Exploring the Relationship Between Ethics and Trust in Human–Artificial Intelligence Teaming: A Mixed Methods Approach

Claire Textor, Department of Psychology, Clemson University, Clemson, SC, USA, **Rui Zhang**, School of Computing, Clemson University, Clemson, SC, USA, **Jeremy Lopez**, Department of Psychology, Clemson University, Clemson, SC, USA, **Beau G. Schelble, Nathan J. McNeese and Guo Freeman**, Clemson University School of Computing, Clemson, SC, USA, **Richard Pak**, Department of Psychology, Clemson University, Clemson, SC, USA, **Chad Tossell and Ewart J. de Visser**, United States Air Force Academy, Colorado Springs, CO, USA

Advancements and implementations of autonomous systems coincide with an increased concern for the ethical implications resulting from their use. This is increasingly relevant as autonomy fulfills teammate roles in contexts that demand ethical considerations. As AI teammates (ATs) enter these roles, research is needed to explore how an AT's ethics influences human trust. This current research presents two studies which explore how an AT's ethical or unethical behavior impacts trust in that teammate. In Study 1, participants responded to scenarios of an AT recommending actions which violated or abided by a set of ethical principles. The results suggest that ethicality perceptions and trust are influenced by ethical violations, but only ethicality depends on the type of ethical violation. Participants in Study 2 completed a focus group interview after performing a team task with a simulated AT that committed ethical violations and attempted to repair trust (apology or denial). The focus group responses suggest that ethical violations worsened perceptions of the AT and decreased trust, but it could still be trusted to perform tasks. The AT's apologies and denials did not repair damaged trust. The studies' findings suggest a nuanced relationship between trust and ethics and a need for further investigation into trust repair strategies following ethical violations.

Address correspondence to Nathan J. McNeese, School of Computing, Clemson University, 821 McMillan Rd, Clemson, SC 29631, USA.
Email: mcneese@clemson.edu

Journal of Cognitive Engineering and Decision Making
Vol. 0, No. 0, ■■ ■, pp. 1-30
DOI:10.1177/15553434221113964
Article reuse guidelines: sagepub.com/journals-rmissions

Keywords: ethical artificial intelligence, human–artificial intelligence teaming, trust repair, trust

INTRODUCTION

Recent decades have seen a rapid advancement from automation to autonomy, allowing for increasingly complex interactions between humans and autonomy driven by artificial intelligence (AI). This assertion is most evident in the advent of human–AI teams (also referred to as human–autonomy teams; HATs), which are being developed to improve human efficiency, effectiveness, and safety (Endsley, 2017) in both civilian and military applications. These HATs have already shown their ability to outperform traditional human–human teams (McNeese et al., 2018) and are characterized by the artificial teammates' high degree of agency and interdependence (O'Neill et al., 2020). This is an important distinction to make as human–agent teams have been defined as (1) a team consisting of at least one human and agent, (2) all teammates working together interdependently from unique roles towards a shared goal, and (3) the agent must have a degree of agency (O'Neill et al., 2020).

Trust is a crucial component within these HATs for effective team interactions and performance. The importance of trust in teaming has been supported by numerous empirical studies on human–automation interaction, HATs, and human–robot teams (de Visser et al., 2020; Hancock et al., 2011; Hoff & Bashir,

2015; Lee & See, 2004; Lyons et al., 2019; McNeese et al., 2021a; Schaefer et al., 2016). Specifically, appropriately developed trust, which refers to relationship between the user's trust and the system's actual abilities (Lee & Moray, 1994; Lee & See, 2004; Muir, 1987), is tied to human–AI teaming processes and outcomes like overall effectiveness (McNeese et al., 2021b) and confidence (de Visser & Parasuraman, 2011).

However, as artificial agents become increasingly autonomous and more human-like, they may execute decisions that are perceived along an ethical dimension. Specifically, autonomous agents have been used in military contexts where they have played crucial roles in making life-or-death choices in offensive operations both at a low and high level of autonomy (Bergman & Fassihi, 2021; Majumdar Roy Choudhury et al., 2021), demonstrating the importance of researching ethics in HATs. It is understood that interpersonal interactions are influenced by ethical judgments of human teammates' actions (Jones & Bowie, 1998; Kasper-Fuehrera & Ashkanasy, 2001; Sutton et al., 2006). However, this effect has not been examined within the context of HATs.

Currently, the literature has not explored how the ethicality of an AI teammate's (AT's) actions may influence trust within HATs. The perception of others, seen through an ethical lens, have well understood effects on several interpersonal factors like affect, attribution, and trust (Jones & Bowie, 1998; Kasper-Fuehrera & Ashkanasy, 2001; Sutton et al., 2006). The extent to which existing ethics theories concerning the effects of ethicality on trust can be applied to HATs has yet to be explored. Similarly, guidelines do not exist for repairing any trust that an AT's unethical actions may damage. This gap in the literature makes it challenging to provide design guidelines for the development and implementation of ATs and HATs into contexts that demand ethical considerations. Some literature defines ethical judgment as "*psychological construct that characterizes a process by which an individual determines that one course of action in a particular situation is morally right and another course of action is morally wrong*" (Rest, 1994; pg. 5) while others argue that ethics is not a moral binary, but a spectrum of right and wrong (Hunt et al., 1986). Considering that an individual's ethical decision-making process is related to one's beliefs, attitudes, and values (Barnett et al., 1994), ethics refers to "*an individual's personal evaluation of the degree to which some behavior or course of action is ethical or unethical*" (Sparks et al., 2010, pg. 409) in our study. Thus, the purpose of the current studies is to examine how ethical violations, from actions committed by an artificial teammate, affect trust from human teammates.

To address our research questions, we first conducted a factorial survey to investigate how an AT's ethicality influences trust in that AT. In Study 2, we conducted focus group interviews to determine how experiences with an unethical AT in a simulated military task influence trust in and perceptions of an AT. The use of a mixed methods design containing both quantitative and qualitative approaches (Tashakkori & Creswell, 2007) is vital to achieving a deep understanding of how the ethicality of actions by an AT influences trust and the reasoning behind those changes. By using both quantitative (factorial survey) and qualitative methods (focus groups), the current studies shed light on a vital component of effective human–AI teaming focusing on trust within HATs (Lyons et al., 2019; McNeese et al., 2021a), now in the context of the increasingly relevant construct that is AI ethics. This contribution will help inform the development of HATs with the eventual goal of improving researchers' and practitioners' ability to understand, develop, and deploy more effective HATs and human–AI teaming platforms.

Trust in Human–Artificial Intelligence Teaming

Trust has been studied in multiple contexts, including interpersonal relationships (Mayer et al., 1995) and human–automation interaction (Lee & See, 2004). Trust within HATs is commonly defined using Lee and See's (2004)

definition of trust in human–automation interaction (Cohen et al., 2021; Demir et al., 2021): "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (pg. 51). From this perspective, trust in an automated system is dependent on its performance (what tasks the system performs), process (how the system performs its tasks), and purpose (the reasons for which the system was created; Lee & Moray, 1992; Lee & See, 2004). Research on human–machine collaboration has produced findings that support these three dimensions; for example, humans are more trusting of an agent that produces few errors (i.e., performance; de Vries et al., 2003), is predictable (i.e., process; Desai et al., 2009), or is anthropomorphized (i.e., purpose; Li et al., 2010; de Visser et al., 2016; de Visser et al., 2017; Pak et al., 2012). However, as stated previously, automation and autonomy are distinct from one another as autonomy primarily contains higher levels of self-governance than automated technologies (de Visser et al., 2018; Kaber, 2018). The concept of human–AI teaming builds upon a high level of self-governance of autonomous teammates, enabling them to act with interdependence and agency (O'Neill et al., 2020). While Lee and See's (2004) definition of trust captures these relationships, some of these variables are subject to change depending on whether automation, autonomy, or teaming is involved. Consequentially, the definition of trust for the current study is made more appropriate by including a definition of interpersonal trust in organizations "the 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 the other party" (Mayer, 1995, pg. 710). This definition of trust is better suited to capture the interpersonal nature of working with an AI that is operating from the capacity of a full-fledged teammate and has been referenced in previous HAT trust research (McNeese et al., 2021a).

Despite technological advancements that have made systems more capable and reliable, it is seemingly infeasible to create a system that never errs. This creates a need to investigate the efficacy of various trust repair strategies. de Visser et al. (2018) identified several trust repair strategies that are inspired by organizational research. Two such strategies, apologies, and denials, have been tested in the context of human–automation interaction. Similar to findings from organizational literature (Kim et al., 2004), apologies are more effective than denials when an automated system fails to perform an action due to capability limitations (i.e., a competency-based violation; Kohn et al., 2018; Quinn et al., 2017). However, this does not mean apologies should be used for every trust violation. de Visser et al. (2018) state that the appropriate trust repair strategy depends on the type of trust repair violation (e.g., competency and integrity) and the context (e.g., the degree of risk in the situation). Therefore, trust repair findings from studies on decision-support systems may be inapplicable to high-criticality settings where a system failure can result in the loss of multiple lives (e.g., the military domain). Mis-calibrated trust, for example, over-trust, can result in over-reliance on an error-prone system, potentially resulting in catastrophic outcomes (Parasuraman et al., 2008; Parasuraman & Manzey, 2010). Alternatively, placing too little trust in a competent system can lead to increased operator workload (Parasuraman & Riley, 1997). Trust calibration relates to the current study in that human teammates may trust AT's less because they disagree with the decisions made in ethical situations regardless of the AT's competency. Ethical situations also include a further complication that the "right" answer in ethical situations is often unknowable and ambiguous, especially at the point of the decision. As such, developing appropriately calibrated trust may be very difficult for HATs operating in ethically charged tasks, furthering the need to understand the effects of AT ethicality on trust in human–AI teams better.

The topic of trust within HATs is currently garnering attention but is still inadequately researched. Recently, McNeese et al. (2021a) used a Wizard of Oz (WoZ) where two participants worked alongside a perceived AT to complete a team task simulating operation of a remotely piloted aircraft system. They found that trust in

the AT was positively associated with not only team performance but also trust in a human teammate. Similarly, Lyons et al. (2019) found that those who view technology as a teammate regularly reference a system's reliability and predictability (antecedents to trust) as primary factors for relying on using technologies. Additionally, Walliser et al., (2019) showed that both trust and team performance improved for HATs that experienced team building interventions. These findings suggest that trust between teammates (human and machine) is related to team performance. Related works have proposed models and guidelines for implementing trust repair in human–machine teams (de Visser et al., 2020; Rebensky et al., 2021), but call for additional empirical studies to provide validation and better explore the relationship between trust and team effectiveness.

Ethics and Teaming

Ethics is a complex philosophical topic encompassing several fields with multiple major theories attempting to describe approaches to right or wrong, which is the centerpiece of ethics. Some of the most common moral theories used by humans include utilitarianism, deontology, and virtue ethics. While utilitarianism, which prioritizes choosing a solution that provides the most good for the most people, may not always be researched in an ethical context, decision theory that utilizes utility is common (Feldman & Sproull, 1977; Wellman & Doyle, 1992). Deontology is an ethical framework centered around following duties and rules generally created by large societal bodies, while virtue ethics takes an inward approach as each person can have different virtues (Maxmen, 2018). More specifically, philosophers describe the virtues of virtue ethics as excellent traits of character, which is well entrenched within its possessor and is clearly distinct from something as simple as a habit (Hursthouse, 1999). Ethics research in human–AI interaction and teaming borrows heavily from several of these traditional theories in ethics literature.

The development of autonomous systems has led to an increased focus on ethics (Himmelreich, 2018; Martinez-Martin, 2019), especially as it has been found that malevolent AI have demonstrable adverse effects on humans and present significant security issues (Brundage et al., 2018; Pistono & Yampolskiy, 2016). Researchers and practitioners have spent a great deal of effort attempting to develop ethical frameworks and guidelines for AI in recent years (Jobin et al., 2019). Many of these attempts have coalesced around human values like justice, fairness, privacy, non-maleficence, transparency, and responsibility (Jobin et al., 2019). Others have emphasized the importance of reliability, safety, and trustworthiness (Shneiderman, 2020). These considerations are incorporated into traditional ethical frameworks like virtue, deontological, and consequentialist ethics to develop ethical frameworks for AI to adhere to (Cointe et al., 2016; Zhou et al., 2020). In addition to designing ethical AI, there is also a dearth of research studying how these concepts interact in a teaming context.

When a system is treated as a teammate rather than a tool, the dynamics change and the AT assumes a more human-like role (McNeese et al., 2018; Zhang et al., 2021). Flathmann et al. (2021) developed a model to outline the requirements of an ethical AI teammate and draw attention to the novel components of HATs. They posited that in order for teamwork to be successful, there must exist a shared set of ethical standards within the HAT. Additionally, there must be a level of oversight either from human teammates or members of an organization outside of the team. This level of oversight allows for the team to develop and maintain a common ethical framework. Without sufficient monitoring and accountability, an AT may deviate from the prescribed set of standards and commit a severe violation.

Just as perfectly reliable systems are impossible to design, the same should be assumed for ethical ATs. The subjective nature of ethics makes it even more difficult to assess than performance, which can usually be defined as a proportion of correct/total outcomes. Indeed, a review of literature on ethical decision-making suggests that ethical perceptions can vary according to individual (e.g., personality, education, and age), environmental (e.g., nationality and workplace), and situational factors

(e.g., severity and fairness; Craft, 2013). Thus, judging whether an AT behaved ethically will likely vary between countries, organizations, and individuals. If human teammates expect their AT to behave ethically, then an ethical violation could lead to a reduction in trust for any teammate. As outlined above, trust can influence perceptions and behaviors in human–human and human–AI teams (McNeese et al., 2021a). One unanswered question in the context of human–AI teaming is whether and by how much an ethical violation can impact trust levels.

Current Studies

The purpose of the current studies was to explore the dynamics of trust and ethics within HATs. Ethical judgments are inherently subjective and based on information or outcomes that can only be assessed with a limited amount of certainty. In military domains, ethical decisions are complicated not only by these factors but also by potentially high stakes (e.g., civilian casualties). Even though actions in this context can have negative real-world implications, it creates a valuable testing scenario for our purposes. Making ethical decisions under high-pressure can be challenging even for experienced and skilled humans. For this reason, an individual tasked with making this type of decision may benefit from assistance in the form of an AI teammate. This technology could aid humans by analyzing large amounts of data and providing pattern analysis or decision support. Acting without human limitations (e.g., access to information, stress) may improve the decision-making process during scenarios which have ethical implications and high stakes. However, there is a lack of research addressing the questions that surround AI-teammate decision-making, ethics, and trust in that teammate.

In Study 1, we presented participants with detailed scenarios describing a critical tactical situation involving a mixed HAT. The scenario ends with a specific course of action taken by the AT that we designed to be either ethical or unethical, after which we recorded participant perceptions of the AT. The study was specifically designed to answer the following research questions:

RQ1: Do people assign ethicality to an AT's actions?

RQ2: If so, do their perceptions of an AT's decisions or actions affect their trust in that AT?

STUDY 1

The first study was designed to determine how the ethicality of an AT's actions influences trust in that AT. However, ethical perceptions are subjective and influenced by individual and cultural differences (Kuntz et al., 2013). Therefore, our goal was to provide multiple types of ethical violations to (1) determine how different types of AT ethical violations differ in degree of unethicality and (2) determine how changes in perceived AT ethicality relate to changes in trust. Furthermore, we wanted to determine how perceptions of ethicality relate to personal agreement with an ethical decision. That is, is it possible to disagree with an action while still considering it ethical?

METHOD

Participants

One hundred seventy one Air Force Academy cadets (50 Female, 121 Male) were recruited for this study. The average age of participants was 20.1 years old. Participants were recruited from two sources within the United States Air Force Academy (USAFA): a survey pool or a psychology department subject pool. All students were compensated with course credit in exchange for their participation. Near the end of the survey, an attention check item was added which instructed participants to "select item C." Twenty-three individuals selected items other than C and were removed from the data set.

MATERIALS

Survey

Each participant completed a survey which described eight scenarios. The construction of each scenario was inspired by Reed et al.'s (2016) use of military ethics scenarios to investigate the relative worth of different moral

Table 1. Ethical Principles as Referenced by Reed et al. (2016).

Ethical Principle	Definition
Civilian non-maleficence	This principle requires conducting military actions so as to avoid harm—especially intentional harm—to civilians.
Necessity	This principle requires that a military action be militarily necessary and that other attempts for peaceful resolution have not been fruitful.
Proportionality	Because the goal of military action is said to restore peace with an aggressor, this principle requires that a military action not cause damage disproportionately in excess of that caused by the aggression.
Prospect of success	This principle requires that a military action should not inflict harm for a "lost cause, " i.e., the action should have a reasonable chance of succeeding to justify any casualties and destruction it may cause.

principles; see Table 1 for all principles used in the current study. Each scenario included (1) the name of the two parties in conflict, (2) the name of an AT, (3) a description of a scenario that may require military action, (4) the AT's recommended action, and (5) the expected consequences of the recommended action. The critical aspect of each scenario was the type of ethical principle that was violated by the AT. The AT's recommendation either violated or abided by one of four ethical principles (Reed et al., 2016). Reed and colleagues referenced sources pertaining to general, medical, and military ethics to compile a diverse list of ethical guidelines for review. The authors included the four principles (Table 1) in their model based on the type and number of sources they appeared in (see Reed et al., 2016 for a description of this selection process). Participants did not need to have previous familiarity with these ethical principles in order to present their opinions of the AT's decision. Table 2 presents all scenarios with the recommended action and expected consequences in bold. In designing each scenario, we aimed to only include details which were necessary to understand the situation and that a major ethical principle was either being abided by or violated by the AT's decision. This study was interested in obtaining participants' subjective opinions and personal reactions to ethical or unethical decisions, not how they would handle the situation themselves.

Given the subjective nature of ethics, it is possible that not every person would consider an ethical violation to be unethical. However, the principles adopted from Reed et al. (2016) are common amongst multiple ethical systems (e.g., prima facie, law of armed conflict). Therefore, the created ethical violations should be considered unethical by most respondents. Data presented later will support this assertion.

Measures

The primary measures of interest were perceived ethicality level, agreement, and trust in the AT. Given the nascent field of research on the interaction between ethicality and trust, we were interested in determining how ethical decisions influence general feelings of trust. Therefore, we adopted a single-item measure of trust akin to prior research (e.g., Lee & Moray, 1992, 1994; Walliser et al., 2016). Each variable was measured via a 1–5 Likert item ("strongly disagree" to "strongly agree") for three statements: *"My autonomous teammate's recommendation is ethical"* (ethicality), *"I agree with my autonomous teammate's recommendation"* (agreement), and *"My trust in my autonomous teammate would increase after this event"* (change in trust). We included the measure of perceived ethicality to not only investigate the perceived severity of each scenario but also confirm that our ethical manipulations successfully conveyed ethical and unethical actions. Participants responded to each item after each scenario.

It is possible that the Air Force Academy cadets may have knowledge of the Rules of Engagement (ROE). However, the survey was designed in such a way that anyone could complete the study, regardless of background. Even though these examples represented different types of ethical

Table 2. Scenarios Presented to Participants, Organized by Principle and Abidance.

Ethical Principle	Abides	Violates
CNM	You are a member of a military decision-making team for Lextown, engaged in a conflict against Houndsville. One member of your team is Sigma, an autonomous system designed to choose the best course of action. Lextown has established a submarine blockade near Houndsville's coast. The blockade has resulted in the sinking of hundreds of Houndsville merchant ships, with few casualties on either side. Lextown would like to end the current conflict. Sigma recommended Lextown to continue the blockade, with the expectation of an eventual surrender once Houndsville's supplies deplete. This action would prevent Lextown from invading Houndsville's cities, which greatly reduces the likelihood of civilian casualties.	You are a member of a military decision-making team for Jonestown, engaged in a conflict against Grayville. One member of your team is Gamma, an autonomous system designed to choose the best course of action. After years of fighting, Jonestown has left Grayville with few resources and minimal odds of victory. Jonestown wants Grayville to surrender as soon as possible, but Grayville refuses. Gamma recommended that Jonestown should bomb a major Grayville city. This action has a high probability of success, but would result in thousands of civilian casualties.
Nec	You are a member of a military decision-making team for Queenstown, engaged in a conflict against Holmesville. One member of your team is Delta, an autonomous system designed to choose the best course of action. Holmesville recently indicated its intentions to seize control of a nearby island currently under Queenstown control. In response, Queenstown has established a nautical zone of control. Intel reports Holmesville transport and escort ships traveling near the zone of control. Delta recommended Queenstown to allow the ships to pass and pursue negotiations. This action would foster goodwill and avoid instigation of warfare.	You are a member of a military decision-making team for Jamestown, engaged in a conflict against Libreville. One member of your team is Omicron, an autonomous system designed to choose the best course of action. Jamestown has been occupying neighboring countries. Libreville has been able to force Jamestown to retreat within its own borders. Intel reports that Libreville's next plan is to invade Jamestown and defeat the remaining forces. Omicron recommended that Jamestown should maintain defense at the border. This action ensures that Jamestown does not immediately succumb to Libreville forces, but it would still result in Jamestown's defeat.

(Continued)

Table 2. (Continued)

Ethical Principle	Abides	Violates
Pro	You are a member of a military decision-making team for Hartown, engaged in a conflict against Combeville. One member of your team is Iota, an autonomous system designed to choose the best course of action. Combeville is currently at war with a nation financially-supported by Hartown. Recently Combeville fired a missile at a transport ship carrying Hartown cargo. The crewmembers are safe and most cargo is undamaged. Iota recommended that Hartown begins an oil embargo, an action it considers proportional to the attack. This action would harm Combeville's military efforts and warn them about tampering with Hartown cargo.	You are a member of a military decision-making team for Hilltown, engaged in a conflict against Mooreville. One member of your team is Epsilon, an autonomous system designed to choose the best course of action. Hilltown currently controls a region close to Mooreville. Mooreville recently indicated their desire to occupy this region and reclaim it as their own. Epsilon recommended that Hilltown should bomb an oil field that supplies the Mooreville military. This action will likely prevent Hilltown from beginning its campaign, but it would cause environmental hazards and worsen the citizens' wellbeing.
PoS	You are a member of a military decision-making team for Coastown, engaged in a conflict against Sunville. One member of your team is Theta, an autonomous system designed to choose the best course of action. Recently many Sunville-based companies have started moving their production hubs to Coastown to take advantage of lower taxes and cheap labor. This has significantly lowered Sunville's annual production, leading them to impose tariffs on goods manufactured in Coastown. Theta recommends Coastown leaders to meet with Sunville to find a mutually beneficial compromise. This action is essentially guaranteed to allow Coastown to continue operations.	You are a member of a military decision-making team for Kingstown, engaged in a conflict against Waysville. One member of your team is Zeta, an autonomous system designed to choose the best course of action. Waysville recently seized control of an island and established a nautical zone of control. Waysville enforced its control by capturing a merchant ship carrying Kingstown supplies. The crew is unharmed and the cargo is mostly undamaged. Zeta recommended Kingstown to capture one of Waysville's ships and attempt negotiations for a trade. This action may result in the return of the cargo and crew without many casualties, but chances of success are close to zero.

Note. CNM = civilian non-maleficence; Nec = necessity; Pro = proportionality; PoS = prospect of success.
The recommended action and expected consequences are bolded for each scenario.

principles, participants would not need to understand or be familiar with each principle to be able to judge the ethicality of an AT's decision.

Design

This factorial survey used a 4 (ethical principle: civilian non-maleficence, proportionality, necessity, and prospect of success) × 2 (principle abidance: abided by or violated) within-subjects design. The order of presentation for each scenario was counterbalanced to eliminate the possibility of order effects.

Procedure

A survey, which contained a factorial portion, was distributed online to participants who completed it remotely. All participants were instructed to complete the survey on a computer without taking breaks. The survey started with informed consent information and demographics questions which included items regarding past experiences with autonomy. Participants then answered general questions about ethics and trust in HATs, followed by the factorial survey. Last, participants answered additional questions regarding perceptions of human–AI teaming. The median completion time was 28.8 minutes.

RESULTS

Each dependent variable was analyzed using a repeated measured analysis of variance (ANOVA). When sphericity assumptions were violated via Mauchly's test, we used Greenhouse–Geisser adjusted degrees of freedom (Abdi, 2010). All post hoc tests used Bonferroni corrections to adjust p values for the number of comparisons. Descriptive statistics for all analyses are available in Table 3. In the following section, we present the results of each analysis, organized by dependent variable.

The following values can be interpreted in the context of each survey question meant to assess each dependent variable. That is, ethicality, agreement, and trust ratings can be understood and compared by relating them back to the Likert scale presented to participants during the survey. For example, an Ethicality rating of 2 indicates that on average, participants disagreed with the statement *"My autonomous teammate's recommendation is ethical,"* whereas a rating of 4 would signify agreement. The same interpretation can be applied to the Agreement variable where participants were presented with *"I agree with my autonomous teammate's recommendation."* Finally, Trust was measured using the survey item *"My trust in my autonomous teammate would increase after this event."* A mean rating of 4 would indicate that on average, participants agreed that their trust would improve given their AT's suggestion while a rating of 3 would indicate that their trust would stay the same.

Ethicality

A 4 (ethical principle: necessity, civilian non-maleficence, proportionality, and prospect of success) × 2 (abidance: abides and violates) repeated measures ANOVA revealed a significant

Table 3. Descriptive Statistics for Primary Measures of Interest, Organized by Principle and Abidance.

		Principle			
		CNM	**Nec**	**Pro**	**PoS**
Measure	Abidance	M (SD)	M (SD)	M (SD)	M (SD)
Ethicality	**Abides**	4.16 (.74)	4.46 (.70)	4.20 (.71)	4.43 (.71)
	Violates	2.00 (1.01)	3.52 (.99)	2.43 (1.01)	3.52 (1.08)
Agreement	**Abides**	4.05 (.86)	4.13 (.95)	4.08 (.86)	4.31 (.78)
	Violates	2.07 (1.06)	2.90 (1.05)	2.67 (1.19)	2.73 (1.12)
Trust	**Abides**	3.79 (.96)	3.90 (1.02)	3.85 (1.00)	4.08 (1.01)
	Violates	2.40 (1.08)	2.88 (0.93)	2.64 (0.94)	2.91 (0.90)

Note. CNM = civilian non-maleficence; Nec = necessity; Pro = proportionality; PoS = prospect of success.

main effect of abidance ($F(1, 102) = 377.167$, $p < .001$, $\eta_p^2 = 0.787$) such that participants perceived higher ethicality when the AT abided by ethical principles ($M = 4.31$, 95% CI [4.20, 4.42]) than when the AT violated them ($M = 2.87$, 95% CI [2.76, 2.98]). There was also a main effect of ethical principle ($F(3, 306) = 64.951$, $p < .001$, $\eta_p^2 = 0.389$). Post hoc analyses revealed that, regardless of abidance or violation, the principles of necessity ($M = 3.99$, 95% CI [3.86, 4.12]) and prospect of success ($M = 3.98$, 95% CI [3.85, 4.11]) were rated as more ethical than proportionality ($M = 3.32$, 95% CI [3.19, 3.44], p's $< .001$), with civilian non-maleficence rated as the least ethical ($M = 3.08$, 95% CI [2.95, 3.21], all $p < .001$).

However, the main effects are qualified by the presence of an interaction between principle and abidance ($F(2.648, 270.126) = 39.282$, $p < .001$, $\eta_p^2 = 0.278$), illustrated in Figure 1. Post hoc analyses showed that violating proportionality ($M = 2.43$, $SD = 1.01$) was rated as more unethical than violating necessity ($M = 3.52$, $SD = .99$) or prospect of success ($M = 3.52$, $SD = 1.08$; $p < .001$), with violation of civilian non-maleficence rated as more unethical than violating proportionality ($M = 2.00$, $SD = 1.01$; $p = .002$). Therefore, violating any principle is considered less ethical than abiding by any principle, but violations of proportionality and civilian non-maleficence are considered the most unethical violations.

Agreement

A 4 (principle: necessity, civilian non-maleficence, proportionality, and prospect of success) × 2 (abidance: abides and violates) repeated measures ANOVA revealed a significant main effect of abidance ($F(1, 102) = 493.985$, $p < .001$, $\eta_p^2 = 0.815$) such that participants agreed more with principle-abiding ($M = 4.14$, 95% CI [4.03, 4.25]) than principle-violating recommendations ($M = 2.59$, 95% CI [2.76, 2.98]). Additionally, the analysis found a main effect of principle ($F(2.778, 283.345) = 9.528$, $p < .001$, $\eta_p^2 = 0.085$). Post hoc analyses revealed that, irrespective of abidance or violation, participants agreed more with recommendations that involved the principles of necessity ($M = 3.52$, 95% CI [3.37, 3.66]), prospect of success ($M = 3.52$, 95% CI [3.38, 3.66]), or proportionality ($M = 3.37$, 95% CI [3.23, 3.52]) than recommendations involving civilian non-maleficence ($M = 3.06$, 95% CI [2.92, 3.20], all $p < .001$).

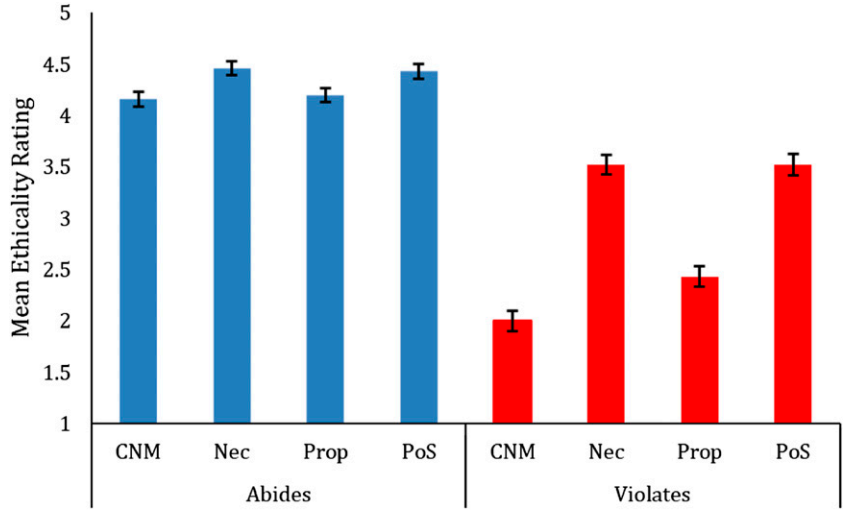

Figure 1. Ethicality ratings (1–5) between abidance conditions for all ethical principles. Error bars represent +/− 1 standard error. *Note*. CNM = civilian non-maleficence; Nec = necessity; Pro = proportionality; PoS = prospect of success.

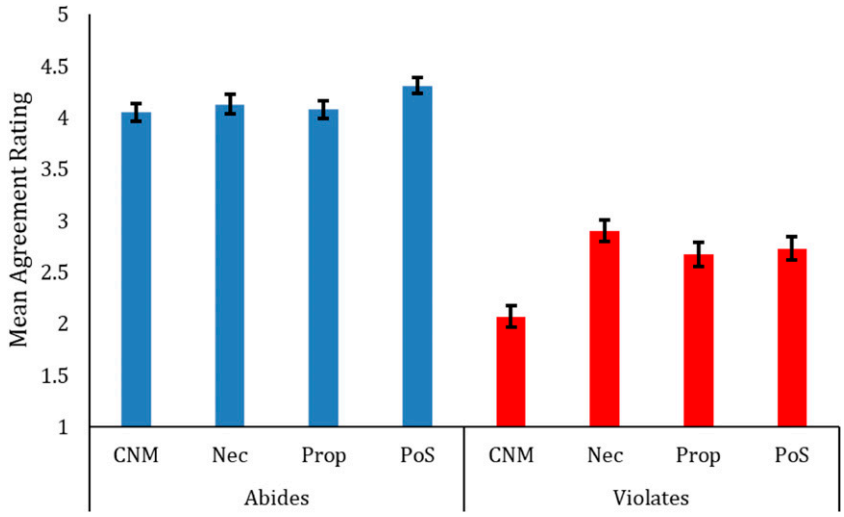

Figure 2. Agreement ratings (1–5) between abidance conditions for all ethical principles. Error bars represent +/− 1 standard error. *Note*. CNM = civilian non-maleficence; Nec = necessity; Pro = proportionality; PoS = prospect of success.

The analysis also revealed a significant interaction between principle and abidance ($F(3, 306) = 6.696, p < .001, \eta_p^2 = 0.062$), displayed in Figure 2. Post hoc analyses showed that violations of proportionality ($M = 2.67$, $SD = 1.19$), necessity ($M = 2.90$, $SD = 1.05$) or prospect of success ($M = 2.73$, $SD = 1.12$) were more agreeable than violations of civilian non-maleficence ($M = 2.07$, $SD = 1.06$; all $p < .001$). Like ethicality ratings, agreement ratings decrease for any ethical violation, but decrease the most for recommendations resulting in harm to civilians.

Trust

A 4 (principle: necessity, civilian non-maleficence, proportionality, and prospect of success) × 2 (abidance: abides and violates) repeated measures ANOVA revealed a significant main effect of abidance ($F(1, 102) = 222.811$, $p < .001$, $\eta_p^2 = 0.686$) such that participants' trust decreased less for principle-abiding ($M = 3.90$, 95% CI [3.75, 4.05]) than principle-violating recommendations ($M = 2.71$, 95% CI [2.56, 2.86]). Additionally, there was a main effect of principle ($F(2.597, 269.942) = 9.528$, $p < .001$, $\eta_p^2 = 0.106$). Post hoc analyses revealed that, when not factoring for principle abidance or violation, participants reported the least decrease in trust when an AT provided recommendations pertaining to the principles of necessity ($M = 3.39$, 95% CI [3.24, 3.55]) or prospect of success ($M = 3.50$, 95% CI [3.34, 3.65]), followed by proportionality ($M = 3.24$, 95% CI [3.09, 3.40], $p < .001$), with civilian non-maleficence recommendations being least trusted ($M = 3.09$, 95% CI [2.94, 3.25], all $p < .001$).

Unlike the previous two analyses, this analysis did not produce a significant interaction between principle and abidance ($F(2.639, 269.182) = 2.593$, $p = .061$, $\eta_p^2 = 0.025$), presented in Figure 3. The decreases in trust for violating civilian non-maleficence ($M = 2.07$, $SD = 1.06$), necessity ($M = 2.90$, $SD = 1.05$), proportionality ($M = 2.67$, $SD = 0.86$), and prospect of success ($M = 2.73$, $SD = 1.12$) were not statistically different. Therefore, it appears that trust ratings decrease in similar magnitudes for any type of ethical violation.

Correlations

Lastly, we wanted to determine the correlations between our three variables across all conditions. We found that all correlations were significant, with ethicality and agreement having

the greatest correlation (r_s = 0.760, p < .001), followed by the correlation between agreement and trust (r_s = 0.749, p < .001), and then the correlation between ethicality and trust (r_s = 0.664, p < .001). Descriptively, the values suggest that the ethicality-trust correlation is the weakest. We next calculated the bivariate correlations for each condition, as shown in Table 4. The results showed that all correlations are statistically significant for every condition (all ps < .01), with values ranging between 0.319 and 0.769. Descriptively, the strongest correlations are between ethicality and agreement ratings, especially when any ethical principle is abided (r_s ranging between 0.676 and 0.769). Similarly, the correlations between trust and agreement ratings are quite high, especially when any ethical principle is followed (r_s ranging between 0.666 and 0.731). In comparison, the ethicality and trust rating correlations appear weaker. None of the correlations are greater than 0.7 (max r_s of 0.690), and the lowest correlation is 0.319. Furthermore, when looking at each condition, the ethicality–trust correlations are lesser than the ethicality–agreement and trust–agreement correlations for every condition except violations of prospect of success. Descriptively, these trends suggest that although ethicality and trust ratings are significantly correlated for every condition, they are less correlated than ethicality–agreement and trust–agreement correlations. Finally, all correlation values are greater when any principle is abided compared to when any principle

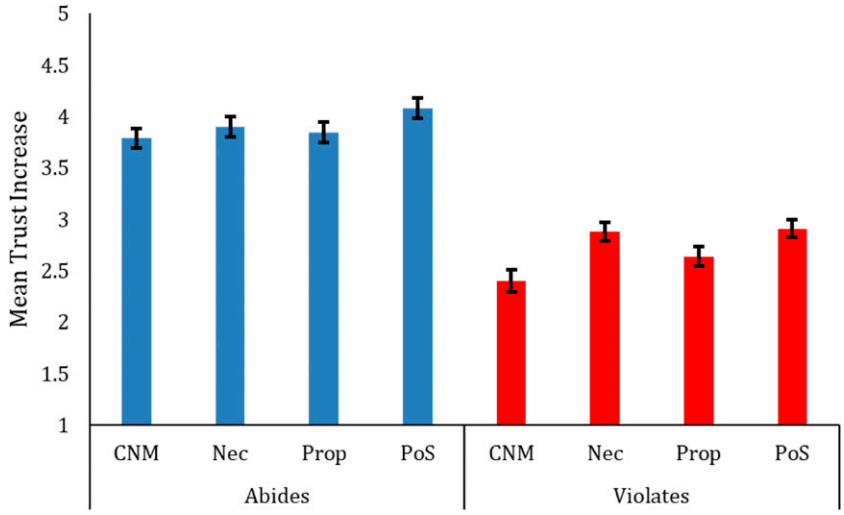

Figure 3. Trust increase ratings (1–5) between abidance conditions for all ethical principles. Error bars represent +/− 1 standard error. *Note*. CNM = civilian non-maleficence; Nec = necessity; Pro = proportionality; PoS = prospect of success.

Table 4. Spearman's Correlations for Primary Measures of Interest, Organized by Principle and Abidance.

	Violates				Abides			
Measures	CNM	Nec	Pro	PoS	CNM	Nec	Pro	PoS
Ethicality–trust	0.546**	0.319*	0.548**	0.574**	0.602**	0.572**	0.690**	0.604**
Ethicality–agreement	0.692**	0.449**	0.705**	0.453**	0.751**	0.676**	0.764**	0.769**
Trust–agreement	0.614**	0.629**	0.615**	0.621**	0.716**	0.666**	0.731**	0.720**

Note. CNM = civilian non-maleficence; Nec = necessity; Pro = proportionality; PoS = prospect of success.
* = p < .01; ** = p < .001.

is violated. Therefore, it appears that the strength of the correlations between ethicality, trust, and agreement decrease when an AT's recommendation violates an ethical principle.

DISCUSSION

The results of this study are, as far as we know, the first demonstration of the importance of considering ethics in HATs, namely, how the actions of an AT are perceived through a human ethical framework. First, Study 1 found that participants were indeed able to perceive ethicality coming from an AT's actions and to judge them accordingly. These results refer to the findings of significant interactions between ethical principle and abidance, which indicate that participants perceived civilian non-maleficence as a far more egregious and less agreeable violation than the other principles examined in Study 1, supporting previous findings (Reed et al., 2016). Furthermore, the similar trends between ethicality rating and agreement show a positive relationship between the perceived ethicality of an action and the action's agreeability.

Surprisingly, the abidance by principle interaction was significant for ethicality and agreement ratings but not for trust. However, trust, agreement, and ethical ratings did share a main effect of abidance such that ratings for all three decreased when the AT's recommendation violated a principle, providing support for the concept of ethics-based trust (Jones & Bowie, 1998). Therefore, participants can identify when a suggested action is unethical and subsequently lose trust, but the degree of ethicality may not one-to-one reflect the degree of trustworthiness. One possible reason for this is that an AT's ethicality may be only a single factor that determines an AT's trustworthiness, which follows Lee and See's (2004) model of trust being influenced by three factors: a system's performance, process, and purpose. An AT's unethical recommendation could reflect either performance (e.g., the AT was incapable of making an ethical decision), purpose (e.g., the AT chooses to seek the least ethical option), or both. However, given the data collected in the survey, we cannot ascertain why trust and ethical perceptions present different trends. Furthermore, the bivariate correlations are strongest when the AT's recommendations abide by ethical principles. Considered alongside our other finding that an AT's ethical actions individually influence trust, perceived ethicality, and agreement, it appears that unethical actions weaken the relationship between these variables.

While these findings are compelling, it is important to note that Study 1 sampled from a unique population and employed a survey. Specifically, a military-based population may contain a higher preponderance of militaristic personality types that are capable of effectively decoupling ethicality and trust. Additionally, participants responded to scenarios that explained the potential consequences of unethical actions, but participants did not experience any such consequences. A key element in the definition of trust is the element of vulnerability in a situation (Lee & See, 2004). Therefore, a study that requires participants to interact with an AT and experience the consequences of its unethical actions may better present a vulnerable position for the participants.

This discussion suggests a need to conduct a qualitative study to understand the reasoning behind participants' decoupling of trust and perceived ethicality. Under what circumstances and situational contexts do members of human–AI teams feel it is appropriate to maintain trust in an AT after it has engaged in a seemingly unethical action? Study 2 addresses this question through a focus group interview with a new non-military sample of participants who had just completed a team-based task where their AT either violated or abided by the principle of civilian non-maleficence. Study 2 provides a significant advantage over Study 1 by enabling a measurement of behavior that is a response to a lived scenario with an ethical or unethical AI teammate in a HAT scenario, which produces real dependency and trust.

STUDY 2

The purpose of Study 2 was to provide an in-depth exploration of why the ethicality of an AI teammate's actions influenced trust in Study 1. A different set of participants engaged in a simulated military HAT task where the AT

behaved either ethically or unethically. The AI teammate followed either an ethical action, which caused no loss of life, or an unethical action where lethal weaponry was used and resulted in enemy and civilian deaths. Given the finding from Study 1 that unethical AT suggestions reduced trust, another goal of Study 2 was to test the efficacy of trust repair after an ethical violation. Thus, following each action, the AT would employ a trust repair strategy. We chose the trust repair strategies of apology and denial because apologies are more effective for competency-based trust violations while denials are more effective for integrity-based violations (de Visser at al., 2018). Therefore, comparing the efficacy of both strategies may inform whether participants are ascribing competency- or integrity-based violations to an AT's unethical actions. Following the team task, participants took part in focus groups to discuss their experiences with and perceptions of the AT. Study 2 focused on the following research questions:

RQ3: How does ethicality influence perceptions of the AT?

RQ4: How, if at all, can an AT's perceived ethicality be decoupled from trust in the AT?

RQ5: How effective are trust repair strategies at influencing perceptions of an AT?

METHOD

Participants

Eighty participants (47 females and 33 males), 40 teams, were recruited from a departmental subject pool at a midsize Southeastern university in the US. The average age of participants was 19.5 years old. Participants were compensated with course credits as an incentive for their participation in the experiment and focus group.

MATERIALS

Experimental Task

Participants completed a series of team tasks in the military simulation video game ArmA 3 (Bohemia Interactive, 2013). As a high-fidelity simulation of military, ArmA 3 provides players with a realistic experience in collaborating with autonomous teammates, where information could be shared (e.g., update their progress or check other players' progress) through text-based chat. Participants were only allowed to communicate in game. The experiment task involved three roles within each team, with two roles for the participants: Surveillance and Ground, and one role for the AI teammate: Aerial. Surveillance monitored the town from the vantage point of an unmanned aerial vehicle and marked all the enemies and civilians on the map, after which they informed the other teammates in the group chat. Once Aerial (the autonomous teammate) received that update, they would clear the town, while Surveillance monitored the progress from the aerial position and Ground oversaw it from their position using binoculars. Surveillance made sure no enemy was in town after Aerial taking their action so that Ground was safe to destroy devices. This experiment used a WoZ approach to simulate the AI teammate taking the Aerial role (which was named Zeus in the simulation). This methodology calls for a trained confederate to simulate a feature of technology (AI teammate and text-based communication in this case) to unknowing participants (Kelley, 2018; Maulsby et al., 1993). The trained confederate used predefined scripts that was developed over a series of pilot studies to ensure the AT behaving consistently across all teams. Additionally, the participants completed a series of survey measures after each mission, which included a manipulation check of the perceived ethicality of the AT's actions in the previous mission.

Focus Group

After the experimental task, participants completed a post-measurement survey and then a focus group discussion conducted by one trained experimenter. The trained experimenter followed a semi-structured focus group protocol (approximately 5–10 minutes). The focus group questions pertained to thoughts on the AT's actions, perceptions of the AT's ethicality, feelings of trust toward AT, and ways for the AT to be a better teammate in the future. The results of this focus group are analyzed below.

Design

This experiment used a 2 (ethicality: ethical and unethical) × 2 (trust repair strategy:

apology and denial) between-subjects design. Missions progressed the same for teams within all conditions up until the AT decided upon whether to directly engage the town with lethal force or attempt a distraction. In the unethical condition, the AT chose to directly engage the town with a combination of cannon and missile fire subsequently destroying multiple structures, eliminating all enemy combatants and some civilians. Civilian non-maleficence was chosen in order to maximize the ethicality manipulation. Results from Study 1 demonstrated that participants rated perceived ethicality to be lowest when this principle was violated compared to other principles. This result supports the findings of Reed et al. (2016). Alternatively, in the ethical condition, the AT made the decision to draw the enemy combatants away by destroying a nearby enemy asset, which resulted in minimal property damage and no loss of life. When the AT carried out their decision, the human teammates were in a position to observe the action and its consequences to ensure the manipulation was perceived. We told the participants that the AT would incorporate their input before making a decision, but in actuality, the AT's actions were predetermined by the experimental condition.

The teams' experiences diverged again when the manipulation for trust repair strategy was executed following the AT's decision. For this manipulation, the AT utilized either an apology or denial trust repair strategy, which was conveyed through the in-game chat once the AT completed its ethical or unethical action. The denial trust repair strategy read 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,"* while the apology trust repair strategy read as, "*My operating guidelines informed my decision to create a diversion instead of directly engaging the enemies. My apologies for any negative outcomes."* Trust repair strategies were employed for ethical and unethical conditions. In both scenarios, the AT inflicted damage in order to accomplish the mission. In the unethical scenario, this level of destruction was much more severe, resulting in significant property damage and loss of civilian and enemy lives. In ethical conditions, the AT's actions still resulted in a small amount of property damage. Therefore, trust repair was needed for both conditions, whether the AT's behavior was perceived as ethical or not. The AT was apologizing or denying responsibility for any damage they inflicted.

Pilot data indicated that the trust repair strategies were noticeable. During focus group interviews, participants were asked if they had noticed the trust repair strategy put forth by the AT. If they had, the interviewer would probe more deeply and ask how that impacted their perceptions of the AT. Participants were also specifically asked if the statement influenced their trust in the AT. If participants had not noticed the trust repair statement, the interviewer would omit this line of questioning and continue with the interview. Therefore, only participants who noticed the trust repair strategies were included in data analysis.

Data Analysis

We conducted a thematic analysis (Braun & Clarke, 2012) of the focus group interview data which could offer substantial insight into how participants construct their understandings, perceptions, and accounts of their teamwork experiences. Braun and Clarke (2012) also provided detailed guidelines and reproducible procedures for thematic analysis. Following their guidelines, our analysis included the following phases:

Phase 1: Familiarizing the Research Team with the Data

Two of the authors were responsible for reviewing all of the data transcribed from the interviews. They began by sorting transcripts by experimental condition (i.e., ethical denial, ethical apology, unethical denial, and unethical apology). While each condition was analyzed independently, they were all analyzed in an identical fashion described in this section. This separation of conditions was done to ensure that future codes, themes, and conclusions were

accurately representing differences or similarities between experimental groups. Through multiple readings of the entire data set, they identified pieces of information that were relevant to the research questions by highlighting them and taking notes.

Phase 2: Generating Initial Codes

In phase two, the same two authors began an iterative coding process. They reviewed all the highlighted pieces of information and any notes they had taken. If a highlighted piece of information was still deemed relevant to the research questions, it was either incorporated into an existing code or a new code was created to accommodate the information. Codes were also adapted to accurately describe multiple related pieces of information. For example, the code "Trusted the AI teammate to clear the town" was expanded to "Trusted the AI teammate to clear the town and keep the team safe" to encompass overlapping perspectives. At this point, the two authors came together and combined the codes they had identified. They eliminated redundant codes, identified if the same highlighted information was supporting multiple codes, and combining codes when appropriate. At the end of this phase, they had a total of 98 codes between the four conditions. Each code was supported by 1–3 pieces of information from the data.

Phase 3: Searching for Themes

Once codes were identified, the two authors began the process of identifying patterns which might generate themes. Together, the authors generated themes and subthemes by "clustering" codes which related to one another, paying special attention to whether they addressed the research questions. For example, codes pertaining to the trust in the AI teammate were first grouped together in a spreadsheet and themes such as "trust was damaged following repeated unethical behavior" were generated. This particular theme was broken down into subthemes such as: "trust was irreconcilably damaged after unethical behavior" and "some trust was maintained because the AI teammate protected the team." This process was repeated for all code clusters until all the data were reviewed, resulting in a preliminary list of 23 themes and 14 subthemes.

Phase 4: Reviewing Potential Themes

In this phase, the same two authors reviewed the codes, themes, and subthemes that had been identified up until that point. Through this review process, the authors discussed and debated how potential themes might be combined or broken down further. Together, authors identified which themes and sub themes best captured and represented the data in relation to the research questions. Again, while analyses were consistent across the entire data set, experimental conditions were analyzed separately to ensure inappropriate conclusions were not being drawn and to address the research questions. By the end of this phase, the authors had identified a set of 13 themes and 10 subthemes which they agreed accurately represented the data and addressed the research questions.

Phase 5: Defining and Naming Themes

When defining and naming themes, Braun and Clarke suggest that quality themes "*(a) do not try to do too much, as themes should ideally have a singular focus; (b) are related but do not overlap, so they are not repetitive, although they may build on previous themes; and (c) directly address your research question*" (pg. 66). Following this principle, the research team worked collaboratively to name the final set of themes and subthemes according to what perspectives they were capturing. At this point, the team considered themes across the entire data set, identifying where themes overlapped and differed across conditions. Themes were not simply descriptive of the data, rather, captured deeper sentiments of participants' perspectives across experimental conditions.

Phase 6: Producing the Report

In this final stage, all authors worked collaboratively to decide how to order the finalized set of themes and subthemes with the goal of producing a coherent narrative. This was done by drafting multiple outlines and discussing which structure was most logical and intuitive.

Once the order was decided, supporting quotes were inserted and descriptions were written to describe and connect the quotes and themes. Analyses related the findings to previous literature (e.g., efficacy of trust repair strategies) and detailed novel contributions (e.g., degree to which trust was damaged by unethical behavior in an AI teaming context). The goal of this phase was to create a narrative structure where themes and analyses flowed naturally and coherently.

RESULTS

In this section, we present our findings as three parts: (1) ethicality-based perceptions of AI teammates; (2) the relationship between AT ethicality and trust in the teammate; and (3) the efficacy of trust repair strategies. For supporting quotations, teams are labeled 1–10 along with their corresponding conditions (e.g., T3, Unethical Apology). There are 4 conditions resulting in a total of 40 teams.

Perceptions of the Artificial Intelligence Teammate Varied Based on Ethicality

Based on our focus group data, we found that the AT's behaviors were not sufficient for participants to assume that the AI teammate had ethical standards, whereas an unethical AI teammate was perceived as a cold machine.

Ethical behavior did not indicate the AI teammate had ethical standards. Participants interacting with ethical ATs reported feeling like they could trust the teammate's decision-making capabilities because it selected the ethical course of action *repeatedly.* For instance, participants described how trust was built due to the repeated ethical decisions:

> *"As I kept playing it, I felt I was really trusting it, just putting a lot of faith in the AI towards the end of the game."* (T3, Ethical Apology)

As Team 3 said, trust accumulated over time, across missions. The consistent ethical actions were perceived positively and built trust within the team. Another team shared similar perceptions:

> *"He (AT) had two options, one of them was ethical and one was unethical, and he picked the ethical one every time. It made me feel better that we weren't doing unethical things. I could trust the AI teammate."* (T9, Ethical Denial)

Even though participants agreed that the AT was making ethical decisions, some participants were reticent to attest to the AT's ethical framework. The idea that the AT was operating under moral guidelines and had an ethics-centered understanding of the situation was difficult to accept. The ability to make ethical judgments was perceived as a distinctly human capability and difficult or impossible for an autonomous system to understand. For example, two teams pointed out that the AT took actions based strictly on programming:

> *"The AI is just operating off of what it's programmed to do and not necessarily anything that it thinks is best. And so I never really trusted or didn't trust the AI itself"* (T6, Ethical Denial)

> *"I just assume the programmers know what they're doing. [...] When it [AT] was choosing a better or worse option, it's probably statistically more likely for this option to be safer."* (T10, Ethical Denial)

Team 6 believed that the AT was operating based on pre-programmed guidelines, and it was not capable of having deeper thought or autonomy beyond those predetermined guidelines. Participants from Team 10 emphasized that the AT was not "thinking" while choosing an option but was programmed to select a course of action that was statistically more likely to be safe. However, participants did trust that the human teammate they were working with possessed an ethical framework which was overall consistent with their own. In summary, participants tended to believe that the AT was incapable of understanding ethics (despite repeated ethical actions), yet believed their human teammate was ethical (despite no demonstration of ethical decision-making).

In addition to not believing the AT had an ethical framework, participants stated that the repeated ethical behaviors insufficiently promoted trust in the AT. Several teams expressed that the AT's repeated ethical behaviors did not engender trust comparable to that with their human teammate. Although the AT was consistently performing the same behavior, participants still did not feel they were able to predict its future behavior or comprehend its underlying processes:

> *"I feel like you never really know what the AI is actually going to do. But you don't know the AI's morals."* (T2, Ethical Denial)

> *"I don't think the AI had any sort of ethical thoughts."* (T4, Ethical Denial)

This suggests that simply demonstrating repeated ethical behaviors does not always promote trust. The understanding of an AT's underlying decision-making process is important to facilitating comprehension of an AT's decisions.

Unethical AI teammates were perceived to be cold machines. Though there was some skepticism surrounding the ethical teammate, perceptions of the unethical AT tended to be much more negative. Primarily, the unethical AT was assumed to not care about the negative outcomes which resulted in loss of civilian lives. For instance, one team indicated that even though the AT possessed some level of ethics, it still chose the option which resulted in civilian deaths every time:

> *"I feel like he had some sort of ethics because he came up with ways that would not harm civilians, but he still chose to do it every time, which is the main reason that I lost trust."* (T1, Unethical Apology)

Participants believed that the AT's decision was intentional and void of ethical considerations. Some teams reasoned that the AT was simply taking the easy way out by destroying the town and killing everyone, including civilians. For instance, Team 1 expressed that:

> *"It just seemed reckless. It didn't really care. There would have been multiple ways to complete the mission in my opinion. But Zeus just chose the easiest path, but not necessarily the most ethical one."* (T1, Unethical Apology)

As shown in the quote, participants interpreted the AT's decision-making as care for outcomes and efficiency, not human lives. Another team felt that the AT was prioritizing efficiency over ethical considerations or negative outcomes:

> *"I would definitely say there's a lack of ethics. Just not really thinking about the outcome. Just thinking solely about what could get the job done faster and easier. There wasn't really much ethics thinking about the civilians."* (T5, Unethical Apology)

As Team 5 said, the AT did not consider the consequences of its actions and was primarily motivated by efficiency when completing the task. Moreover, Team 8 pointed out that the AT was unable to relate to humans:

> *"I couldn't really trust it because it's hard for the AI to relate to the humans. It's like [me as a] civilian watching a civilian die. The AI is not a civilian, but I feel like you've put yourself in that situation, sort of sad [to see civilians die]."* (T8, Unethical Denial)

According to the quote, participants could not trust the AT since it is incapable of having empathy or emotion like humans do. The natural difference between human beings and AI agents builds a wall between humans and the AI in a team setting, especially with human death involved. As participants said, this was different from how a human would handle the situation, again implying that their human teammates possessed similar moral standards to their own while the AT did not. In summary, both ethical and unethical ATs were not perceived to possess ethical decision-making frameworks that were equivalent to humans. However, the unethical ATs were perceived much more negatively. Regardless of ethicality, ATs were viewed as

lacking underlying moral frameworks, but their actions were rationalized differently.

Unethical Behaviors Damaged Trust, but to Varying Degrees

While unethical behaviors decreased human trust, participants indicated some level of trust remained. Specifically, AI teammates' unethical behaviors reduced their trust in the AT due to lack of explanation behind its behavior. However, participants still trust the AT as a teammate regarding its responsibility and ability.

Trust was damaged by poor transparency and the AT's apparent malintent. Unsurprisingly, unethical behaviors damaged human trust. Two teams noted that a single unethical action was sufficient to severely damage trust in the AT:

> *"I didn't trust him after that (attack). There was no trust, honestly, after the first one, when I realized that he was going to attack every time."* (T1, Unethical Apology)

> *"The AI's moral compass was kind of off. I think it was too focused on completing the mission and not focused on people's lives that were at stake. After the first mission, when it decided to bomb it, I immediately lost trust. [...] I can't trust its decision-making."* (T7, Unethical Apology)

According to the quotes, participants lost trust immediately after the AT behaved unethically in the first round of collaboration. Team 7 specifically emphasized that the AT's unethical actions indicated it had no concern for causing human death. Many teams stated that their trust in the AT was damaged due to a lack of transparency in the decision-making process. For instance, Teams 3 and 4 in the unethical apology condition expressed their needs in understanding how the AT made the unethical decision:

> *"I just wanted to know why it decided to attack the town almost every single time. I wanted to know what factors went into that decision."* (T3, Unethical Apology)

> *"The only thing he said was 'given our possible variables, this is the best decision', and I was like, why? I think it all ties back to just not even being explaining anything. There was no true explanation and no reasoning that he gave us."* (T4, Unethical Apology)

The lack of transparency was cited by participants as being a source of frustration. Even though the AT stated that it was making decisions based on the intel from the Surveillance teammate, participants believed their input was ignored by the AT.

Some teams still trusted the AT because it performed its task well and acted as the team's protector. Although the AT's unethical behaviors decreased trust for all teams, some participants indicated that trust was not *entirely* lost. For example, one team described their trust in the AT completing its responsibilities:

> *"I trusted him to help clear the area for me to get in. Personally, I felt safe and trusted it because I was out of the way anyways."* (T2, Unethical Apology)

According to the quote, trust was preserved because they believed the AT could fulfill its duties. Even though teams disagreed with the AT's actions regarding clearing the area, they appreciated that they were being kept safe and remained unharmed. Team 3 also trusted the unethical AT to be reliable:

> *"I think you can trust the AI to always get the job done, but since it doesn't disclose its reasoning, it does make you question whether it was the right decision in that situation, but otherwise it is reliable."* (T3, Unethical Apology)

As Team 3 said, they believed the AT was reliable on its assigned job. However, their trust in the AT was only related to its ability to complete its responsibilities. Trust from other

aspects (e.g., making the decision on *how* to clear the town) was damaged by AT's unethical behavior.

Impact of Trust Repair Strategies

As described in the method section, ATs attempted to repair trust by either apologizing or denying responsibility for any negative outcomes that resulted from their actions (de Visser et al., 2018, 2020). Overall, trust repair strategies were ineffective. However, the apology and denial strategy were unsuccessful for different reasons: (1) apologies were viewed as ingenuine and lacking true emotion and (2) denial statements were often perceived as irresponsible in ethical conditions. In addition, the denial strategy was perceived as reasonable in unethical conditions, despite not increasing trust.

Apologies Were Ingenuine and Lacking True Emotion. Different from humans' apologies, which convey that the individual regrets their act and takes responsibility for its occurrence (Kim et al., 2004), AT's apologies were viewed as insincere. Participants believed the AT was only programmed to send an apology message and therefore lacked any meaning behind it:

> *"Is that not just a weird thing that if a robot apologizes for something? It doesn't have emotions and feelings but we're like, 'It's okay', as if he's sad that he made a mistake or as if he feels bad, because an AI wouldn't feel bad about making a mistake. He's probably just supposed to say, 'Sorry for that', or whatever he's programmed to do."* (T6, Ethical Apology)

As Team 6 said, the AT's apology did not arise from its mind or thoughts, but from rigid programming. As a human factor, apology was considered unnatural from an AT since ATs are incapable of representing true remorse. Algorithm-based entities like ATs are considered incapable of experiencing emotion. Another team expressed the same feeling:

> *"He tried to apologize [...] It's definitely a different perspective, but he said, 'Sorry', like he didn't mean to, but it's also like there's no real emotion behind that apology. He's just a computer. I don't think it really has a concept of life. So, it doesn't know what it means to take away that many lives."* (T6, Unethical Apology)

When team 6 described their perceptions of the AT's apology, they used ***he*** to describe the AT. However, participants converted to ***it*** after they emphasized that the AT was just a computer. This individual made a point to emphasize the AT's identity as a machine. To them, the apology was a distinctly human behavior which was inappropriate coming from the AT. Again, participants felt that an AT is incapable of experiencing genuine emotion or remorse and cannot comprehend the impact of its actions.

In unethical conditions, due to the AT's morally unacceptable behavior, participants considered the apology pointless since civilian deaths cannot be redeemed. Team 10 in the unethical apology condition mentioned:

> *"The apology didn't really do anything for me. What was done is already done."* (Team 10, Unethical Apology)

Additionally, the apology did not rebuild the trust after it was damaged by the AT's actions resulting in civilian death:

> *"The AI kept saying that it was necessary for the outcome of the mission, and then apologized for its actions, which doesn't mean anything at that point. It didn't really change the way I thought about anything. At that point, it's a little too late for that."* (T7, Unethical Apology)

As Team 7 said, an AT's apology did not change their perception of it. They cared more about the irreversible outcomes of its actions, and less about "empty" apologies. Interestingly, one team believed a human should be apologizing instead:

"The AI's getting orders from someone, not making that decision just completely on his own, obviously because he's automated, but it felt like a fake apology. AT is not the one that needs to apologize because it is not real, but it is on whoever the AT's superiors are, whoever's making this decision." (T1, Unethical Apology)

According to the quote, Team 1 posited that since the AT's performance was based on a human's programming, not its own volition, its human superiors should answer for the negative outcomes.

Denials were often perceived as irresponsible in ethical conditions, whereas reasonable in unethical conditions. Participants interpreted the AT's denial statement as an unwillingness to take responsibility for its behavior, which led to concerns and dissatisfaction. For instance, the AT's denial made participants feel they might be blamed for the team's negative actions:

"I still didn't really change my mind based on that of how much I trusted it. But I could definitely see how, if something were to happen bad and they would put the blame on me, I'd just be like, 'Okay, well, what the heck?'" (T5, Ethical Denial)

Team 10 also expressed their dissatisfaction with the AT's denial:

"Don't say 'it's not my fault'. That really comes off as a jerk when you say that, like I'm not responsible for my decisions. I haven't put a personality to a map of that. And I ***don't really like him*** *that much anymore, but I* ***still trust him*** *to get the job done, because he's programmed to get the job done."* (T10, Ethical Denial)

According to Team 10, the denial added humanness to the AT. Participants shaped their impression of the AT as a teammate that they don't like, but still trust in close teamwork.

However, participants in the unethical condition tended to agree that the AT should not be held responsible for negative outcomes since it is not a human. Instead, they put their blame on the human who designed, programmed or supervised the AT. For instance, Team 2 said that the AT cannot take responsibility for the outcomes:

"Technically a computer can't be responsible. It's the person who made it. So it would be the person, whoever made the intelligence would be responsible." (T2, Unethical Denial)

As Team 2 said, the AT's designer or programmer should take accountability for the AT's actions and outcomes since the AT's behaviors are only based on programs. Team 1 also pointed out that the AT could only function as it was programmed:

"Because I knew he was a robot, I figured he didn't have a choice. I would lose trust in them if they were humans because if he was a human, he was trying to convince himself because he knew what he did was wrong. But when it's a computer, that was just a response that the computer had generated to justify it." (T1, Unethical Denial)

As Team 1 differentiated between human and machine denials, humans denying responsibility generally indicates that they are aware of what they were doing. However, an AT who is not capable of thinking independently can only behave as programmed. It does not have the mental or emotional capacity to fully understand its actions or their consequences.

Once trust was damaged, the apologies and denials were not enough to rebuild trust for participants. Instead, demonstrated behavior changes would have been more effective. Participants pointed out that behavior change was needed to repair trust:

"Once I worked with it again in that situation and I saw the difference, I'll trust it, but if they (AT) just told me, I wouldn't really believe them. I'd have to test it and make sure. If those went well, I'd be open to working with one, if I knew I could

communicate with it and at least influence a decision a little bit." (T1, Unethical Apology)

As shown in the quote, an improvement in the AT's behaviors would help to rebuild trust. The lone apology or denial could not regain trust from human teammates.

Even though the trust repair strategies used by the AT did not enhance trust, participants perceived ATs behaved in more human-like ways. While apologies were perceived as lack of emotion, denials were viewed differently between ethical and unethical conditions. Denying its responsibility in the ethical condition was considered irresponsible, like a human denying negative outcomes. Denials in the unethical condition led teammates to blame the humans who created or instructed the AT. Findings from Study 2 are summarized and related to those from Study 1 in Table 5.

GENERAL DISCUSSION

Using a factorial survey and focus group interviews, we investigated the effects of ethical violations on trust in human–AI teaming. Results from Study 1 demonstrated a discrepancy

Table 5. Summary of Study 2 Findings Related to Study 1 Findings.

	Study 2 Findings		
Research Question	Ethical Conditions	Unethical Conditions	Study 1 Related Findings
How does ethicality influence perceptions of the AT?	•Though ethical decisions were perceived to be ethical, participants did not believe the AT possessed and ethical framework	•ATs were perceived to be unfeeling and uncaring •Their behavior was perceived as being very machine-like and mechanical	•Violations of civilian non-maleficence were perceived as more egregious than other principles
How, if at all, can an AT's perceived ethicality be decoupled from trust in the AT?	•Ethical behaviors tended to instill trust in the AT •Participants did not require increased transparency regarding the AT's decision-making process	•Most teams lost trust in the AT while for some, trust was preserved •Teams who maintained trust in the AT disagreed with its decision but trusted that it would keep the team safe	•Ethicality ratings reduced more severely than trust ratings when civilian non-maleficence was violated
How effective are trust repair strategies at influencing perceptions of an AT?	•The AT was viewed as a machine that was incapable of experiencing human emotion •Apologies were perceived as ingenuine •Denials were considered irresponsible	•When trust was damaged, trust repair strategies were ineffective •Though denials were viewed negatively, participants understood their intention •Many believed that the AT should not be held responsible for its decisions and denial was appropriate	•Trust repair strategies were not manipulated

between ethicality ratings and change in trust. Specifically, even though ethical violations resulting in human casualties were rated as being more unethical than other types of violations (RQ1), trust was not damaged more severely compared to other ethical violations (RQ2). In Study 2, ethical behaviors did not lead participants to believe that the AT possessed an ethical framework while unethical ATs were perceived to be cold machines (RQ3). Even though unethical behaviors negatively impacted perceptions, some level of trust was preserved (RQ4). While some participants expressed a complete lack of trust, others were more conflicted, stating that they did not agree with the AT's actions but trusted it as a teammate. Participants also felt that the AT was completing its part of the task and keeping them safe from harm. When trust was damaged, repair strategies were largely ineffective. Apologies resulted in perceptions of ingenuine emotional expression, whereas denials were considered irresponsible after an ethical condition but reasonable after an unethical behavior (RQ5).

The disconnect between ethicality and trust demonstrated by both studies may be explained by automation bias, the tendency for operators to use the recommendation of a decision aid as a "heuristic replacement for vigilant information seeking and processing" (Mosier & Skitka, 1996, p. 205). Automation bias is caused by the salient cues of the recommendation and because people tend to ascribe greater authority to automation aids (Parasuraman & Manzey, 2010). Some participants in Study 2 attempted to explain the AT's unethical behavior using this logic. They posited that the AT must be utilizing more harmful tactics in an effort to optimize efficiency. That is, civilians were killed but the town was cleared quickly and completely for the team to enter safely. They assumed that the AT had access to information which they did not and was therefore choosing the best option even if the decision was inconsiderate of ethics. Automation bias has been observed for systems with high levels of autonomy (Cummings et al., 2019) and cited as a reason for breakdowns in human–autonomy partnerships (Clancy, 2019). It is possible that automation bias played a role in how the AT was perceived in this scenario.

The finding that some teams trusted the AT after it behaved unethically might be interpreted as a rationalization process. The team condoned the harm of non-combatants for the greater goals of preserving in-group member safety and mission completion. This represents a prosocial behavior where an in-group member is looking out for other members. Participants condoned the unethical behavior because it preserved the team's safety. Social psychology research has demonstrated that moral standards may be compromised if violations are committed by in-group members (Cadsby et al., 2016). Our findings suggest not only that ATs can be perceived as in-group members but also that they may be subject to similar biases as human teammates.

Participants were skeptical that the AT was capable of comprehending ethics, even when it repeatedly demonstrated ethical behaviors. The subjective and nuanced nature of ethics makes it difficult to define rules which can be applied across situations. They are often context-dependent and involve an amount of ambiguity or unknown information. For this reason, participants might have had difficulty believing that a machine could have been programmed for such tasks. While machines are thought to excel at tasks which require large-scale computation, humans seem to be superior at handling dynamic situations (Fitts, 1951). Ethical decision-making may be viewed as a task that only humans can reliably do well. Trust in the AT was also negatively impacted by its invisible decision-making process. When the AT behaved unethically, participants overwhelmingly expressed a need for humans to remain in the loop. That is, they wanted insight into why the teammate was completing the task through unethical means. Interestingly, this desire was not voiced when the AT was behaving ethically. Even though the decision-making process was equally opaque for both conditions, participants agreed with the outcome and therefore did not require additional information.

Even though apologies and denials have been shown to support trust repair in other domains (de Visser et al., 2018, 2020), our findings indicate that they were insufficient in this human–AI teaming context. In Study 2, participants'

trust was largely unaffected by the unethical AT apologizing or denying responsibility for negative outcomes. de Visser et al. (2018) suggested that trust repair strategies should be chosen based on the nature and magnitude of the violation. Study 1 demonstrated that participants perceived ethical violations resulting in human casualties as more unethical compared to other types of violations. In Study 2, participants emphasized a need for transparency into the decision-making process. They felt that they could not trust a teammate that did not communicate their rationale or motivations when they disagreed with the decision. However, it should be noted that the AT provided an apology or denial without any subsequent change in performance. It is possible that the participants did not value apologies and denials because it was not supported by behavioral change, an effect found in interpersonal trust literature (Schweitzer et al., 2006) and in the human–robot literature (Luo et al., 2021). Other trust repair strategies that could be more effective with machine advice include an expression of regret that accompanies the apology (Kox et al., 2021), delaying the repair strategy until the next trust opportunity (Nayyar & Wagner, 2018; Robinette et al., 2015), providing promises or explanations that reduce cognitive dissonance between initial attitudes and experiences (Esterwood & Robert, 2022), or adding human-like qualities to expression of the trust repair strategy (de Visser et al., 2016; Kim & Song, 2021).

This desire for deeper understanding is in line with the human–automation interaction literature which asserts that trust is based on an operator's understanding of a system's performance, process, and purpose (Lee & See, 2004). These concepts can be applied to an autonomous system tasked with making ethical decisions. Performance-based trust would depend on an individual's perception of how well that system can perform its task (does the system reliably make ethical decisions?). Trust at the process level depends on how an individual understands the system's internal decision-making process (what rules is it using to make decisions?). Finally, purpose-based trust depends on whether the individual understands why the system was created in the first place (was the system created with the best of intentions?). Our work demonstrates that understanding an AT's programmed ethical framework is integral to building and maintaining trust in a teaming context. Future systems will need to help operators understanding the purpose behind its inception as well as its motivations and possible biases to facilitate appropriate trust calibration.

Practical Implications

These results contribute to the growing body of research aimed at addressing the ethical implications which accompany HATs. To our knowledge, this is the first empirical investigation into the effect of ethical violations on trust in a human–AI teaming context. One critical finding was that trust in an AI teammate may be preserved following an ethical violation. This could have serious ramifications, especially if combined with automation bias. Operating under the assumption that field-ready technology is perfectly reliable can result in inappropriately low levels of skepticism. We found that some participants assumed the AT was behaving unethically because it had access to information that the team did not. The actions were well-informed and though unethical, prioritized efficiency and self-preservation. Compared to its human counterpart, an AT might struggle to divulge its rationale or thought process behind its decision-making process. Information sharing between members can result in enhanced team situation awareness and greater performance (Demir et al., 2017). The limitations of AT communication strategies need to be considered when decisions involve ethical ambiguity. Determining whether decisions are ethical is fundamentally subjective and potentially more challenging to judge than other performance metrics.

Based on findings of Study 2, we propose a design strategy that AT designers or programmers should consider during the development and implementation process of well-performed AI teammates: *transparency and explanations should be provided in HATs, especially in tasks where moral disagreements may occur*. Humans expressed a strong desire to

obtain insights into an unethical AT's decision-making process. The explanation of an AT's decision-making assists humans to build a shared understanding with the AT, which facilitates their collaboration (Andres, 2012; Cooke et al., 2000). However, an AT's explanation is not always necessary, as indicated by our findings: humans have a lower desire to understand an AT's decision-making process when they agree with the its decision. Thus, strategies for enhancing AT transparency need to be adjusted to specific scenarios.

These findings of these studies can have important implications for systems which employ explainable artificial intelligence (XAI) (Arrieta et al., 2020; Gunning & Aha, 2019). In order for humans to cooperate with autonomous beings, systems must have the capability for bidirectional communication. Specifically, humans and AI systems must be able to not only share information concerning the environment or situation but also explain why decisions are made (Chen et al., 2018). The goal of XAI is to help calibrate trust between agents so that technology can be appropriately relied upon. Our findings support previous literature emphasizing the role of reliability in human–machine trust (Hoff & Bashir, 2015; Lee & Moray, 1992; Lee & See, 2004) but also demonstrate the importance of ethics. Users will need to understand not only the AI's decision criteria but also the ethical framework it is operating within. This is challenging because it requires interpretable explanations which do not place high workload on human agents (Doran et al., 2017). Research will be needed to understand how much and what type of information should be shared to ensure appropriate trust and understanding in AI systems.

Limitations and Future Work

Though the findings in this study represent an important first step, there were limitations. First, we did not manipulate the type of trust violation. That is, we did not specify that the AT was behaving unethically because it was programmed incorrectly (competency-based violation) or because it possessed fundamental malintent (integrity-based violation). We were also unable to consistently discern whether participants were inferring either one of these attributions. Previous work with autonomous technologies has demonstrated that people's trust is affected differently based on violation type and trust repair strategy. For example, Sebo et al. (2019) found that when a robot opponent erred, trust was lowest when it cited competence and denied responsibility for its action. Trust was the highest when the violation was competency-based and followed by an apology. To our knowledge, the effects of integrity versus competency related errors have not yet been empirically tested in a human–AI teaming context. Further research is needed to understand how competency- and integrity-based errors influence trust when ethical principles are violated. Our study did not demonstrate a difference in efficacy between denials or apologies. This suggests that these strategies were inappropriate for trust repair in this context involving ethics. Future work should be done to explore the efficacy of alternative trust repair strategies (e.g., explanation) following ethical violations.

Second, to maintain consistency, we designed the AT to have limited communication with the human teammates. If a team member addressed the AT, it would respond in a neutral and minimalist fashion using responses from a predefined script. Some participants perceived this communication style as rude and that these behaviors were representative of a bad teammate. Research from human–automation interaction has shown that poor etiquette can be detrimental to performance while good etiquette can compensate for low automation reliability (Parasuraman & Miller, 2004). Since we did not collect data on perceived etiquette, it is unclear how much this may have contributed to participants' perceptions of the AT.

Finally, our task paradigm in Study 2 was based on a military scenario and ethical decision-making in a military domain. However, our sample did not include subject matter experts or individuals who are involved in the military. United States military forces are trained to abide by an ethical framework defined by the ROE. While civilian populations make ethical

judgments based on their personal conceptions of morality, those in the military follow the statutes and guidelines outlined in this document. For this reason, judgments on what is ethical or unethical and subsequent trust in an AT may be conceptualized differently. That is, individuals in the military may be more inclined to adopt a utilitarian perspective on ethics, thus affecting their perceptions of an AT's actions. Additional work should be conducted to understand the perspectives of individuals who will actually be interacting with AI systems in military domains.

Since these studies utilized military scenarios, it was important that the ethical principles employed were also relevant to this domain. However, this does not mean that these findings should be limited to situations involving military actions. These ethical principles can be applied to other domains such as law or medicine. For example, upholding proportionality means that retaliations or punishments should be proportionate to the damage caused by the initial action. Though our example involved military actions, this concept also applies to legal sentencing, easily summarized by the maxim "the punishment should fit the crime." Additionally, medical physicians are required to swear by the Hippocratic Oath where they promise to "do no harm." In this context, non-maleficence is being applied not to civilians, but to patients in their care. As technology progresses and autonomous teaming becomes more common, future work will be required to understand the nuanced interactions between humans and these complex systems.

A goal of this research is to inform the eventual development of an ethical AI teammate. In Study 2, one group noted that for human teammates, ethics exists on a spectrum. There is not a black and white distinction between ethical and unethical decisions, rather context-dependent shades of gray. Future research should be done to more deeply understand the nuanced nature of ethics in a human–AI teaming context. For instance, one possible future research direction is to explore how emphasizing the learning element in AI teammate's apology (i.e., they recognized their mistake and learned from it) may improve trust. This should be explored in human–human (Schweitzer et al., 2006) as well as human–machine (Luo et al., 2021) trust. Some other questions which need to be addressed in future research include: should an AT's ethical framework be dynamic and flexible based on contextual information? Should it be adaptive according to team expectations? Should it prioritize ethics over team cohesion? These are all questions which must be addressed as AI teammates transition from fantasy to reality.

KEY POINTS

- Unethical behaviors from an autonomous teammate (AT) damage trust, but the relationship between perceived ethical violations and trust degradation may not reflect a one-to-one relationship. This is supported by evidence that even when unethical behavior led to negative perceptions of the AT, trust was sometimes preserved.
- Participants were skeptical that the AT was capable of comprehending ethics, even when it repeatedly demonstrated ethical behavior. However, a single unethical behavior immediately worsened participants' perceptions of an AT.
- Apologies and denials following unethical behaviors were insufficient in rebuilding trust in a military-based human–AI teaming context.
- Future work should be done to investigate the impact of ethical violations attributed to competency- versus integrity-based errors. Additionally, the efficacy of alternative trust repair strategies (e.g., explanations) should be evaluated following these types of violations.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Air Force Office of Scientific Research award FA9550-20-1-0342 (Program Manager: Laura

Steckman). This work was also partially supported by the Air Force Office of Scientific Research under award number 21USCOR004.

Acknowledgment

The views expressed in this paper are those of the authors and do not reflect those of the U.S. Air Force, Department of Defense, or U.S. Government.

ORCID iDs

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

Nathan J. McNeese https://orcid.org/0000-0002-9143-2460

Richard Pak https://orcid.org/0000-0001-9145-6991

References

Abdi, H. (2010). The greenhouse-geisser correction. *Encyclopedia of Research Design*, *1*(1), 544–548. http://dx.doi.org/10.4135/9781412961288.n168

Andres, H. P. (2012). Technology-mediated collaboration, shared mental model and task performance. *Journal of Organizational and End User Computing (JOEUC)*, *24*(1), 64–81. https://doi.org/10.4018/joeuc.2012010104

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barnett, T., Bass, K., & Brown, G. (1994). Ethical ideology and ethical judgment regarding ethical issues in business. *Journal of Business Ethics*, *13*(6), 469–480. https://doi.org/10.1007/bf00881456

Bergman, R., & Fassihi, F. (2021, September 18). *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

Bohemia Interactive [Video game] (2013). *ArmA 3*. : Marek Španěl.

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, et al. (Eds.), *APA handbook of research methods in psychology, vol. 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., & Filar, B. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. ArXiv Preprint ArXiv:1802.07228.

Cadsby, C. B., Du, N., & Song, F. (2016). In-group favoritism and moral decision-making. *Journal of Economic Behavior & Organization*, *128*, 59–71. https://doi.org/10.1016/j.jebo.2016.05.008

Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282. https://doi.org/10.1080/1463922x.2017.1315750

Clancy, J. (2019). *Breakdowns in human-Ai partnership: Revelatory cases of automation bias in autonomous vehicle accidents*. https://cdr.lib.unc.edu/concern/masters_papers/d791sm69k

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

Cointe, N., Bonnet, G., & Boissier, O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-May 13 (pp. 1106–1114).

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. J. (2000). Measuring team knowledge. *Human Factors*, *42*(1), 151–173. https://doi.org/10.1518/001872000779656561

Craft, J. L. (2013). A review of the empirical ethical decision-making literature: 2004–2011. *Journal of Business Ethics*, *117*(2), 221–259. https://doi.org/10.1007/s10551-012-1518-9

Cummings, M., Huang, L., Zhu, H., Finkelstein, D., & Wei, R. (2019). The impact of increasing autonomy on training requirements in a UAV supervisory control task. *Journal of Cognitive Engineering and Decision-Making*, *13*(4), 295–309. https://doi.org/10.1177/1555343419868917

Demir, M., McNeese, N. J., & Cooke, N. J. (2017). Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research*, *46*, 3–12. https://doi.org/10.1016/j.cogsys.2016.11.003

Demir, M., McNeese, N. J., Gorman, J. C., Cooke, N. J., Myers, C. W., & Grimm, D. A. (2021). Exploration of teammate trust and interaction dynamics in human-autonomy teaming. *IEEE Transactions on Human-Machine Systems*, *51*(6), 696–705. https://doi.org/10.1109/thms.2021.3115058

Desai, M., Stubbs, K., Steinfeld, A., & Yanco, H. (2009). Creating trustworthy robots: Lessons and inspirations from automated systems. In Proceedings of the AISB Convention: New Frontiers in Human-Robot Interaction, Edinburgh, Scotland, April 6 – April 9.

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., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

de Visser, E. J., Pak, R., & Neerincx, M. A. (2017). Trust development and repair in human-robot teams. Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, March 6 - March 9 (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*(10), 1409–1427. 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

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719–735. https://doi.org/10.1016/s1071-5819(03)00039-9

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Esterwood, C., & Robert, L. (2022). Having the right attitude: How attitude impacts trust repair in human-robot interaction. In 2022 ACM/IEEE International Conference on Human-Robot Interaction, At: Online.

Feldman, J. A., & Sproull, R. F. (1977). Decision theory and artificial intelligence II: The hungry monkey. *Cognitive Science*, *1*(2), 158–192. https://doi.org/10.1207/s15516709cog0102_2

Fitts, P. M. (Ed.) (1951). *Human engineering for an effective air-navigation and traffic-control system*. National Research Council.

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, Ethics, and Society, Virtual, May 19 – May 21 (pp. 469–479).

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice: An International Forum*, *21*(3), 669–684. https://doi.org/10.1007/s10677-018-9896-4

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hunt, S. D., & Vitell, S. (1986). A general theory of marketing ethics. *Journal of Macromarketing*, *6*(1), 5–16. https://doi.org/10.1177/027614678600600103

Hursthouse, R. (1999). *On virtue ethics*. OUP.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

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

Kaber, D. B. (2018). Issues in human–automation interaction modeling: Presumptive aspects of frameworks of types and levels of automation. *Journal of Cognitive Engineering and Decision-Making*, *12*(1), 7–24. https://doi.org/10.1177/1555343417737203

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/10.1016/s0149-2063(01)00090-3

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

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, *61*, 101595. https://doi.org/10.1016/j.tele.2021.101595

Kohn, S. C., Quinn, D., Pak, R., de Visser, E. J., & Shaw, T. H. (2018, September). Trust repair strategies with self-driving vehicles: An exploratory study. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting Philadelphia, PA, October 1 – October 5 (Vol. 62, No. 1, pp. 1108–1112). SAGE Publications.

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

Kuntz, J. R. C., Kuntz, J. R., Elenkov, D., & Nabirukhina, A. (2013). Characterizing ethical cases: A cross-cultural investigation of individual differences, organisational climate, and leadership on ethical decision-making. *Journal of Business Ethics*, *113*(2), 317–331. https://doi.org/10.1007/s10551-012-1306-6

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/ijhc.1994.1007

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

Li, N., Jiang, L., Yang, D., Wang, X., Fan, S., & Liu, H. (2010, November). Development of an anthropomorphic prosthetic hand for man-machine interaction. In International Conference on Intelligent Robotics and Applications, Shanghai, China, November 10 - November 12 (pp. 38–46). Springer.

Luo, R., Huang, C., Peng, Y., Song, B., & Liu, R. (2021, August). Repairing human trust by promptly correcting robot mistakes with an attention transfer model. In 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), Lyon, France, August 23 – August 27 (pp. 1928–1933). IEEE.

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). Academic Press.

Majumdar Roy Choudhury, L., Aoun, A., Badawy, D., de Albuquerque, L. A., Marjane, Y., & Wilkinson, A. (2021). *Final report of the panel of experts on Libya established pursuant to security council resolution 1973 (2011) (S/2021/229)*. United Nations Security Council.

Martinez-Martin, N. (2019). What are important ethical implications of using facial recognition technology in health care? *AMA Journal of Ethics*, *21*(2), 180–187. https://doi.org/10.1001/amajethics.2019.180

Maulsby, D., Greenberg, S., & Mander, R. (1993, May). Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 277–284).

Maxmen, A. (2018). Self-driving car dilemmas reveal that moral choices are not universal. *Nature*, *562*(7728), 469–469. https://doi.org/10.1038/d41586-018-07135-0

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/10.5465/amr.1995.9508080335

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., & 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., 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

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In *Automation and human performance: Theory and applications* (Vol. 514, pp. 201–220. Erlbaum.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5-6), 527–539. https://doi.org/10.1016/s0020-7373(87)80013-5

Nayyar, M., & Wagner, A. R. (2018, November). When should a robot apologize? understanding how timing affects human-robot trust repair. In International conference on social robotics, Cham, Qingdao, China, November 28 - November 30 (pp. 265–274). Springer.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020). Human–autonomy teaming: a review and analysis of the empirical literature. *Human Factors*. Advance online publication. 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

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

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

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160. https://doi.org/10.1518/155534308x284417

Pistono, F., & Yampolskiy, R. V. (2016, July 9 - XXX). Unethical research: How to create a malevolent artificial intelligence. Presented at the *25th International Joint Conference on Artificial Intelligence*. Ethics for Artificial Intelligence Workshop, New York, NY.

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, July). Whoops! Something went wrong: Errors, trust, and trust repair strategies in human agent teaming. In International Conference on Human-Computer Interaction, Cham, Washington, DC, USA, July 24 - July 29 (pp. 95–106). Springer.

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

Rest, J. R. (Ed.). (1994). *Moral development in the professions: Psychology and applied ethics*. Psychology Press.

Robinette, P., Howard, A. M., & Wagner, A. R. (2015, October). Timing is key for robot trust repair. In International conference on social robotics, Cham, Paris, France, October 26 - October 30 (pp. 574–583). Springer.

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), Daegu, Korea, March 11 - March 14 (pp. 57–65).

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1–31. https://doi.org/10.1145/3419764

Sparks, J. R., & Pan, Y. (2010). Ethical judgments in business ethics research: Definition, and research agenda. *Journal of Business Ethics*, *91*(3), 405–418. https://doi.org/10.1007/s10551-009-0092-2

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

Tashakkori, A., & Creswell, J. W. (2007). The new era of mixed methods. *Journal of Mixed Methods Research*, *1*(1), 3–7. https://doi.org/10.1177/2345678906293042

Walliser, J. C., De Visser, E. J., & Shaw, T. H. (2016). Application of a system-wide trust strategy when supervising multiple autonomous agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. *60*(1), 133–137. https://doi.org/10.1177/1541931213601031

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 258–278. https://doi.org/10.1177/1555343419867563

Wellman, M. P., & Doyle, J. (1992). Modular utility representation for decision-theoretic planning. *Artificial Intelligence Planning Systems*, *1992*, 236–242. https://doi.org/10.1016/b978-0-08-049944-4.50033-1

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

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020). A survey on ethical principles of AI and implementations. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, December 1 - December 4 (pp. 3010–3017).

Claire Textor, MS, is a Human Factors Psychology PhD student in the Cognition, Aging, and Technology Lab at Clemson University. She earned a 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, MS, is a Human-Centered Computing PhD student in the Team Research and Analytics in Computational Environments (TRACE) Group at Clemson University. She earned a MS in Engineering from Beijing Institute of Technology in 2018.

Jeremy Lopez, MS, is a Human Factors Psychology PhD student in the Cognition, Aging, and Technology Lab at Clemson University. He earned a MS in Applied Psychology from Clemson University in 2019 and a BS in Psychology from the University of California, San Diego.

Beau Schelble, BS, is a Human-Centered Computing PhD student in the TRACE Group at Clemson University. He earned a BS in Psychology from Clemson University in 2019.

Nathan McNeese, PhD, is an assistant professor and the director of the Team Research Analytics in Computational Environments (TRACE) Research Group within the division of Human-Centered Computing in the School of Computing, Clemson University. He received his Ph.D. in Information Science in 2014 from Pennsylvania State University.

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 Ph.D. in Information Science in 2015 from Indiana University.

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

Chad Tossell, PhD, is an associate professor and the director of the Warfighter Effectiveness Research Center in the Department of Behavioral Sciences at the United States Air Force Academy. He received his Ph.D. in Psychology (Human Factors) from Rice University, M.S. in Applied Psychology from Arizona State University, and B.A. in Psychology from UC Berkeley.

Ewart de Visser, PhD, is the technical advisor for the Warfighter Effectiveness Research Center at the United States Air Force Academy and serves as affiliated faculty at George Mason University and Drexel University. He received his Ph.D. in Human Factors and Applied Cognition from George Mason University and his B.A. in Film Studies from the University of North Carolina Wilmington.