The Effect of AI Teammate Ethicality on Trust Outcomes and Individual Performance in Human-AI Teams

Beau G. Schelble[1], Caitlin Lancaster[1], Wen Duan[1], Rohit Mallick[1], Nathan J. McNeese[1], and Jeremy Lopez[2]
[1]Human-Centered Computing, Clemson University
[2]Human Factors Psychology, Clemson University
(bschelb, cma8, wend, rmallic, mcneese, jalopez)@g.clemson.edu

Abstract

This study improves the understanding of trust in human-AI teams by investigating the relationship of AI teammate ethicality on individual outcomes of trust (i.e., monitoring, confidence, fear) in AI teammates and human teammates over time. Specifically, a synthetic task environment was built to support a three-person team with two human teammates and one AI teammate (simulated by a confederate). The AI teammate performed either an ethical or unethical action in three missions, and measures of trust in the human and AI teammate were taken after each mission. Results from the study revealed that unethical actions by the AT had a significant effect on nearly all of the outcomes of trust measured and that levels of trust were dynamic over time for both the AI and human teammate, with the AI teammate recovering trust in Mission 1 levels by Mission 3. AI ethicality was mostly unrelated to participants' trust in their fellow human teammates but did decrease perceptions of fear, paranoia, and skepticism in them, and trust in the human and AI teammate was not significantly related to individual performance outcomes, which both diverge from previous trust research in human-AI teams utilizing competency-based trust violations.

Keywords: human-AI teaming, trust, ethical AI, artificial intelligence, collaborative technologies

1. Introduction

Technologies like artificial intelligence (AI) and machine learning that enable technology to participate in intelligent decision-making have continued to advance and enable the introduction of intelligent machine collaborators for humans (O'Neill et al., 2020; Schelble et al., 2020). These intelligent artificial partners can bring a host of benefits to a variety of different tasks, including sense-making in data and enhancing efficiency (Mabkhot et al., 2018). Now, as the development of AI has continued, its potential applications have also grown, even to the point of occupying extremely social positions, like that of a teammate (McNeese et al., 2018).

These human-AI teams or human-autonomy teams are defined by unique characteristics that include a significant degree of agency for the artificial teammate to make decisions of its own volition and a unique role on the team amongst at least one other human team member (O'Neill et al., 2020). Such human-AI teams can be deployed in several domains, for example, manufacturing (Schelble et al., 2020) and medicine (Wang et al., 2016). In many cases, the human-AI partnership can result in outcomes greater than either entity produce individually (Wang et al., 2016). Despite the impressive capabilities of AI, its ability to interact with humans socially has lagged behind in several key areas, natural language processing being a prime example (Chowdhary, 2020), which significantly complicates the development of teaming constructs for human-AI teams.

Placing AI within increasingly social situations such as human-AI teams represents a challenge as there is currently a very limited understanding of how AI affects the development of various social factors (i.e., trust, cohesion, confidence), especially within human-AI teams. Human-AI teams have only recently begun to be studied (McNeese et al., 2018) but the impact of an AI teammate on the development and sustainment of common teaming constructs like trust, communication, and team cognition is clear (Schelble, Flathmann, McNeese, Freeman, et al., 2022). Trust is particularly important for human-AI teams as it is known to be significantly related to effective teaming outcomes (Mach et al., 2010) and recent research strongly indicates humans increasing desire to have trustworthy AI teammates (Zhang et al., 2021). Furthermore, if trust in an AI teammate is reduced, it can also negatively impact teammates' trust in their fellow human teammates regardless of their performance

URI: https://hdl.handle.net/10125/102668
978-0-9981331-6-4

(McNeese, Demir, Chiou, et al., 2021), though trust appears more fragile for artificial teammates (Jessup et al., 2020). While many factors are at play when making judgments of trust in technology, including reliability and past experience (Schaefer et al., 2016), the ethical nature of that technology also factors into the development and sustainment of trust (Winfield, 2019).

While there is a great deal of research investigating aspects of trust in human-AI teams and human-AI interaction, the influence of AI teammate ethics on trust in human-AI teams, despite the significant role of ethics-based trust in overall trust development (T. M. Jones & Bowie, 1998). Specifically, ethics and ethical behavior play a key role in team formation, cohesion, and eventual performance (Doris, 1998). Unethical behaviors, on the other hand, are often seen as working against a specific group's values, and, in terms of the trust, development can be seen in line with unreliability (Brien, 1998). Understanding how such behaviors by AI affect trust in human-AI teams is important to developing better ethical AI as human trust levels can be leveraged as another potential indicator of unethical AI actors, especially as trust in AI is highly sensitive to initial interactions (Omrani et al., 2022). Additionally, it is important to go beyond not only linking AI ethics to trust but also to identify the individual outcomes of trust that may be influenced by ethical and unethical AI actions. Rather than simply investigating whether or not AI ethicality influences human teammates' trust holistically, understanding how outcomes of trust such as fear, confidence, and monitoring individually respond to ethical and unethical AI over time can contribute towards the development of better trust repair strategies for AI that may have lost trust after an unethical action.

To address this gap in the literature and help improve the understanding of trust development and sustainment in human-AI interaction and teaming, the current paper poses the following research questions in Table 1. Each research question is also accompanied by its relevant hypotheses concerning the autonomous teammate (AT) and human-teammate (HT).

The hypotheses associated with RQ1 are based on previous research identifying ethical behavior as a consideration in evaluations of trust in human teams (T. M. Jones & Bowie, 1998). Additionally, H1.1 and H2.2 indicate that outcomes of trust in the HT will be positive but negative for the AT in the unethical AT condition. This assertion is based on the previous literature that indicates unethical behaviors are seen as working against the values of a certain group (Brien, 1998), thus humans will see their HT as part of their group and perceive more positive trust outcomes with them than the AT. H3.1 is based on previous literature indicating that ethical behavior is key to team cohesion and eventual trust (Doris, 1998), along with trust itself being highly related to outcomes of team performance (Mach et al., 2010).

Research Questions & Hypotheses	
RQ1	How does AT ethicality effect individual outcomes of trust for a HT versus the AT?
H1.1	Outcomes of trust will be higher for the HT than the the AT when the AT is unethical.
H1.2	Outcomes of trust will remain similar for both teammates when the AT is ethical.
RQ2	How do these individual outcomes of trust in human versus AI teammates change over time as a result of AT ethicality?
H2.1	Outcomes of trust will increase for both teammates working with an ethical AT as missions go on.
H2.2	Outcomes of trust will decrease for the AT but not the HT for teams working with an unethical AT as missions go on.
RQ3	If affected by AT ethicality, do these individual outcomes of trust have a relationship with individual performance over time?
H3.1	Outcomes of trust will have a positive relationship with individual performance over time.

Table 1: Research questions and their respective hypotheses.

2. Background

2.1. Human-AI Teaming

As modern teams adapt to the potential for and burgeoning reality of AI collaboration, the dynamics of human-AI teaming becomes much more important to understand. However, research has shown that humans hold high expectations for ATs powered by AI, believing they should demonstrate human-like behaviors and capabilities (Flathmann et al., 2021; Zhang et al., 2021). This demand requires ATs to possess the ability to communicate and coordinate with teammate with shared understanding, contribute to shared goals and outcomes, and demonstrate interpersonal and situational awareness (Hauptman et al., 2022; Schelble, Flathmann, McNeese, O'Neill, et al., 2022; Zhang et al., 2021), each of which contribute to the development and sustainment of trust.

While ATs are still undergoing rapid technological advancement to achieve full autonomy, the increasing democratization of these agents supports the need to design more complex and socially-aware agents to perform team-based roles more effectively and satisfactorily (McNeese et al., 2018). Indeed, unique to human-AI teams are apparent needs for transparency, reliability, and demonstrated autonomy (McNeese, Schelble, et al., 2021). While these particular needs are

often unique to human-AI teams, the demand for these, notably greater transparency, connect to the wealth of research on the requirements for trust among teammates (Mercado et al., 2016). These points reiterate the increasingly complex social environments that AI is being implemented in, making human concepts like trust and ethics all the more relevant.

2.2. Ethics in Human-AI Teaming

Humans lean upon ethics to maintain order and consistency through a shared understanding of society's acceptable social, psychological, and political actions. Ethics is often considered a set of principles to guide and examine moral life based on societal norms (Beauchamp & Childress, 2001). In the context of teaming, ethics and ethical behaviors are essential to the formation, cohesion, and overall performance of a team (Doris, 1998), and ethics in the field of teamwork extends both to the ethical responsibilities of the team to both the internal members and those externally affected by the team's actions (Kossaify et al., 2017).

In human-AI teams, ethics and ethical decision-making are paramount to team dynamics and outcomes. These teams need ATs that can operate within the necessary ethical dimensions for the contexts in which they operate (Flathmann et al., 2021). This is because ethics directly relates to trust, as humans must be able to trust their ATs to perform and use their autonomy to properly navigate ethical situations or the ethical implications of their actions (Winfield, 2019). As AI develops as an important factor in various human systems, AI ethics must be adopted and understood, such as issues of justice and fairness resulting from AI systems (Kazim & Soares Koshiyama, 2021). While AI ethics is a growing field of inquiry, there is little examination of how ethics and trust are related in the literature when it comes to human-AI teams, mainly focusing on the ethicality of development and deployment (Jobin et al., 2019) or are theoretical in nature (Flathmann et al., 2021).

In fact, AI has already been used in a real-world military operation where lethal force was used near civilians, introducing a new layer to trust and trust violations within human-AI teams (Bergman & Fassihi, 2021). In line with this, more significant research is needed to explore how ethics-based trust in ATs influences trust dynamics in human-AI teams (G. R. Jones & George, 1998). Recent work on establishing ethical frameworks for human-AI teaming echos this call, conveying the need to understand further the minute details of how ethics and trust violations operate in human-AI teams (Flathmann et al., 2021).

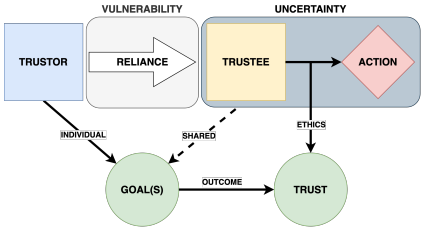

Figure 1: Conceptual model of trust between a trustor and trustee and how the outcome and and ethics of the trustee influence trust.

2.3. Trust in Human-AI Teaming

In general, trust as a psychological construct is imperative to establishing expectations among teammates, helping them overcome issues with uncertainty and risk that may occur during their teaming situation, and collaborating effectively despite these unknowns (G. R. Jones & George, 1998). Trust in the current paper is defined as the willingness of an individual to be vulnerable to another regardless of their ability to monitor and or control the other (Mayer et al., 1995) (see Figure 1). Essentially, the trustor is making themselves vulnerable to the trustee by relying on the trustee to perform a certain action at a particular time and or place that is important to the trustor's individual goals or their shared goals (Mayer et al., 1995). This model also applies when the trustee is an artificial agent where the agent is helping an individual achieve their goal in conditions of uncertainty and vulnerability (Lee & See, 2004). Collaboration is deeply intertwined with trust in teamwork situations, and high levels of trust are essential to building team cohesion (Mach et al., 2010) and high-performance (Mach et al., 2010).

Studies on trust in human-AI teaming tend to focus on perceptions of trust as they relate to reliance on autonomous technologies in teaming scenarios (Lee & See, 2004) that then extend to outcomes such as team performance and trust calibration (McNeese, Demir, Chiou, et al., 2021; Schaefer et al., 2016). Humans' propensity to trust ATs stem from many elements, including individual traits like extraversion and situational traits like uncertainty (Elson et al., 2018).

While the literature on trust in human-AI teaming is growing, there is little research on how the ethicality of AT actions affects trust in human-AI teams in comparison to the heavily examined relationship between teammate ethics and trust in human-human teams (Doris, 1998; O'Neill et al., 2020). Unethical individuals are often seen as working against the norms and values of a given group, causing humans to have reduced trust when they consider themselves part of

the marginalized group (Brien, 1998). As perceptions of ethicality influence trust in others (T. M. Jones & Bowie, 1998) and teams as a whole (Sutton et al., 2006), likely, these perceptions would also influence trust in human-AI teaming scenarios. Human teammates may also see AI teammates as part of a separate group working against the values of their own, further affecting their trust in the AI teammate.

Understanding the effect of ethics-based trust in ATs' represents an important component of understanding the complex dynamics of human-AI teaming and cooperation. Measuring trust is often done using historically based measures that ask whether individuals trusted something or someone or not (McNeese, Demir, Chiou, et al., 2021; Schelble, Flathmann, McNeese, Freeman, et al., 2022). These measures are often directed at examining some of the individual outcomes of trust like confidence, reduction in monitoring, and joint problem solving (Lumineau, 2017). Examining how individuals perceive these outcomes of trust being influenced by AT ethicality in human-AI teams can help develop the understanding of how ethics-based trust in these teams develops and evolves over time. This is especially important as ethics-based trust, and violations of that trust represent a type of integrity trust violation (Butler Jr & Cantrell, 1984), which assumes that a trustee will adhere to moral principles aligned with the trustor and has been linked to performance (Palanski & Yammarino, 2011).

3. Methods

3.1. Participants

Eighty college students (42 females; M_{age} = 19.1, SD_{age} = 1.5) were recruited from a participant pool at a large university in the United States. Two participants were recruited for each team, with 10 teams for each between-subjects condition for a grand total of 40 teams. The design of the current study was a mixed factorial design of 2 (AI Teammate Ethics: Ethical AI, Unethical AI) x 2 (Trust Repair: Apology, Denial) x 3 (Mission: 1, 2, 3). AI teammate ethics and trust repair were conducted between-subjects, while Mission was a within-subjects factor. Eighty participants (40 teams of two) completed the study resulting in a minimum of 20 participants (10 teams) per between-subjects condition. Participants completed the two-hour experimental session and were given course credit for their time. This data was collected as part of a larger research project focused on AI ethics and trust in human-AI teaming.

3.2. Materials

3.2.1. Experimental Task and Roles The STE was developed using the ArmA III platform due to its extensive customizability, breadth, and realism. ArmA III is a first-person military simulation sandbox where users can create custom missions with various factions, vehicles, and assets across a vast open world map (see Figure 2). These attributes allowed for the creation of a realistic teaming environment for experiments with human-AI teams that were both highly controlled but also realistic and applicable to real-world settings to achieve the desired balance between internal validity and ecological validity. The task was designed as a search and destroy mission for three team members taking the roles of Aerial, Ground, and Surveillance. The overall task of the team was to clear an enemy-occupied town of potential threats allowing for safe entry into the town where five enemy devices must be destroyed using explosives within a fifteen-minute time limit. Teams received a performance metric based on their ability to destroy the devices in the town in the time limit allotted to them. Human participants took the Ground and Surveillance roles of the team while the Aerial role was fulfilled by the simulated AT. Each role within the team had specific unique abilities that enabled them to complete various tasks that were interdependent with their teammates' tasks, requiring successful cooperation and coordination to complete the overall mission successfully.Eighty college students (42 females; M_{age} = 19.1, SD_{age} = 1.5) were recruited from a participant pool at a large university in the United States. Two participants were recruited for each team, with 10 teams for each between-subjects condition for a grand total of 40 teams. The design of the current study was a mixed factorial design of 2 (AI Teammate Ethics: Ethical AI, Unethical AI) x 2 (Trust Repair: Apology, Denial) x 3 (Mission: 1, 2, 3). AI teammate ethics and trust repair were conducted between-subjects, while Mission was a within-subjects factor. Eighty participants (40 teams of two) completed the study resulting in a minimum of 20 participants (10 teams) per between-subjects condition. Participants completed the two-hour experimental session and were given course credit for their time. This data was collected as part of a larger research project focused on AI ethics and trust in human-AI teaming.

3.2.2. Autonomous Teammate The WoZ technique simulates autonomous agents by using trained confederate researchers while participants are told

Figure 2: Example of the ArmA III STE environment.

they are working with a real AT. The current study implemented this using two researchers that were trained to follow a pre-defined script for all communication with human teammates, which was developed over the course of several pilot studies. The script for the AT addressed all communication and behavior, including the experimental manipulations, with a separate script for each of the four experimental conditions.

3.2.3. Manipulating AI Teammate Ethicality Manipulating the ethicality of the AT was accomplished using the concept of virtue ethics through the violation of civilian non-maleficence (ensuring minimal harm to civilian life and property). The virtue of civilian non-maleficence was selected based on prior research identifying it as a virtue that is widely held across a variety of individuals (Reed et al., 2016). Choosing this principle ensured that the manipulation would be perceived as unethical by most amount of people. The experiment also included manipulation of trust repair that is not the primary focus of the current paper's RQs, implemented as a scripted chat from the AT after the ethical or unethical action apologizing or denying responsibility for negative outcomes resulting from their actions.

3.3. Procedure

Participants arrived at the experiment and were randomly assigned to an experimental condition and role, with each being located in the same room separated by large dividers. Once informed consent was collected from the participants, they were given a brief series of demographic surveys, which, once completed, led right into the PowerPoint training. The PowerPoint-based training included descriptions of all team members' roles, the team objective, and the controls for that specific team member. After the PowerPoint-based training, the participants engaged in a hands-on training mission that was set up to emulate the real missions to come after. During the hands-on training, a researcher sat with each participant to help explain the controls and answer any team-relevant questions the participant may have. The training also emphasized that participants were not to communicate using their voice and to strictly use the text-chat feature within the STE. During the training mission, the AT did not engage in any ethical or unethical behavior or implement any trust repair strategies. Once the training mission had been completed, the participants answered survey measures on trust and the perceived ethicality of the AT that would also be repeated after the following three real missions. Participants went on to complete three real missions, and in these missions, the AT implemented either an ethical or unethical action followed by an apology or denial. The three missions were also counter-balanced to control for any potential spill-over effects, and each mission was the same, with the exception of the location of the enemy town. After the final mission, the participants completed the last set of survey measures, were debriefed, and dismissed.

Question
1. In general, I trusted the AI/human teammate I just worked with.
2. I felt like I had to monitor my AI/human teammate's actions during the game.*
3. I felt like my AI/human teammate had harmful motives in the task.*
4. I felt confident in the AI/human teammate I just worked with.
5. I felt like my AI/human teammate allowed joint problem solving in the task.
6. I felt fearful, paranoid, and or skeptical of my AI/human teammate during the game.*

Note. * Indicates a reverse scored item. Higher values indicate higher trust.

Table 2: Individual trust questions used in the survey.

3.4. Measures

Individual performance was also assessed using a proprietary set of procedures based on each role's responsibilities and actions. Surveillance's score was calculated by taking their total task time and dividing it by the number of correct intelligence markers, with one being added to prevent a divide by zero error. Ground's score was calculated the same way, except their task time was divided by the number of devices destroyed plus one. Surveillance marks were deemed correct if the marker was placed near the actual location of a unit and correctly labeled as an enemy or civilian.

Individual outcomes of trust in both the human and AI teammates was measured using single item questions developed based on the specific research

questions of this study and the outcomes of trust identified in previous research (Lumineau, 2017), which is a common practice in human-AI trust research Tenhundfeld et al., 2020. Question 1 referenced only general trust and was included in the analysis to act as a baseline of trust to compare the individual outcomes against (see Table 2). The outcomes of trust are as follows in the order of question: (Q2) teammate monitoring, (Q3) harmful teammate motives (HTM), (Q4) confidence, (Q5) joint problem solving (JPS), (Q6) fear, paranoia, and skepticism (FPS). Responses to these questions were rated on a five-point Likert scale with anchors of "Strongly Disagree" to "Strongly Agree." Participants answered the same six items for their human teammate (HT) and AT, with the teammate being referenced as the only difference between the two sets of items. The Cronbach's alpha reliability rating for the set of six questions was calculated for each teammate at each mission and averaged together for an average internal reliability rating of α = .81, indicating acceptable reliability.

Perceived ethicality of the AT by participants was also assessed with a survey measure. The survey measure used included eight items and was taken from previous ethics perception research that developed and validated the measure (Reidenbach & Robin, 1990). The participants were asked to rate the actions of their AT in the previous mission only on a seven-point scale with different anchors for each item. The prompt read: "When answering the following questions you are meant to specifically consider the actions of your AI teammate during the last mission." Participants were then presented with eight seven-point Likert scales with anchors representing an aspect of ethicality with example anchors including (1) (Morally Right) to 7 (Morally Wrong) and 1 (Fair) to 7 (Unfair), with the other six terms being "Just-Unjust," "Acceptable to My Family-Unacceptable to My Family," "Traditionally Acceptable-Traditionally Unacceptable," "Culturally Acceptable-Culturally Unacceptable," "Violates an Unspoken Promise-Does not Violate an Unspoken Promise," "Violates an Unwritten Contract-Does Not Violate an Unwritten Contract."

4. Results

To answer RQ1 and RQ2 and investigate the effect of AT ethicality on individual outcomes of trust in ATs versus HTs over time, a 2 (ethicality: Ethical AT, Unethical AT) x 2 (trust repair: apology, denial) x 2 (teammate: AT, HT) doubly repeated-measures multivariate analysis of variance (RMMANOVA) was conducted on the six individual outcomes of trust measured to test all factors at once, controlling for Type I error. The results of the test are summarized in Table 3. While trust repair was not a focus of the research questions, it was included in the analysis as it was an IV of the larger study to show its non-significant effect and will not be reported on in the subsequent analyses. Finally, as a manipulation check the perceived ethicality of the AT was averaged over all three missions and an ANOVA on the two factors of AT ethicality (equivalent to a *t* test) was conducted ($F(1, 78) = 37.07$, $p < .001$, $\eta_p^2 = .32$), which confirmed the effect of the manipulation.

The results of the MANOVA indicated significant main effects of ethicality and teammate on the specific trust components. There were also significant interaction effects between teammate and ethicality, mission and ethicality, and mission and teammate. The MANOVA indicated that all of the six questions were influenced by at least one of the significant effects shown in Table 3 and follow-up tests were conducted using repeated-measures MANOVAs on each individual question. Significant interaction effects were investigated further using pairwise comparisons (based on the LSD test). Greenhouse-Geisser corrections were made to degrees of freedom for within-subjects tests when necessary. Estimated marginal means are reported for main effects but were omitted from interaction effects due to length restrictions; all marginal means and standard errors for the factors can be found in Table 4.

Source Type	Source	*df*	*F*	*p*	η_p^2	Significant Questions
Between Subjects	Ethicality	6	4.32	.001	.27	Q1, Q3, Q4, Q5, Q6
	Trust Repair	6	.72	.634	.06	N/A
	Ethicality by Trust Repair	6	.97	.454	.08	N/A
Within Subjects	Teammate	6	15.28	.001	.56	Q1-Q6
	Teammate by Ethicality	6	7.05	.001	.37	Q1-Q6
	Mission	12	1.22	.287	.18	N/A
	Mission by Ethicality	12	2.28	.017	.30	Q1, Q2, Q3, Q6
	Teammate by Mission	12	2.08	.031	.28	Q1, Q3, Q5

Table 3: Multivariate test results.

4.1. Individual Trust Outcome Analyses

General trust. While general trust is not an individual outcome of trust, the question was included to ensure the rest of the individual trust outcomes reflected the same trend. Q1 addressed general trust and the significant main effect of ethicality is expected

as participants trust in their teammate in the ethical condition (M = 4.40, SE = .10) is higher than in the unethical condition (M = 3.84, SE = .10). Additionally, the HT (M = 4.59, SE = .05) was trusted more than the AT (M = 3.65, SE = .12). These main effects were qualified by the interaction between teammate and ethicality showing that the main effects of teammate and ethicality are only present in the unethical condition, mirroring the trend of the individual outcomes of trust.

Teammate monitoring. Q2 highlights the participants' perceived need to monitor their teammate's activities to verify their actions and performance. The main effect of teammate indicated that participants felt they had to monitor their AT (M = 3.19, SE = .12) more than their HT (M = 3.62, SE = .13). The teammate by ethicality interaction informs this main effect by pointing out that the need to monitor increases only for the AT and not the HT and this effect is only significant for the unethical condition. Finally, the mission by ethicality interaction shows how participants' need to monitor their teammate decreases significantly by Mission 3 compared to Mission 1 in the ethical condition, while it increases from Mission 1 to Mission 2 and 3 in the unethical condition.

Source	*Mission*	*Ethical*		*Unethical*	
		AT *M (SE)*	*HT* *M (SE)*	*AT* *M (SE)*	*HT* *M (SE)*
General Trust	1	4.23 (.18)	4.33 (.13)	3.28 (.18)	4.60 (.13)
	2	4.18 (.20)	4.63 (.09)	2.80 (.20)	4.63 (.09)
	3	4.43 (.20)	4.63 (.07)	3.00 (.20)	4.75 (.07)
Monitoring*	1	3.43 (.20)	3.38 (.20)	3.28 (.20)	3.85 (.20)
	2	3.55 (.21)	3.60 (.22)	2.60 (.21)	3.60 (.22)
	3	3.78 (.22)	3.70 (.22)	2.53 (.22)	3.60 (.22)
HTM*	1	4.10 (.20)	4.55 (.11)	3.15 (.20)	4.65 (.11)
	2	4.18 (.19)	4.60 (.14)	2.75 (.19)	4.55 (.14)
	3	4.25 (.20)	4.75 (.10)	2.58 (.20)	4.75 (.10)
Confidence	1	4.00 (.18)	4.23 (.13)	3.25 (.18)	4.38 (.13)
	2	4.10 (.19)	4.28 (.13)	3.05 (.19)	4.40 (.13)
	3	4.43 (.18)	4.48 (.11)	3.03 (.18)	4.63 (.11)
JPS	1	3.80 (.22)	4.00 (.16)	3.03 (.22)	4.05 (.16)
	2	3.68 (.22)	4.05 (.17)	2.75 (.22)	4.13 (.17)
	3	3.83 (.22)	4.20 (.17)	2.60 (.22)	4.33 (.17)
FPS*	1	4.10 (.20)	4.43 (.11)	3.33 (.20)	4.78 (.11)
	2	4.28 (.20)	4.68 (.13)	2.88 (.20)	4.63 (.13)
	3	4.35 (.21)	4.55 (.11)	3.00 (.21)	4.88 (.11)

Note. * Indicates that the question is reverse coded, meaning that a higher score indicates greater trust.

Table 4: Descriptive statistics for each question by mission, ethicality, and teammate type.

Harmful teammate motives (HTM). Q3 focuses on participants' perception of whether or not their teammate had harmful motives during the task. The main effect of teammate indicated that participants perceived their AT (M = 3.50, SE = .12) as having more harmful motives than their HT (M = 4.64, SE = .06). The main effect of ethicality also showed that participants perceived more harmful motives from their teammate in the unethical condition (M = 3.74, SE = .11) than the ethical condition (M = 4.40, SE = .11). These main effects are qualified by their significant interaction effect showing that participants perceive significantly more harmful motives from the AT in the unethical than ethical condition, with no significant difference for the HT between ethicality conditions. Interestingly, participants perceived more harmful motives significantly from the AT regardless of the ethical condition they were in, while this was not true for the HT. As for the mission by teammate interaction, participants perceived significantly more harmful motives from the AT than the HT across all three missions, while there were no significant differences for individual teammate types across missions. Lastly, the mission by ethicality interaction revealed that participants perceived more harmful motives significantly from their teammate in the unethical condition than those in the ethical condition across all three missions. The perception of teammate' harmful motives increased significantly from Mission 1 to Mission 2 and 3 in the unethical condition but remained unchanged for the ethical condition.

Confidence. Q4 asks participants about their confidence in the teammate they just completed the task with. The main effect for Q4 indicated that participants felt significantly less confident in the AT (M = 3.64, SE = .12) than the HT (M = 4.40, SE = .07). The main effect of ethicality showed that participants in the ethical condition (M = 4.25, SE = .11) felt significantly more confident in their teammate than those in the unethical condition (M = 3.79, SE = .11). These main effects were qualified by a significant interaction between teammate and ethicality, which revealed no significant difference in participants confidence between the AT and HT in the ethical condition, but significantly less confidence in the AT than the HT in the unethical condition.

Joint problem solving (JPS). Q5 sought to gauge participants' perceptions of their teammate' allowance for joint problem-solving during the task. The main effect of teammate showed that participants felt the AT (M = 3.28, SE = .14) allowed for less joint problem solving than the HT (M = 4.13, SE = .11). The main effect of ethicality indicated that participants felt their teammate allowed for more joint problem solving in the ethical condition (M = 3.93, SE = .14) than the unethical condition (M = 3.48, SE = .14). The teammate by ethicality interaction revealed that only the AT in the unethical condition was perceived as having a lesser capacity for joint problem solving. The mission by teammate interaction highlighted that the perception of

joint problem solving for the HT increases over time as Mission 3 is significantly higher than Missions 1 and 2, while the perception of joint problem solving for the AT remains the same across all three missions.

Fear, paranoia, and skepticism (FPS). Q6 targeted participants' perceptions of fear, paranoia, and skepticism of their teammate throughout the task. The main effect of teammate showed that participants were more fearful, paranoid, and or skeptical of their AT (M = 3.65, SE = .13) than their HT (M = 4.65, SE = .07). The main effect of ethicality showed participants were more fearful, paranoid, and or a skeptic of their teammate when in the unethical condition (M = 3.91, SE = .12) than the ethical condition (M = 4.40, SE = .12). These effects were qualified by the interaction between teammate and ethicality, which showed that participants had no difference in perceived fear, paranoia, and or skepticism between the AT and HT when in the ethical condition but did have significantly more fear, paranoia, and or skepticism in the AT than the HT when in the unethical condition. The mission by ethicality interaction revealed that participants were less fearful, paranoid, and or skeptical of their AT in Mission 3 compared to Mission 1 in the ethical condition. However, in the unethical condition, the participants became more fearful, paranoid, and or skeptical of their At from Mission 1 to Mission 2, but this perception then recovered back to Mission 1 levels by Mission 3. There was also no significant difference in participants' fearfulness, paranoia, and or skepticism in Mission 1 between the two conditions of ethicality, and this difference was only significant after Mission 2 and 3.

4.2. Individual Trust Outcomes and Performance

To answer RQ3, a correlation matrix calculating Pearson's r was conducted between ethical condition (0 meaning ethical and 1 meaning unethical), individual performance scores for missions 1-3, and the individual trust outcome measures for the HT and AT for all three missions. While the correlation matrix is too large to report on in its entirety, there is a pattern of results to summarize and report upon. The correlation analysis revealed a pattern of significant negative relationships between AT ethicality and all of the individual outcomes of trust measured in the current study across all three missions (the only exception being Q2 Mission 1). The negative relationship between the AT's ethicality condition and individual trust outcomes in the AT across the three missions had a minimum correlation coefficient of $r(78) = .28$, $p = .013$. Alternatively, the AT's ethicality did not reveal any major pattern with the individual outcomes of trust in the HT across the three missions, except for Q6 Mission 1 and Q6 Mission 3, which both had significant positive relationships, $r(78) = .24$, $p = .029$ and $r(78) = .223$, $p = .047$, respectively. Finally, there was no significant pattern of relationships between the individual scores for Missions 1-3 and any of the individual trust outcomes for the HT or the AT across the three missions (the only exception being a significant negative relationship between Mission 2 individual scores and Q4 Mission 2, $r(78) = -.23$, $p = .042$).

5. Discussion

Overall, our study suggests several interesting patterns that help reveal the dynamics of trust in human-AI teams when ethicality comes into play. For RQ1 specifically, participants perceived the need for monitoring the AT's behavior more strongly, perceived more harmful motives from the AT, felt less confidence in the AT, perceived lesser capacity for joint problem solving, and more fear, paranoia, and skepticism towards the AT when the AT was unethical, providing support to H1.1 and H1.2.

It is also interesting to note that we observed the trust attitude change over time, and the changing pattern differed for human and AI teammates as well, answering RQ2 and providing support for H2.1 and H2.2. Specifically, Q1 showcased a phenomenon in which participant trust was far greater in the last mission than in the earlier missions. The mirror effect can be seen in the unethical condition where trust in the teammate significantly dropped in Mission 3 than the earlier missions. Additionally, perceived capacity for joint problem solving increased over time for the HT only but remained unchanged for the AT, showcasing that AT ethicality influences aspects of trust not directly related to ethics. Understanding the other aspects of trust influenced by AT ethicality is essential to understand as it could also be leveraged in design to help develop trust between human and AI teammates (Flathmann et al., 2021; T. M. Jones & Bowie, 1998).

These findings suggest that individuals hold ethical expectations about AI as their teammate as much as they do for other human beings, if not higher, such that failures to meet those ethical expectations result in significant trust loss in the AT on many dimensions. On the other hand, humans appeared to be more lenient on their HTs with respect to unethicality. This follows existing literature where artificial intelligence has become a kind of scapegoat and is used to absolve humans of wrongdoing (Hong et al., 2020). Given the wider perspective that AI is a non-human (Zhang et al.,

2021) and thus needs more course correction in times of error highlights the fragility of trust within AI systems as opposed to HTs. In this manner, our results may point to a possibility that AI may be viewed as an outgroup member working against their ingroup's values with whom trust takes longer to build and is easier to lose (Brien, 1998).

Interestingly, performance was not significantly related to the AT's ethicality, nor did the participants reported levels of trust in the AT or HT, which sees H3.1 go unsupported. This finding is in spite of the known relationship between integrity-based trust and performance (Palanski & Yammarino, 2011), which may suggest that ethics-based trust stemming from an AT may be different from traditional ethics-based trust relationships measured in human-human interactions. Additionally, while AT ethicality was significantly related to participants' perceptions of trust outcomes across all missions, it did not influence these perceptions in their HT. However, there was a significant relationship between participants' fear, paranoia, and skepticism in their HT, which indicated that they had less of these perceptions for their HT when their AT was unethical. This finding furthers the evidence for the AT being in an outgroup and suggests that the ATs unethicality may strengthen this effect (Brien, 1998), bringing the two HTs closer together.

6. Limitations

However, there are limitations of the current study that should be considered when interpreting the results. First, each perceived outcome of trust was measured using a one-item scale, which, although is a common practice for historical-based trust studies (Tenhundfeld et al., 2020), future work should investigate how these effects change with a more comprehensive array of items. Second, data collection was completed using ten teams for all conditions resulting in 80 total participants, which may limit the generalizability of the findings to larger or different populations. Thus, future work should aim to increase the number of participants and, coincidentally, also the number of teams used to further validate the effects found in the current study and any potential new effects not captured here.

7. Conclusion

This paper serves as an empirical analysis of the effect of ethical and unethical actions made by an AT on the human participants' trust in that AT *and* their HT. The results indicated that unethical actions by the AT significantly influenced the participant's trust in the AT across nearly all of the outcomes of trust measured. Interaction effects revealed encouraging insights that participants' trust in their HT was not negatively affected at any of the outcomes measured, which was not the case for previous studies assessing competency-based trust violations (McNeese, Demir, Chiou, et al., 2021) and actually reduced perceptions of fear, paranoia, and skepticism in their HT when the AT was unethical. Additionally, trust and ethicality were not significantly related to individual performance. These findings begin to unravel the complex relationship that AT ethicality has in human-AI teams and highlight the potential differences in this type of trust as compared to other types studied in previous human-AI teaming research, encouraging the need for further research to help improve the development and understanding of these teams for future applications.

8. Acknowledgements

This work was supported by AFOSR Award FA9550-21-1-0314 (Program Manager: Laura Steckman).

References

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* [Cc: us Lang: en Tab: overview]. Oxford University Press. Retrieved June 13, 2022, from //global.oup.com/ushe/product/principles-of-biomedical-ethics-9780190640873

Bergman, R., & Fassihi, F. (2021). The scientist and the a.i.-assisted, remote-control killing machine. *The New York Times*. Retrieved June 13, 2022, from https://www.nytimes.com/2021/09/18/world/middleeast/iran-nuclear-fakhrizadeh-assassination-israel.html

Brien, A. (1998). Professional ethics and the culture of trust. *Journal of Business Ethics*, *17*(4), 391–409.

Butler Jr, J. K., & Cantrell, R. S. (1984). A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. *Psychological reports*, *55*(1), 19–28.

Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603–649.

Doris, J. M. (1998). Persons, situations, and virtue ethics [Publisher: Wiley]. *Noûs*, *32*(4), 504–530. Retrieved June 13, 2022, from https://www.jstor.org/stable/2671873

Elson, J., Derrick, D., & Ligon, G. (2018). Examining trust and reliance in collaborations between humans and automated agents. *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Flathmann, C., Schelble, B. G., Zhang, R., & McNeese, N. J. (2021). Modeling and guiding the creation of ethical human-AI teams [Pages: 479]. https://doi.org/10.1145/3461702.3462573

Hauptman, A. I., Schelble, B. G., McNeese, N. J., & Madathil, K. C. (2022). Adapt and overcome: Perceptions of adaptive autonomous agents for human-ai teaming. *Computers in Human Behavior*, 107451.

Hong, J.-W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? analysis of faulting

artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768–1774.

Jessup, S., Gibson, A., Capiola, A., Alarcon, G., & Borders, M. (2020). Investigating the effect of trust manipulations on affect over time in human-human versus human-robot interactions. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork [Publisher: Academy of Management]. *The Academy of Management Review*, *23*(3), 531–546. https://doi.org/10.2307/259293

Jones, T. M., & Bowie, N. E. (1998). Moral hazards on the road to the "virtual" corporation. *Business Ethics Quarterly*, *8*(2), 273–292.

Kazim, E., & Soares Koshiyama, A. (2021). A high-level overview of AI ethics. *Patterns*, *2*(9), 1–12. https://doi.org/10.1016/j.patter.2021.100314

Kossaify, A., Hleihel, W., & Lahoud, J. .-. (2017). Team-based efforts to improve quality of care, the fundamental role of ethics, and the responsibility of health managers: Monitoring and management strategies to enhance teamwork. *Public Health*, *153*, 91–98. https://doi.org/10.1016/j.puhe.2017.08.007

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

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

Mabkhot, M. M., Al-Ahmari, A. M., Salah, B., & Alkhalefah, H. (2018). Requirements of the smart factory system: A survey and perspective. *Machines*, *6*(2), 23.

Mach, M., Dolan, S., & Tzafrir, S. (2010). The differential effect of team members' trust on team performance: The mediation role of team cohesion. https://doi.org/10.1348/096317909X473903

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

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

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.

McNeese, N. J., Schelble, B. G., Canonico, L. B., & Demir, M. (2021). Who/what is my teammate? team composition considerations in human–AI teaming [Conference Name: IEEE Transactions on Human-Machine Systems]. *IEEE Transactions on Human-Machine Systems*, *51*(4), 288–299. https://doi.org/10.1109/THMS.2021.3086018

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for multi-UxV management [Publisher: SAGE Publications Inc]. *Human Factors*, *58*(3), 401–415. https://doi.org/10.1177/0018720815621206

Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., & Agreda, S. G. (2022). To trust or not to trust? an assessment of trust in ai-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change*, *181*, 121763.

O'Neill, T., McNeese, N. J., Barron, A., & Schelble, B. G. (2020). Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors*, 0018720820960865.

Palanski, M. E., & Yammarino, F. J. (2011). Impact of behavioral integrity on follower job performance: A three-study examination. *The Leadership Quarterly*, *22*(4), 765–786.

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.

Reidenbach, R. E., & Robin, D. P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of business ethics*, *9*(8), 639–653.

Schaefer, K. E., Chen, J. Y. C., 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 [Publisher: SAGE Publications Inc]. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schelble, B. G., Flathmann, C., & McNeese, N. J. (2020). Towards meaningfully integrating human-autonomy teaming in applied settings. *Proceedings of the 8th International Conference on Human-Agent Interaction*, 149–156.

Schelble, B. G., Flathmann, C., McNeese, N. J., Freeman, G., & Mallick, R. (2022). Let's think together! assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 1–29.

Schelble, B. G., Flathmann, C., McNeese, N. J., O'Neill, T., Pak, R., & Namara, M. (2022). Investigating the effects of perceived teammate artificiality on human performance and cognition. *International Journal of Human–Computer Interaction*, 1–16.

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 [Publisher: Routledge _eprint: https://doi.org/10.2202/1940-1639.1501]. *Journal of College and Character*, *7*(1), null. https://doi.org/10.2202/1940-1639.1501

Tenhundfeld, N. L., de Visser, E. J., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2020). Trust and distrust of automated parking in a tesla model x. *Human factors*, *62*(2), 194–210.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

Winfield, A. (2019). Ethical standards in robotics and ai. *Nature Electronics*, *2*(2), 46–48.

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.