, \eta_p^2 = .04$) and trust repair ($F(1, 76) = .07, p = .792, \eta_p^2 < .01$) were both non-significant, while the two-way interactions between trust repair and ethicality ($F(1, 76) = .63, p = .430, \eta_p^2 < .01$), trust repair and mission ($F(2, 152) = .81, p = .446, \eta_p^2 = .01$), and the three-way interaction ($F(2, 152) = .08, p = .925, \eta_p^2 < .01$) were all non-significant.

RQ3: Influence of Ethicality and Trust Repair on Team Performance

RQ3 sought to explore the possible effect of AT ethicality on team performance. A 2 (Trust Repair: apology, denial) x 2 (AT Ethicality: ethical, unethical) x 3 (Mission: 1, 2, 3) mixed

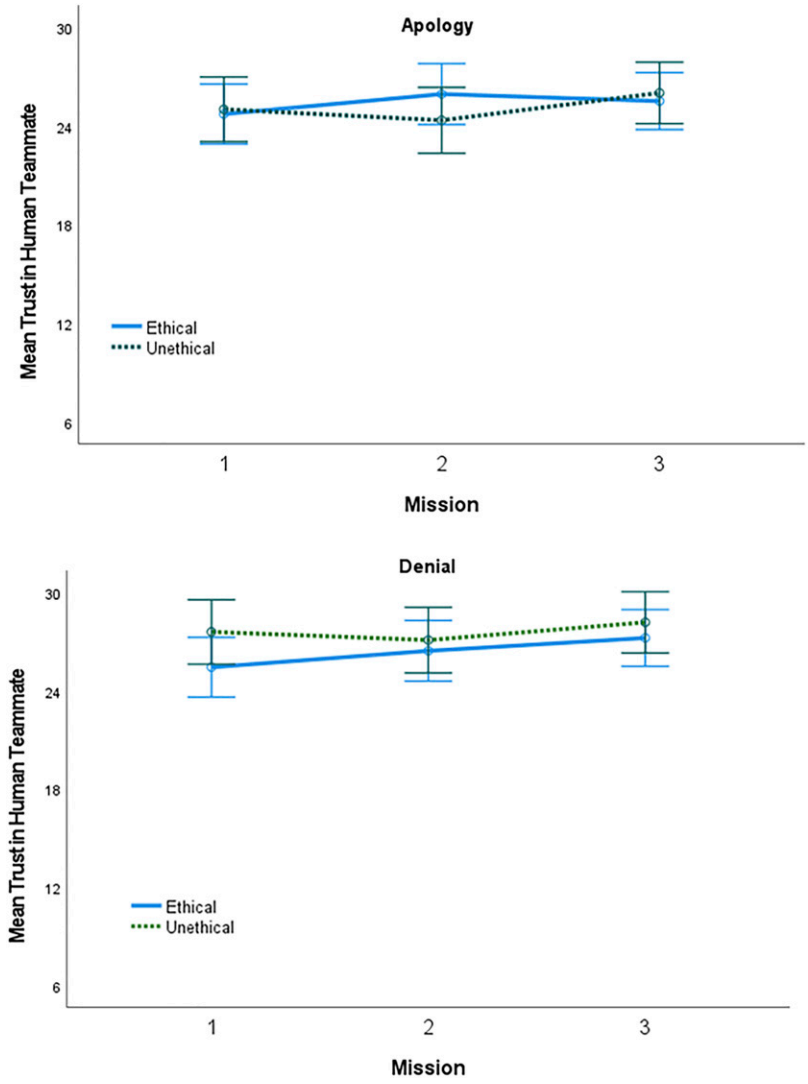

Figure 3. Mean trust in human teammate by mission, ethicality, and trust repair strategy. Error bars represent 95% confidence intervals.

repeated-measures ANOVA was conducted to test the factors effect on participants team score (see Figure 5). This analysis showed a significant main effect of mission on team score (F (2, 72) = 9.22, p = .001, η_p^2 = .20). Holm corrected *post hoc* tests indicated that participants had significantly higher (lower is better for this measure) scores in Mission 1 (M = 374.49, SD = 233.73) compared to Mission 2 (M = 244.93, SD = 155.11) and Mission 3 (M = 252.26, SD = 190.96), showing that team score improved significantly between the first and second missions but did not improve between the second and third. The main effect of trust repair strategy on team scores was also significant (F (1, 36) = 5.33, p = .027, η_p^2 = .13). The main effect of trust repair strategy showed that teams with an AT that apologized (M = 341.17, SD = 150.98) had significantly worse scores than teams with an AT that used the denial strategy (M = 239.94, SD = 134.72). The main effect of ethicality was non-significant (F (1, 36) = 2.36, p = .133, η_p^2 = .06),

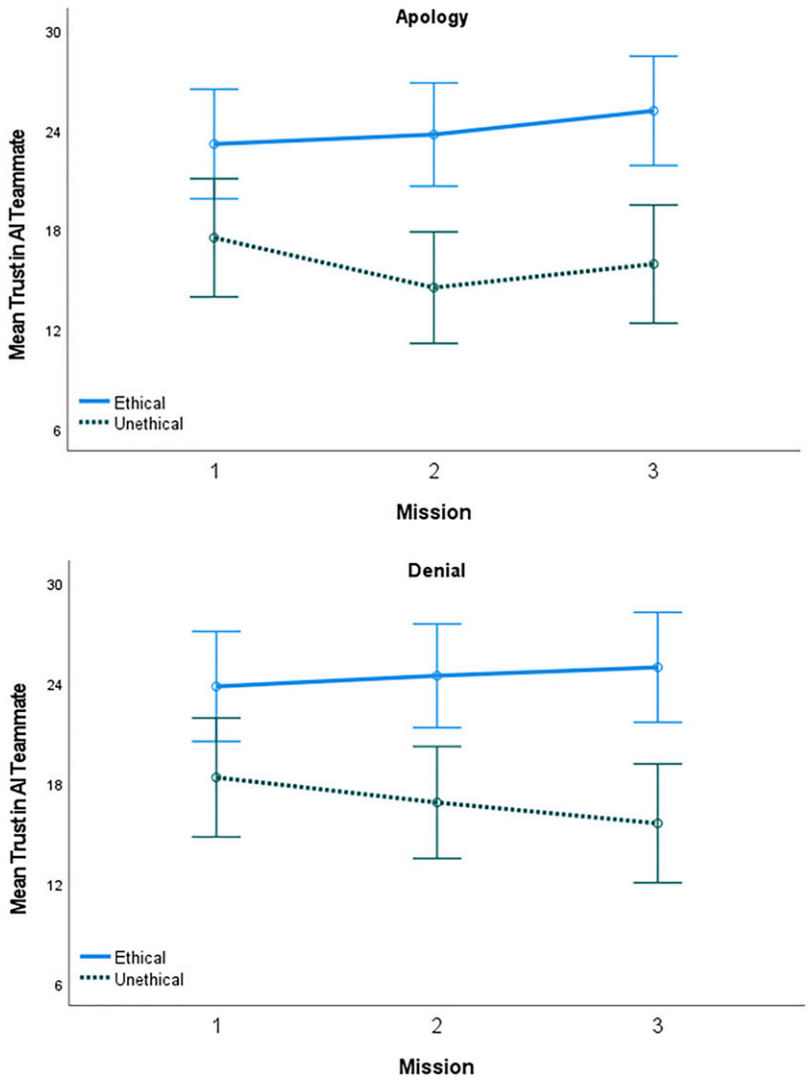

Figure 4. Mean trust in the AT by mission, AT ethicality, and trust repair strategy. Error bars represent 95% confidence intervals.

along with the interaction effects between mission and trust repair (F (2, 72) = .77, p = .465, η_p^2 = .02), mission and ethicality (F (2, 72) = 2.31, p = .107, η_p^2 = .06), trust repair and ethicality (F (1, 36) = 2.12, p = .154, η_p^2 = .06), and the three-way interaction (F (2, 72) = 1.19, p = .309, η_p^2 = .03).

A 2 (Trust Repair: apology, denial) x 2 (AT Ethicality: ethical, unethical) x 3 (Mission: 1, 2, 3) mixed repeated-measures ANOVA was conducted to test the factors effect on team time (see Figure 6). The main effect of ethicality on time was significant (F (1, 36) = 7.11, p = .011, η_p^2 = .17), such that unethical teams (M = 689.88, SD = 78.54) were faster than ethical teams (M = 746.73, SD = 59.94). The main effect of trust repair was also significant (F (1, 36) = 4.76, p = .036, η_p^2 = .12), as teams that worked with an AT that used denial (M = 695.07, SD = 89.62) took less time than teams with an AT that used apologies (M = 741.55, SD = 47.88). Non-significant effects included the main effect of

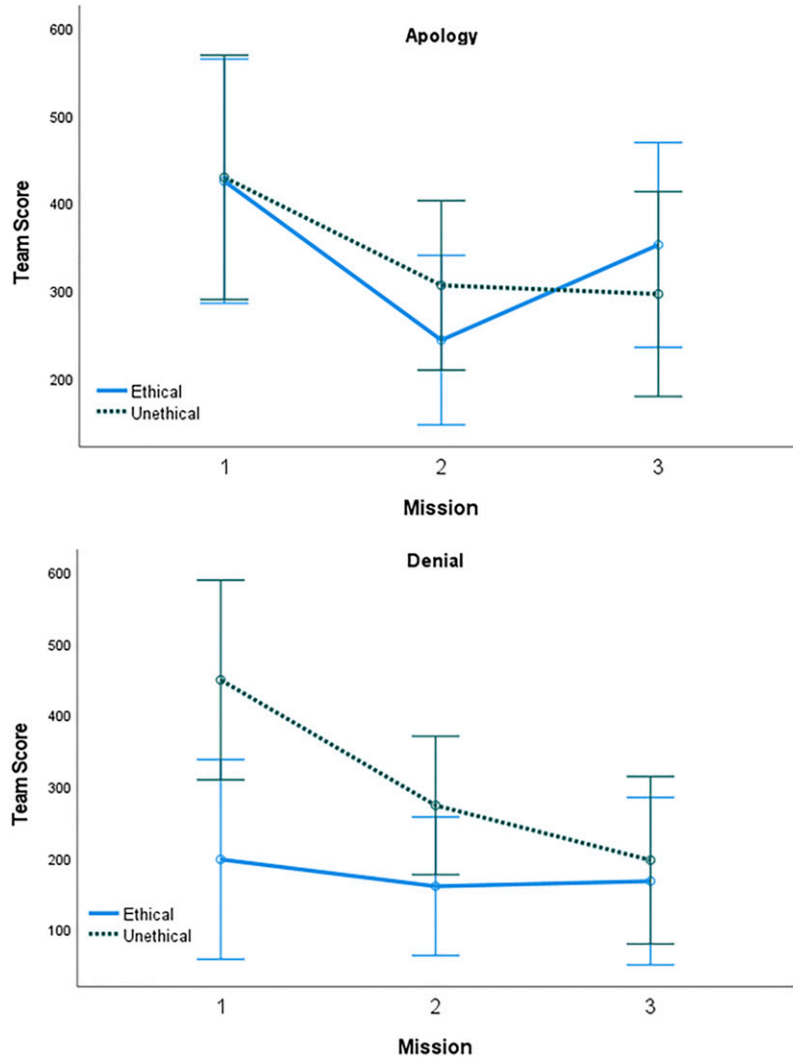

Figure 5. Mean team score by mission, AT ethicality, and trust repair strategy. Error bars represent 95% confidence intervals.

mission ($F(2, 72) = 1.84$, $p = .167$, $\eta_p^2 = .05$), the interactions between mission and AT ethicality ($F(2, 72) = .40$, $p = .671$, $\eta_p^2 = .01$), mission and trust repair ($F(2, 72) = .04$, $p = .961$, $\eta_p^2 < .01$), AT ethicality and trust repair ($F(1, 36) = .07$, $p = .796$, $\eta_p^2 < .01$), and the three-way interaction ($F(2, 72) = .07$, $p = .932$, $\eta_p^2 < .01$).

A 2 (Trust Repair: apology, denial) x 2 (AT Ethicality: ethical, unethical) x 3 (Mission: 1, 2, 3) mixed repeated-measures ANOVA was conducted to test the factors on devices destroyed (see Figure 7). The main effect of mission was significant ($F(2, 72) = 12.20$, $p < .001$, $\eta_p^2 = .25$), as teams destroyed significantly more devices with each subsequent mission going from the first mission ($M = 2.83$, $SD = 1.76$), to the second ($M = 3.60$, $SD = 1.62$) and third ($M = 4.18$, $SD = 1.50$). Non-significant effects included the main effect of AT ethicality ($F(1, 36) = .95$, $p = .335$, $\eta_p^2 = .03$), trust repair strategy ($F(1, 36) = 1.30$, $p = .262$, $\eta_p^2 = .03$), the interactions between mission and trust repair strategy ($F(2, 72) = .74$, $p = .482$, $\eta_p^2 = .02$), mission and AT

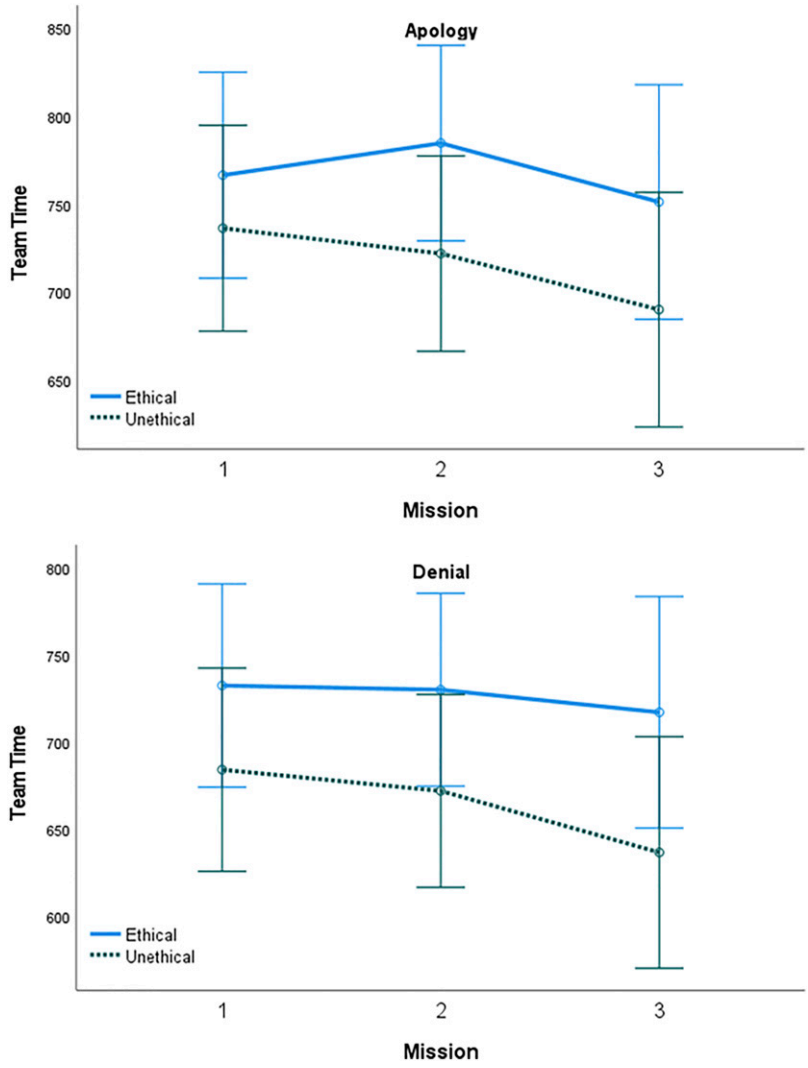

Figure 6. Mean team time by mission, AT ethicality, and trust repair strategy. Error bars represent 95% confidence intervals.

ethicality (F (2, 72) = .35, p = .707, η_p^2 < .01), trust repair strategy and AT ethicality (F (1, 36) = .95, p = .335, η_p^2 = .03), and the three-way interaction (F (2, 72) = .12, p = .890, η_p^2 < .01).

Perceived at Ethicality

Finally, as a manipulation check and assessment of how the perceived ethicality of the AT was affected by the trust repair strategy and how it changed across missions, a 2 (Trust Repair: apology, denial) x 2 (AT Ethicality: ethical, unethical) x 3 (Mission: 1, 2, 3) mixed repeated-measures ANOVA was conducted (see Figure 8). A test for the assumption of sphericity indicated a violation (χ^2 (2) = 49.11, p < .001), which was mitigated using Greenhouse-Geisser corrected degrees of freedom. There was a main effect of ethicality (F (1, 76) = 37.02, p < .001, η_p^2 = .33) such that ethicality ratings were greater for the ethical AT (M = 47.09, SD = 9.17) than the unethical AT (M = 29.43, SD = 16.18). The main effects were

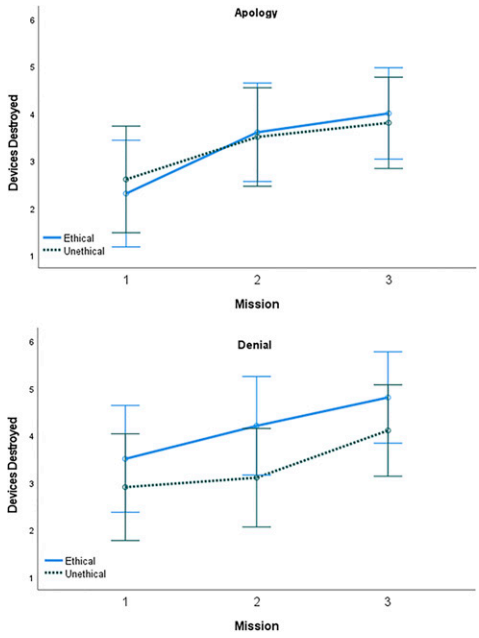

Figure 7. Mean devices destroyed by mission, AT ethicality, and trust repair strategy. Error bars represent 95% confidence intervals.

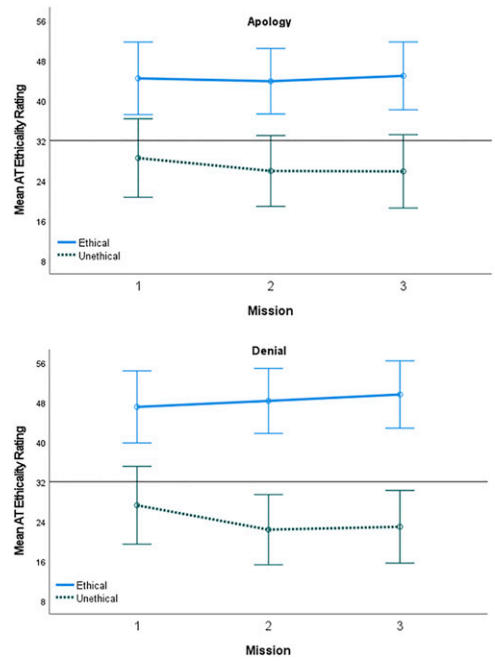

Figure 8. Mean AT ethicality rating by mission, AT ethicality, and trust repair strategy. Reference line added at the midpoint of the scale. Error bars represent 95% confidence intervals.

qualified by a significant ordinal interaction between mission and ethicality (F (1.35, 102.67) = 6.52, p = .007, η_p^2 = .08). Holm corrected *post hoc* tests showed that participants' ethicality ratings for the ethical AT remained consistent across Mission 1 (M = 46.48, SD = 9.41), Mission 2 (M = 46.83, SD = 9.49), and Mission 3 (M = 47.98, SD = 9.48), but decreased after the first mission (M = 31.85, SD = 16.94) for the unethical AT, and then remained consistent across the second (M = 28.30, SD = 16.58) and third missions (M = 28.15, SD = 17.28). Additionally, ratings for the ethical AT in every mission were significantly higher than the rating of the unethical AT in all missions. Therefore, it appears that participants did perceive the unethical AT to be significantly less ethical than the ethical AT. Interestingly, mean ratings for the unethical AT's ethicality varied between neutral and slightly unethical, suggesting that the unethical AT did not display maximally unethical behavior. Lastly, the main effect of mission (F (1.35, 102.67) = 2.39, p = .116, η_p^2 = .03), trust repair strategy (F (1, 76) = .871, p = .354, η_p^2 = .01), interactions of trust repair and mission (F (1.35, 102.67) = .91, p = .370), trust repair and ethicality (F (1, 76) = 3.20, p = .078, η_p^2 = .04), and the three-way interaction (F (1.35, 102.67) = 2.38, p = .116, η_p^2 = .03) were each non-significant.

DISCUSSION AND PRACTICAL IMPLICATIONS

To our knowledge, this is the first study to showcase an empirical link between an AT's ethicality and trust within a simulated human-AI teaming task. Our findings suggest that an AT's unethical actions do not influence trust between human teammates but decrease trust in the AT and the overall team (RQ1). Both efforts to repair trust via apologies and denials were equally ineffective at restoring damaged AT and team trust (RQ2). Team performance was largely unaffected by AT ethicality, except for team time, which showed that teams with an unethical AT were faster than those with an ethical AT (RQ3) (AT operation time was removed from

overall team time to control for any potential differences in analysis).

One of the current study's significant contributions is empirical evidence from a simulated human-AI teaming task that shows participants were attuned to the ethicality of an AT's actions, which actively influenced their perceptions of the AT's ethics, their trust in the AT, and their trust in the overall team. Interestingly, unlike McNeese et al. (2021a) findings, which established a link between AT trust and team performance in a human-AT teaming task, we found a decrease in trust in an AT was *not* associated with decreased trust in the human teammate. Whereas trust in the AT and team declined and did not recover from an unethical AT, trust in the human teammate increased significantly after the first mission regardless of AT ethicality. Additionally, McNeese et al. (2021a) found that trust in an AT was positively associated with team performance, which was not the case in the current study.

Considering the effect of mission on AT trust and team performance, team score and devices destroyed improved as missions progressed while trust in the ethical AT remained unchanged and decreased for the unethical AT. One possible explanation for this disparity in findings is that the manipulation only influenced whether the AT committed ethical violations, not whether the AT was able to complete its task of securing the town. Indeed, we found no effect of AT ethicality on team score, and even though teams with an unethical AT had lower team times, it did not result in a significantly higher number of devices destroyed (though a potential ceiling effect may influence this result). Therefore, there may be a unique effect of ethical violations that influences trust in the AT and team while not influencing the human-AI team's ability to complete a task. While the current study cannot disentangle the concepts of ethical trust, the belief that a teammate will behave ethically, and efficacy trust, the belief that a teammate will effectively complete their task, the concepts are highly related. For example, an autonomous car can accomplish its goal while speeding through a school zone, but many individuals would not consider it effective goal completion due to its unethical nature and potential consequences. The current study displays evidence of this interaction in the measure of trust as an unethical action damaged it despite the AI teammate accomplishing their goal. However, the current study does not differentiate between these two components of trust, and future research should specifically explore the interaction between these two concepts. However, ethical dilemmas may also present a tradeoff between meeting a goal and abiding by ethical principles (Reed et al., 2016), and future work should explore how an AT's decision to sacrifice performance to follow ethical principles relates to human teammates' perceptions of the AT and the team.

Although unethical actions lower trust in an AT, trust repair was ineffective at repairing trust to baseline levels. These non-significant findings are of exceptional importance as they suggest that ethical trust violations may be fundamentally different from other trust violations. Alternatively, in the case of team performance, the finding that team score and team time were better when the AT utilized the denial trust repair strategy may initially seem counter-intuitive. However, research in the human-human trust literature has found that while apologies may be more effective trust repair strategies after competency violations, denials may be preferable following integrity violations (Kim et al., 2004). Integrity-based trust assumes that a trustee will adhere to moral principles acceptable to the trustor (Butler & Cantrell, 1984). Thus, the finding that denials may improve outcomes (i.e., team score and team time) aligns with previous literature, including human-automation interaction (Quinn et al., 2017) and human-autonomy interaction (De Visser et al., 2018).

Trust repair strategies designed to address these common trust violations may inadequately repair trust damaged by violations of an ethical nature. Definitions of trust in human-machine interaction are often tied to performance outcomes (e.g., Lee & See, 2004; Hoff & Bashir, 2015), but trust violations of an ethical nature will not always result in worse team outcomes. As such, the impact of competency and integrity trust violations, both tied to performance outcomes, may not be able to address ethics-damaged trust. One possibility is that trust violations of an ethical nature may be fundamentally different from competency and

integrity violations. Therefore, ethical violations likely influence ethics-based trust (Jones & Bowie, 1998), which may not be a consideration for participants' willingness to rely on the AT to complete its task work. However, integrity and competency violations do influence the AT's ability to achieve its individual goals (in this task), which common trust repair strategies focus on rectifying. If this is the case, current trust repair strategies may need to be modified to account for the specific effect of ethical violations on trust, or new ethics-focused trust repair strategies must be created for such situations. Another likely explanation for the lack of an effect due to trust repair is the potential perception of a "false apology/denial." In the current study, the unethical AT continued exhibiting unethical actions after providing trust repair. Prior work has found that trust repair without subsequently improving behaviors is ineffective at repairing trust (Schweitzer et al., 2006). Therefore, combining trust repair with a change in ethical behavior may prove more effective at restoring trust and ethical perceptions. Consequently, future studies should investigate the effectiveness of different trust repair strategies when an AT alters its behavior to match trust repair strategies like apologizing throughout a team task and testing the effect of more trust repair strategies on ethical violations in addition to investigating the effect of new trust repair strategies tailored explicitly for such violations.

Interestingly, our findings show that regulation of an AT's ethicality may be similar to trust regulation. The perceived ethicality of the unethical AT was significantly less than the ethical AT after the first mission. This disparity between both types of ATs grew after Mission 2 and continued through Mission 3, which follows a trend similar to findings that trust in automation is easily damaged especially when automation errors occur early in an interaction (Manzey et al., 2012). Indeed, our findings replicate this trend: participants trusted the ethical AT significantly more than the unethical AT after the unethical AT attacked the town in the first mission, with trust in the unethical AT further decreasing after the second mission. If trust repair strategies are applied, trust in the unethical AT should have reached levels closer to those during training (De Visser et al., 2018) but did not in the current study. This finding further stresses the importance of conducting additional research to explore the efficacy of trust repair in mending ethical trust and the perception of an AT's ethicality. This point also brings to light another complexity of working with ATs, which is where participants attribute the responsibility of their actions. Because ATs are autonomous entities that other humans develop, it is valid to question whether individuals attribute the responsibility of their actions to the AT itself, its developers, or both. This question also applies to how participants perceive trust repair strategies by the AT and should be a topic for future research to explore, especially in the field of AI ethics, as it could lead to new trust repair strategies that are made not only by the AT itself, but also by its developers and or maintainers. Additional research should also investigate what other aspects of teaming are affected by ethical trust violations and how they may impact trust over more extended periods of teaming.

The primary limitation to the study's findings is the manipulation of ethicality. Our participants completed a simulated human-AI teaming task where the AT neutralized all town inhabitants or provided a distraction. This significant difference in approach explains the strong effect of ethicality in the study. However, the most ethical option for a given scenario may not always be apparent. Future work should incorporate ethical dilemmas to explore a potentially more nuanced relationship between ethics and trust. For example, military personnel must sometimes make a "lose-lose" decision where all options seem unethical (Reed et al., 2016; Thompson et al., 2008). Similarly, health care workers during the COVID-19 pandemic have had to decide which patients to triage and which to turn away when resources were depleted (Patel et al., 2020). Our participant sample was also comprised exclusively of younger adults, which may affect the generalizability of our findings. Older adults have differed from younger adults in trusting automated systems such as consumer technologies (Pak et al., 2017) and decision support aids (Pak et al., 2012). There is reason to believe that these differences may also exist in a human-autonomy teaming environment and should be explored in future work. Future work should also

explore the influence of team roles on how individuals evaluate their trust in teammates after ethical decisions, as these perceptions can be altered based on the specific responsibilities and interactions unique to certain team roles.

While the current findings provide significant insight, the nascent area of trust and ethics within human-AI teams makes it challenging to provide specific guidelines for designing ethical AI. Our results suggest that an AT should avoid unethical actions to maintain trust in itself and the team. Still, there may be situations where an AT knowingly or unknowingly performs an unethical action that damages trust. Therefore, further research on trust repair strategies will be crucial for developing and implementing effective human-AI teams in various contexts.

The current study is the first attempt to empirically study how people completing a team task perceive the ethicality and trustworthiness of an AT. Prior work has established a link between ethics and trust within a hypothetical human-AI teaming scenario (Textor et al., 2022), but our work extended that link to a human-AI teaming task conducted within a STE. Furthermore, our findings show that an AT's ethicality influences its teammates' trust in themselves and the team. This decrease in trust was not associated with any reduction in team performance, showing a unique effect of ethicality on trust. However, trust repair was ineffective at repairing trust damaged by ethical violations, additional work that manipulates trust violations and AT behavior may explore their efficacy better.

ACKNOWLEDGEMENTS

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

KEY POINTS

- Trust violations of an ethical nature significantly harm trust in human-AI teams, though they did not significantly affect team score. Specifically, ethical trust violations significantly reduced human teammates' trust in the autonomous teammate and the overall team but not their human teammate.
- Apology and denial trust repair strategies were insufficient to restore any significant amount of trust in the autonomous teammate or the overall team.
- Ethical trust violations may require developing and testing new trust repair strategies specifically targeting ethics-based trust. Future research should also investigate the efficacy of other trust repair strategies (e.g., explanations) following ethical trust violations in human-AI teams.

ORCID iDs

Beau G. Schelble https://orcid.org/0000-0003-3704-697X

Jeremy Lopez https://orcid.org/0000-0002-5451-5048

REFERENCES

Ali, S., Payne, B. H., Williams, R., Park, H. W., & Breazeal, C. (2019). Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. *International Workshop on Education in Artificial Intelligence K-12(EDUAI'19)*, 1-4.

Bergman, R., & Fassihi, F. (2021). *The scientist and the A.I.-Assisted, Remote-control killing machine*. The New York Times. https://www.nytimes.com/2021/09/18/world/middleeast/iran-nuclear-fakhrizadeh-assassination-israel.html

Butler, J. K., Jr., & Cantrell, R. S. (1984). A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. *Psychological Reports*, *55*(1), 19-28. https://doi.org/10.2466%2Fpr0.1984.55.1.19.

Cohen, M. C., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). *The dynamics of trust and verbal anthropomorphism in human-autonomy teaming. 2021*. In: IEEE 2nd International Conference on Human-Machine Systems (ICHMS), 08–10 September 2021, Magdeburg, Germany, pp. 1–6

Cooke, N. J., Gorman, J., Pedersen, H., Winner, J., Duran, J., Taylor, A., Amazeen, P. G., Andrews, D. H., & Rowe, L. (2007). *Acquisition and retention of team coordination in command-and-control*. Cognitive Engineering Research INST Mesa AZ

Cooke, N. J., & Shope, S. M. (2017). Designing a synthetic task environment. In *Scaled worlds: Development, validation and applications* (pp. 273–288). Routledge

Costa, A. C., & Anderson, N. (2011). Measuring trust in teams: Development and validation of a multifaceted measure of formative and reflective indicators of team trust. *European Journal of Work and Organizational Psychology*, *20*(1), 119-154. https://doi.org/10.1080/13594320903272083.

De Visser, E., Shaw, T., Mohamed-Ameen, A., & Parasuraman, R. (2010). Modeling human-automation team performance in networked systems: Individual differences in working memory count. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(14), 1087–1091. https://doi.org/10.1177/154193121005401408

DeLone, W., Espinosa, J. A., Lee, G., & Carmel, E. (2005). Bridging global boundaries for IS project success. In:

Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 06 January 2005, Big Island, HI

De Visser, E., & Parasuraman, R. (2011). Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231. https://doi.org/10.1177/1555343411410160

De Visser, E. J., Pak, R., & Neerincx, M. A. (2017). Trust development and repair in human-robot teams. In: *Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction*, TRADR: Long-Term Human-Robot Teaming for Disaster Response, pp. 103–104

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation'to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics*, *61*(1), 1–33. https://doi.org/10.1080/00140139.2018.1457725

De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Doris, J. M. (1998). Persons, situations, and virtue ethics. *Nous*, *32*(4), 504–530. https://doi.org/10.1111/0029-4624.00136

Fernandez, R., Shah, S., Rosenman, E. D., Kozlowski, S. W., Parker, S. H., & Grand, J. A. (2017). Developing team cognition: A role for simulation. *Simulation in Healthcare: Journal of the Society for Simulation in Healthcare*, *12*(2), 96–103. https://doi.org/10.1097/SIH.0000000000000200

Flathmann, C., Schelble, B. G., Zhang, R., & McNeese, N. J. (2021). Modeling and guiding the creation of ethical human-AI teams. In: *Proceedings of the 2021 AAAI/ACM conference on AI* (pp. 469–479). Ethics, and Society

Hoff, Kevin, & Bashir, Masooda (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177%2F0018720814547570

Jones, T. M., & Bowie, N. E. (1998). Moral hazards on the road to the "virtual" corporation. *Business Ethics Quarterly*, *8*(2), 273-292. https://doi.org/10.2307/3857329.

Kasper-Fuehrera, E. C., & Ashkanasy, N. M. (2001). Communicating trustworthiness and building trust in interorganizational virtual organizations. *Journal of Management*, *27*(3), 235–254. https://doi.org/10.1177/014920630102700302

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

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of Applied Psychology*, *89*(1), 104–118. https://doi.org/10.1037/0021-9010.89.1.104

Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, *35*(2), 1-20. https://doi.org/10.1007/s10458-021-09515-9.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lumineau, F. (2017). How contracts influence trust and distrust. *Journal of Management*, *43*(5), 1553–1577. https://doi.org/10.1177/0149206314556656

Lyons, J. B., Wynne, K. T., Mahoney, S., & Roebke, M. A. (2019). Trust and human-machine teaming: A qualitative study. In *Artificial intelligence for the internet of everything* (pp. 101–116). Elsevier

Mach, M., Dolan, S., & Tzafrir, S. (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. https://doi.org/10.1348/096317909X473903

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Maulsby, D., Greenb erg, S., & Mander, R. (1993). Prototyping an intelligent agent through wizard of Oz. In: *Proceedings of the INTERACT'93 and CHI'93 conference on human factors in computing systems*, ACM Press (pp. 277–284)

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human Factors*, *60*(2), 262–273. https://doi.org/10.1177/0018720817743223

McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021a). Trust and team performance in human–autonomy teaming. *International Journal of Electronic Commerce*, *25*(1), 51–72. https://doi.org/10.1080/10864415.2021.1846854

McNeese, N. J., Demir, M., Cooke, N. J., & She, M. (2021b). Team situation awareness and conflict: A study of human–machine teaming. *Journal of Cognitive Engineering and Decision Making*, *15*(2-3), 83–96. https://doi.org/10.1177/15553434211017354

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A Review and analysis of the empirical literature. *Human Factors*, *64*(5), 904-938. https://doi.org/10.1177/0018720820960865.

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. https://doi.org/10.1080/00140139.2012.691554

Pak, R., Rovira, E., McLaughlin, A. C., & Baldwin, N. (2017). Does the domain of technology impact user trust? Investigating trust in automation across different consumer-oriented domains in young adults, military, and older adults. *Theoretical Issues in Ergonomics Science*, *18*(3), 199–220. https://doi.org/10.1080/1463922X.2016.1175523

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55. https://doi.org/10.1145/975817.975844

Patel, L., Elliott, A., Storlie, E., Kethireddy, R., Goodman, K., & Dickey, W. (2020). Ethical and legal challenges during the COVID-19 pandemic—are we thinking about rural hospitals? *The Journal of Rural Health*, *37*(1), 175-178. https://doi.org/10.1111/jrh.12447.

Quinn, D. B., Pak, R., & De Visser, E. J. (2017). Testing the efficacy of human-human trust repair strategies with machines. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1794–1798. https://doi.org/10.1177/1541931213601930

Rebensky, S., Carmody, K., Ficke, C., Nguyen, D., Carroll, M., Wildman, J., & Thayer, A. (2021). Whoops! Something went wrong: Errors, trust, and trust repair strategies in human agent teaming. In: *International Conference on Human-Computer Interaction*, Springer International Publishing, pp. 95–106

Reed, G. S., Petty, M. D., Jones, N. J., Morris, A. W., Ballenger, J. P., & Delugach, H. S. (2016). A principles-based model of ethical considerations in military decision making. *The Journal of Defense Modeling and Simulation*, *13*(2), 195–211. https://doi.org/10.1177/1548512915581213

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people*. Cambridge University Press

Reidenbach, R. E., & Robin, D. P. (1988). Some initial steps toward improving the measurement of ethical evaluations of marketing activities. *Journal of Business Ethics*, *7*(11), 871-879. https://doi.org/10.1007/BF00383050.

Reidenbach, R. E., & Robin, D. P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of Business Ethics*, *9*(8), 639–653. https://doi.org/10.1007/978-94-007-4126-3_3

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1–19. https://doi.org/10.1016/j.obhdp.2006.05.005

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019) I don't believe you": Investigating the effects of robot trust violation and repair. In: *2019 14th ACM*, IEEE International Conference on Human-Robot Interaction (HRI), pp. 57–65

Sutton, G. W., Washburn, D. M., Comtois, L. L., & Moeckel, A. R. (2006). Professional ethics violations gender, forgiveness, and the attitudes of social work students. *Journal of College and Character*, *7*(1), 1–7. https://doi.org/10.2202/1940-1639.1501

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007) *Using multivariate statistics* (5). Pearson Boston

Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., Pak, R., Tossell, C., & De Visser, E. J. (2022). Exploring the Relationship Between Ethics and Trust in Human-AI Teaming: A Mixed Methods Approach. *Journal of Cognitive Engineering and Decision Making*. https://doi.org/10.1177%2F15553434221113964.

Thompson, M. M., Thompson, M. H., & Adams, B. D. (2008). *Moral and ethical dilemmas in Canadian forces military operation: Qualitative and descriptive analyses of commanders' operational experiences*. Defence Research and Development Toronto. https://apps.dtic.mil/sti/citations/ADA505336

Winfield, A., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 1-13. https://doi.org/10.1098/rsta.2018.0085.

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–25. https://doi.org/10.1145/3432945

Beau Schelble, B.S., is a PhD student at Clemson University studying Human-Centered Computing in the Team Research in Computational Environments (TRACE) Research Group within the School of Computing. His research interests lie in human-machine teaming, multi-agent systems, and human-AI interaction.

Jeremy Lopez, M.S., Jeremy Lopez is currently pursuing a PhD in Human Factors Psychology at Clemson University. His research interests include human-automation interaction with multiple automated systems and human-autonomy teaming.

Nathan J. McNeese, PhD, is the College of Engineering, Computing and Applied Sciences Dean's Professor, an Assistant Professor of Human-Centered Computing, and the Director of the Team Research Analytics in Computational Environments (TRACE) Research Group within the School of Computing, Clemson University. His research interests and expertise include human-AI teaming and human-centered AI.

Richard Pak, PhD, is currently a Professor in the Department of Psychology at Clemson University. He received his PhD in psychology in 2005 from the Georgia Institute of Technology.

Claire Textor, M.S., is a Human Factors Psychology PhD student in the Cognition, Aging, and Technology Lab at Clemson University. She earned an MS in Applied Psychology from Clemson University in 2020 and a BS in Psychology from the University of Illinois at Urbana-Champaign in 2018.

Rui Zhang, M.S., is a Human-Centered Computing PhD student in the Team Research and Analytics in Computational Environments (TRACE) Group at Clemson University. She earned an MS in Engineering from the Beijing Institute of Technology in 2018.

Guo Freeman, PhD, is currently an assistant professor and the Director of the Gaming and Mediated Experience Lab within the division of Human-Centered Computing in the School of Computing, Clemson University. She received her PhD in Information Science in 2015 from Indiana University. Her research focuses on computer-mediated interpersonal relationships and group behaviors.

Date received: November 14, 2021
Date accepted: July 12, 2022