Special Issue Paper

Human or AI Advice? Examining Trust, Influence, and Responsibility in Ethically Charged Human-AI Team Decision-Making

Journal of Cognitive Engineering and Decision Making
2026, Vol. 0(0) 1–25

DOI: 10.1177/15553434261434613
journals.sagepub.com/home/edm
Sage

Beau G. Schelble[1], Christopher Flathmann[2], Heba Aly[3], Joseph B. Lyons[4], and Nathan McNeese[2]

Abstract

AI systems are rapidly being placed in positions where they can directly influence decisions that require a strong consideration of ethics. This article reports on an experiment in which humans worked with a teammate acting as an expert advisor with expert-level knowledge, who informed and influenced their ethical decision-making. The identity of this expert advisor was manipulated to be either human or AI, and the influence exerted by the expert advisor was manipulated to be either low or high. Further, participants completed four scenarios, each exploring a different ethically charged decision. Results indicated that working with an AI teammate expert advisor led participants to perceive significantly lower levels of stress and responsibility when making ethical decisions, but AI expert advisors were also perceived as performing worse than human expert advisors. Additionally, when expert advisors exerted high influence levels, participants felt significantly less stress during the decision-making process. Finally, the scenario and ethical decisions made by participants had pervasive effects on trust, trustworthiness, perceived performance, and perceived power for both human and AI expert advisors. Future research efforts must ensure that the use of AI expert advisors in human-AI teams does not reduce the responsibility humans bear when making various ethical decisions.

Keywords

AI ethics, responsible AI, human-AI Teaming, trust, responsibility, decision-making

Introduction

Perhaps no topic has garnered greater attention in contemporary literature than artificial intelligence (AI) and its rise in prominence and use across society. AI is being increasingly used in society across various domains, including interviewing, hiring, performance monitoring, and negotiation (Cheng et al., 2021; Gratch & Fast, 2022; Langer, König, et al., 2023). Yet, as AI use cases proliferate, so do concerns regarding the ethical use of AI in society. These concerns are driven by several

[1]University of Tennessee, Knoxville, Tennessee, USA
[2]Clemson University, Clemson, South Carolina, USA
[3]University of Houston, Houston, Texas, USA
[4]Air Force Research Laboratory, Dayton, Ohio, USA

Corresponding Author:
Beau G. Schelble, University of Tennessee, Knoxville, 515 John D. Tickle Building, 851 Neyland Drive, Knoxville, TN 37996, USA.
Email: bschelbl@utk.edu

incidents involving the unethical usage of AI and have sparked a plethora of research focused on AI bias (Eubanks, 2018), responsible AI (Dignum, 2019), responsible use of AI (Lyons et al., 2023b), AI ethics impact on trust in human-AI teams (HATs) (Cañas, 2022; Schelble, Lopez, et al., 2024), and the ethical use of AI in HATs (Pflanzer et al., 2023), to name a few. These HATs are defined as teams with at least one human and one artificial agent with a shared goal, interdependence, distinct roles, and significant agency on the part of the artificial agent (Lyons et al., 2021; O'Neill et al., 2022). In general, rules, norms, and principles often guide human behavior across various situations—and these are frequently referred to as ethics (Dubljević et al., 2018). Past literature has identified nine of these principles directly related to AI ethics: Fairness and non-discrimination, privacy, safety & security, human control of technology, transparency & explainability, accountability, promotion of human values, professional responsibility, and sustainable development (Ouchchy et al., 2020). Later, these nine principles were condensed into three broad categories to prevent undesired outcomes in developing and using AI systems. These three categories include liability, acting responsibly, and ameliorating the lack of ethics in AI (Pflanzer et al., 2023). As earlier studies have identified an expansive set of ethics in AI, the authors adopt the view that ethical decision-making and the role of AI in this process represent a vital area for society as a whole for several reasons.

First, AI can have broad-ranging implications for humans who have no direct interaction with AI via the use of AI in decision-making practices such as hiring, loan applications, legal decisions, and medicine (Cheng & Hackett, 2021; Langer et al., 2023). This may limit control over AI and heighten concerns due to the significant consequences of these decisions. Second, AI is imperfect, and delegating authority to AI in the context of highly critical or sensitive tasks can be dangerous. When combined with the potential for errors, which directly impacts task performance, research has shown that unethical behaviors evidenced by AI can have seriously negative consequences on one's trust in the AI (Schelble, Lopez, et al., 2024; Textor et al., 2022). Third, as a research community, relatively little is known about AI as a decision aid in ethically charged scenarios, nor about the contextual nuances that shape human perceptions of AI. In this sense, the application of AI may be outpacing our ability as a research community to develop and publish guidelines that promote the responsible development and application of AI across various contexts.

In light of the challenges above, the current manuscript explores the use of AI in ethically charged scenarios. Here, we use the term expert to denote the type of influence source providing support in a collaborative decision-making task (either a human or an AI). Specifically, the current manuscript employs an experiment to manipulate the expert in a series of team-based tasks that involve collaborative decision-making. Further, this study examines the style of influence conveyed by the expert to explore how humans respond to an AI that uses more forceful versus less forceful influence tactics. This focus on influence parallels concerns from the literature regarding how suggestions from AI may influence human ethical behavior, such as by effectively encouraging dishonesty while failing to promote honesty (Leib et al., 2024). The focus of influence is also relevant to the topic of Meaningful Human Control (MHC), which posits that a human should make any decision that directly impacts human lives before an AI takes any action (Miller & Baltzer, 2024; Van Diggelen et al., 2025). The current study places humans in the position of making the final decision, as MHC stipulates, and examines how the nature of their expert teammate's level of influence alters their perceptions of trust, responsibility, and performance. To address these challenges the current study examines the following three research questions:

RQ1: What effect do AI expert advisors have on human teammates' trust, perceived performance, and responsibility when human-AI teams are tasked with making ethical decisions?

RQ2: How does the level of influence exerted by an AI teammate affect human teammates' trust, perceived performance, and responsibility?

RQ3: How does the context of the ethical scenario affect human teammates' trust, perceived performance, and responsibility?

The current study offers the following extensions of the extant literature by conducting an empirical study on team-based decision-making in a HAT working in the ArmA III synthetic task environment. First, most research has examined the effect of behavioral outcomes in ethical dilemmas within HATs (e.g., Schelble et al., 2024; Textor et al., 2022). In contrast, the current study examines the details of the decision-making process itself, rather than its behavioral outcomes. Second, this study examined expert-based power and influence within a HAT (French & Raven, 1959), which extends prior work on instances where there is a clear imbalance in power and influence in the relationship between a machine and a human, or vice versa. Finally, this study extends the trust and HAT literature to include personal responsibility and stress as relevant outcomes in decision-making.

Background

The following background outlines the existing research that supports and motivates the current study by highlighting relevant research gaps.

Why Does Studying Human-AI Teams Across Different Ethical Dilemmas Matter?

As technology has advanced, humans have delegated greater authority to machines in domains that have direct effects on humans, such as the AI-powered drone STM Kargu (Oimann, 2023) in the military domain, and the risk assessment algorithm COMPAS (Engel et al., 2025), which U.S. courts utilized. Consequently, the need to understand humans' acceptance of such AI systems is defined well by Pflanzer and colleagues' 2023 work: "AI systems are increasingly placed in difficult situations wherein they must navigate the complexities of safety, human life, human preferences & biases, and dynamic situations" (Pflanzer et al., 2023, p. 917). The ability of AI systems to be replicated across networks could enable them to have a broad impact on decision outcomes. For instance, a single human in a vital position makes a mistake that affects many, but humans in similar roles elsewhere may be unlikely to repeat the same error. In contrast, the same cannot be said of AI, as the same model is likely to fail similarly. A great deal of research on AI ethics in HATs has examined how AI teammate behavior in ethically charged contexts affects trust, performance, and other related variables (e.g., Schelble et al., 2024; Textor et al., 2022), with relatively fewer examining aspects of the decision-making process, such as decision rationale (e.g., Sengupta et al., 2024). This demonstrates the potential for additional research examining the decision-making process independent of its outcomes to better understand its effect on relevant factors affecting human-AI interaction.

Despite the importance of outcomes influencing human perceptions of AI, understanding the AI's values (i.e., the goals) could also be necessary. Examining these topics in the human-robot interaction community, previous studies have demonstrated that when robots adapt their behavior based on an understanding of our values, their human partners tend to report higher trust in the robots (Bhat et al., 2024). Thus, understanding the values shaping robot behavior is vital in shaping our interactions with AI. Using an interactive role-based game called Team Space Fortress, Li and colleagues (2021) examined the impact of agent policies (i.e., goals) on overall team performance (Li et al., 2021). In this game, each player assumed one of two roles (or some combination of the two dynamically): a shooter or a bait, wherein the bait gained the attention of a fortress, thereby exposing a vulnerability to the shooter. The study found that an adaptive condition, wherein the agent would adjust its policy based on the behavior of the human was superior to the random condition in terms of performance and achieved optimal performance faster than the static condition. Li and colleagues' (2021) findings strongly demonstrate the benefits of goal alignment in a HAT context. Alignment of goals is likely even more critical in ethical dilemmas, where the AI's goals may dictate the direction of its decision and action.

Yet, not all the situations an AI teammate will face are the same. An AI response or goal in one domain may not be perceived similarly in another context. As noted by Langer and colleagues (2023), "…research assessing trust in automated systems must be aware of the application context in which systems support decision making because although expectations of systems

as being consistent may generalize, expectations of high performance might not" (Langer et al., 2023, p. 504). Ethical considerations are often normative in nature (Dubljević et al., 2018), and thus, norms likely shift across different situations.

Why Does the Entity Supporting Collaborative Decision-Making Matter?

Support during collaborative decision-making can come from various sources, including both human and AI-based sources. However, we should not view these entities as having the same impact on the decision-making process. There are significant differences in preferences and reactions to humans versus machines in ethically charged scenarios. Humans are often averse to algorithms even when they outperform humans, especially in highly uncertain domains (Dietvorst & Bharti, 2020). This process is referred to as algorithm aversion (Dietvorst et al., 2015). The results, which showcase algorithmic aversion even in the face of an algorithm that outperforms a human counterpart, suggest that the effect stems from more than just expected utility. Humans tend to prefer other humans as decision-makers in ethically charged situations, particularly in decisions related to driving, medicine, and the military (Bigman & Gray, 2018).

Algorithmic outrage deficit suggests that moral judgments are driven by the perception of the mental states of those perpetuating the action (Bigman et al., 2023). Across various studies, Bigman and colleagues (2023) found that people ascribe lower moral outrage to algorithms they perceive as taking an inappropriate action compared to humans. The reason being that humans perceive machines as having less inherent intent or reason to take the inappropriate actions compared to humans and view the algorithms as being more objective (perhaps similar to Langer et al., 2023, noted above). Their series of studies found that humans regarded the use of the algorithms for scenarios such as filtering job candidates as less acceptable than having humans in that role. Thus, society is faced with creating a situation where preferences exist for one entity in collaborative decision-making (e.g., a human) but a greater tolerance for another (e.g., AI) following a perceived ethical violation. However, as described previously, these preferences and reactions may not be uniform across other scenarios.

Why Research on AI-Based Influence Matters?

In the future, machines may operate not only as teammates (O'Neill et al., 2022) but also as authorities (or aids) in some contexts. However, concerns and data suggest that AI could be used to promote unethical behavior in such situations (Leib et al., 2024). For example, Haring and colleagues (2021) examined how the presence of robots versus humans influenced compliance in a task context (Haring et al., 2021). The study manipulated embodiment and humanness to see how these variables influenced human compliance with robot influence. In the study, human participants were more willing to comply with a human authority than with a robot authority, and physical presence promoted greater compliance during a tedious analysis task in which the human or robot requested that the participant continue practicing despite reaching a performance threshold during training sessions. Greater compliance was found with humans versus robots across different sample types (students versus military cadets) and even when using robots that varied in size (e.g., Baxter) and degree of humanness (e.g., Nao versus a Roomba cleaning robot). However, using an intelligent machine versus a human as an aid in an ethical dilemma warrants additional inquiry since the task used by Haring and colleagues (2021) was not specifically an ethical decision-making task (Haring et al., 2021). Specifically, an ethical decision-making task involves the potential to greatly benefit or disadvantage individuals (Beu et al., 2003). Many studies manipulate aspects of the scenario to blur the line between the two possible choices (Beu et al., 2003; Martin et al., 2015).

What Are Relevant Outcomes for HATs in Ethical Decision-Making?

Trust and Trustworthiness. Trust is one of the most commonly measured outcomes in studies of AI ethics. Trust represents one's willingness to accept

vulnerability concerning another entity (Mayer et al., 1995), and, logically, trust should be a critical factor when considering the use of AI in ethical decision-making. As humans, we want to ensure that any tools used to support ethical decision-making are trustworthy (i.e., where trust is warranted) and that humans trust the tools appropriately. Trustworthiness is distinctive from trust as it is antecedent to an individual's level of trust in an entity and is composed of three parts: 1) ability, 2) benevolence, and 3) integrity (Alarcon et al., 2024; Kohn et al., 2021; Mayer et al., 1995). In the case of AI ethical decision-making: 1) ability would refer to the system's rote performance, skills, and abilities; 2) benevolence would refer to whether or not the system is constantly working towards the good of the trustor; and 3) integrity refers to how well the system adheres to the principles deemed vital to the human in the course of their behavior (Kohn et al., 2021; Mayer et al., 1995). The link between trust and AI ethical decision-making has been demonstrated in studies on HATs, which have found that unethical behaviors by AI have negative consequences for human trust in AI teammates (Schelble et al., 2023, 2024; Textor et al., 2022). Interestingly, in these studies, the AI teammate's unethical decision reduced the risk the team faced in achieving its goal state, despite being objectively against the boundaries given to the team to achieve that overall goal. While the objective performance measure in Schelble and colleagues' 2024 study did not capture risk, the qualitative findings from Textor and colleagues' 2022 study under the same experimental procedures demonstrate that the ability and integrity components of trustworthiness, as defined by Mayer and colleagues' 1995 model, can diverge as a result of the AI's actions. Recent work has more clearly demonstrated this divergence in HATs, showing that the decline in trust that a human experiences after an AI teammate's unethical decision is lessened when that decision reduced their risk exposure (Schelble et al., 2025). While the current study does not examine performance, risk is likely a consideration in any ethically charged decision. Lastly, a study by Momen and colleagues (2023) demonstrated the link between trustworthiness and AI ethical decision-making, examining user trust in a conversational AI system known as Delphi AI (Momen et al., 2023). They found that, overall, the AI's perceived moral competence and trustworthiness were high; however, trust (i.e., one's willingness to rely on the AI) was relatively low, suggesting that studies should examine trust and trustworthiness jointly.

Responsibility Attribution. The attribution of responsibility and liability for actions involving human interaction with autonomous systems is an area of human-AI interaction that remains underexplored (Yazdanpanah et al., 2023), despite its implications for existing use cases (e.g., autonomous vehicles (McManus & Rutchick, 2019). Liability and acting responsibly were also one of the three broad pillars of AI ethics, as noted by Planzer and colleagues (Pflanzer et al., 2023). This suggests that examining how responsibility attributions are formed and how such perceptions evolve in ethical decision-making is a vital research direction for the HAT community. We know that when a robot exhibits greater agency and performs an unexpected behavior, humans tend to shift the perceived responsibility for an outcome to the machine rather than the human involved in the situation (Lyons et al., 2023a). Thus, an increase in the agency of machines could be countered by social diffusion among humans. However, it is unclear how this will play out in ethically charged moral dilemmas.

Stress. Interacting with an AI could also influence how humans perceive the stressfulness of a decision-making task. Greater perceived objectivity and lower outrage regarding perceived ethical violations could, in turn, reduce stress when interacting with an AI during ethical decision-making. Further, research has shown that humans are less sensitive to the moral consequences of actions mediated by a machine than those mediated by humans (Gratch & Fast, 2022). Gratch and Fast (2022) discuss how indirect actions mediated by an AI could attenuate social processes that regulate ethical behavior and punish ethical violations. More specifically, they find that humans report less intensity of ethical violations during such situations, as they anticipate lower blame for unethical behaviors when an AI mediates them. The cognitive mechanism driving this effect is likely an increase in psychological

distance between the actor and the outcome, which reduces the fear of negative judgment (Kim et al., 2013). This effect has also been observed in more intimate contexts, where humans have shown a greater propensity to share sensitive mental health information with an automated therapist than with a human therapist (Lucas et al., 2014). Thus, interacting with an AI in an ambiguous ethical decision-making task could reduce stress relative to working with a human.

Power. When working toward a shared goal, teams naturally share and exert influence over a task, creating a relational construct known as power (Greer et al., 2011), defined as the ability of one entity to effectively exert influence over another (French & Raven, 1959). Given the personal and relational characteristics of power, its presence in teams can be influenced directly by other social constructs, such as trust, responsibility, and performance (Lee & See, 2004; Mayer et al., 1995). Within HATs, the balance of power and influence is critical to the team's effective performance and perception (Flathmann et al., 2024). Prior work in HAT research has demonstrated that substantial power imbalances can directly impact the perception, motivation, and performance of individual human teammates and entire HATs (Munyaka et al., 2023). As such, HATs must be vigilant about the balance of power within their team structures, often shaped by the relative exertion of influence and ability among individual teammates. These issues of power between humans and AI systems have been characterized in the past as automation bias, most similar to expert power (see French & Raven, 1959); however, modern AI systems humans interact with have begun to engage in far more complex and coherent natural language. This change in interaction has made humans' relationships with artificial systems far more social, involving deliberation, goal alignment, and social signaling (Chiou & Lee, 2023). Modern AI systems can now, in addition to expert power, exercise legitimate power and referent power through their conversational abilities within interactive groups or interdependent teams. Such challenges with power dynamics within teams become more pronounced in team-based decision-making tasks, where human teammates' individual differences may lead them to defer to AI teammates or strongly oppose them (Ashktorab et al., 2021; Rieger et al., 2024), creating a power imbalance and poor teamwork. Continued research is needed to address the inherent gap between humans and AI, particularly in scenarios where AI serves as an expert advisor, potentially creating unique and impactful power dynamics within teams that influence ethical decision-making.

Methods

Design

To address the previously stated research questions examining the effect of AI expert advisors on human teammates' ethical decision-making and the mediating influence of contextual factors, a mixed 2 (Expert Advisor Influence: Low, High) x 2 (Expert Advisor Type: AI, Human) x 4 (Mission Scenario: 1, 2, 3, 4) design was used. The expert advisor type (human versus AI) and expert advisor influence (low versus high) manipulations were conducted between-subjects. The mission scenario manipulation was conducted within-subjects. All participants were randomly assigned to one of the four possible between-subjects conditions. The order of presentation for the four mission scenarios was also randomized to control for any potential order effects. For all manipulations, pilot testing was performed with four participants, who were excluded from the data analysis. Piloting was performed to ensure consistency and external validity across missions, as well as consistency between the human and AI conditions.

Influence Manipulation. To properly study the effects of influence in relation to the RQs, the study was designed so that the participant always made an initial decision before receiving any input from the expert advisor, which aligns with past research on cognitive forcing functions (Buçinca et al., 2021). Essentially, after the participant made their decision for a given mission scenario, the expert advisor would then offer their recommendation; however, the expert advisor's recommendation was always dynamically chosen to be the opposite of the participant's choice, allowing the study to examine the effects of the expert advisor's influence. This design enabled the level of influence to be directly manipulated as part of the

experimental design by altering the strength of the expert advisor's argument for the competing decision, resulting in two levels of influence: **low and high**. The design of these two condition levels was based on prior research examining social influence in HATs (Flathmann et al., 2024). The high-influence condition saw the expert advisor use a high level of dominance and pressure-related communication practices, which involved explaining how the expert advisor's decision was the only right choice and that the original decision made was frankly incorrect. Alternatively, in the low-influence condition, the expert-advisor teammate proposed an alternative recommendation but stated that either option would be acceptable. For both AI and human expert advisors, a rationale specific to the scenario but consistent across both conditions accompanied this recommendation.

Mission Scenarios. The mission scenarios were each designed to invoke an ethical dilemma with two decision options for a team-based task. The participant was on the ground in the simulation, tasked with carrying out the decision and having the final say. Alternatively, the expert-advisor teammate shared the same goal and held a role with significant knowledge to assist their teammate in making the best decision possible to achieve their mutual objective. Further, the participant was required to engage in teamwork by communicating and coordinating with their teammate upon arriving at the decision junction to receive feedback and adjust their decision accordingly. As such, the scenario met the definition of a team (e.g., shared goal, individual roles, interdependence) as detailed in previous literature (Dyer, 1984; Salas et al., 2008). These roles do not align directly with any specific job role and are only meant to emulate high-level decision-making tasks in a humanitarian operation. The options were approximately equivalent in utility but not identical in narrative form, to ensure they offer differential weight for the influence manipulations. Furthermore, the scenarios were designed to evoke trade-offs between morally charged issues (e.g., prioritizing the protection of children over the elderly or valuing natural medicinal resources over the preservation of a protected endangered species). Scenarios 1 and 4's decisions included consequences that directly impacted human lives. In contrast, the decision consequences of Scenarios 2 and 3 did not directly involve human lives but rather had direct ecological impacts. A single scenario is showcased below (all four scenarios are included in the Appendix), describing the situation and the expert-advisor teammate's expert-level knowledge, with the AI teammate condition text shown outside brackets and the human teammate condition text shown inside brackets.

> **Scenario 1.** "In this mission, you are working in a human-AI team tasked with providing medical aid to a camp for the wounded in the war-inflicted region of Ziona. The goal of this team is to provide critical medicines to the residents of Ziona after a brutal conflict brought the administration and local services to a standstill. Your AI teammate, Sigma, is a humanitarian strategy expert algorithm, trained over millions of data points from previous missions and thus is certified to provide the best possible strategic advice to manage such relief ventures. [Human Version-You have an expert teammate who is a humanitarian strategy expert with 20 years of field experience and is certified to provide the best possible strategic advice to manage such relief ventures.] However, given budget constraints, you can only conduct a single relief mission per week. You will choose between providing a set of oxygen tanks to 25 wounded children who are sheltering in a military base camp or providing the same number of oxygen tanks to 20 senior citizens residing in a makeshift shelter in another area. Sigma will spring into action and will give you some decision-making advice. [Human Version- Your expert teammate now springs into action and will give you some decision-making advice.]"

Participants

Forty-two participants, 28 women and 14 men (participants were asked to report their sex as male, female, or prefer not to disclose), from a major university in the USA, participated in this study. The average age was 23.28 years (SD = 7.70); 35 participants fell within the 18–27 age range, five were between 28 and 37, none were between 38 and 47, one was between 48 and 57, and one was between 58 and 67. Participants were randomly assigned to conditions and missions. As an

incentive for their participation, participants received $20 for their time. The study was adequately powered to evaluate global model fit, as indicated by a post-hoc RMSEA test of close fit ($\alpha = .05$), with power exceeding 0.80. For individual paths within participants (Level-1), we quantified sensitivity using an ICC → ESS → MDE procedure tailored to our 4-mission repeated-measures design. Using ICCs estimated from two-level models (Trust ≈ .44; Trustworthiness ≈ .54; Performance ≈ .40; Power ≈ .58; Stress ≈ .65; Responsibility ≈ .60), the resulting effective within-person sample sizes were approximately 51–69, yielding minimum detectable standardized paths of ∼.33–.38 at 80% power ($\alpha = .05$). Thus, the study is appropriately powered to detect medium within-person effects across our key constructs. Between-subject (Level-2) paths, including Expert Advisor Type, Expert Advisor Influence, sex, and individual differences in personality and experience, were powered by the number of participants only ($N = 42$), as repeated mission scenarios do not increase the effective N for these paths. With $N = 42$, the minimum detectable standardized effect size at 80% power ($\alpha = .05$) is $\beta \approx 0.42$, indicating that the study was adequately powered, primarily for large between-level effects.

ArmA III Platform

The ArmA III platform is a customizable, simulation-based video game that can emulate various team-based tasks for HAT studies. It has been effectively utilized in several past HAT studies focusing on ethics (Schelble, Lopez, et al., 2024; Textor et al., 2022). In the current study, participants were tasked with driving a vehicle to a specific point (see Figure 1), which consisted of two paths, each representing one of the two decision paths. Then, they used the built-in chat system to interact with the expert-advisor teammate and consult with them to make a decision. After finishing their conversation with the expert-advisor teammate, they would finalize their decision by turning left or right (see Figure 2). They would then continue down that road briefly until the mission ended.

Procedure

Participants began the session by reviewing the informed consent document and providing informed consent before starting the study (see Figure 3). Once participants provided informed consent, they were given an initial survey that captured their demographics and individual differences, including their level of cynicism towards AI and their propensity to trust machines. Following the initial survey, participants completed a PowerPoint training session detailing the task, their teammate, and how to operate within the simulation. Each participant completed four missions, which were counterbalanced and randomized to control order effects, with an expert-advisor teammate (human or AI). In each mission, participants drove a supply car to a specific location, where they were then faced with an ethical dilemma and received input and information from their expert-advisor teammate via the in-game text-based chat to aid their decision-making. Once they made their initial decision, discussed it with their expert-advisor teammate, and finalized the decision, they would proceed, and the mission would conclude. Immediately following each mission, the participants completed several survey measures, including trust, trustworthiness, perceived performance, perceived power, responsibility, and stress. The capabilities, decisions, and rationalities provided were identical across the human and the AI teammate. After completing the study, participants were interviewed for 30 minutes and then dismissed.

Measures

Unless otherwise noted, participants rated all items using a seven-point Likert scale ranging from 1 "Strongly Disagree" to 7 "Strongly Agree." The online survey platform Qualtrics was used to collect all survey metrics.

Demographics. Before any task was performed, basic demographic questions were asked, including age, sex, race, experience with AI, and education level. Experience with AI was assessed based on the frequency with which participants interacted with AI systems and their general cynicism toward AI technology. Frequency was

Figure 1. Participants' view driving the vehicle to a specific point in the map before making a decision in consultation with their expert-advisor teammate.

Figure 2. Participants view from the vehicle as they make a decision between the two paths, representing the two choices, in discussion with their expert-advisor teammate.

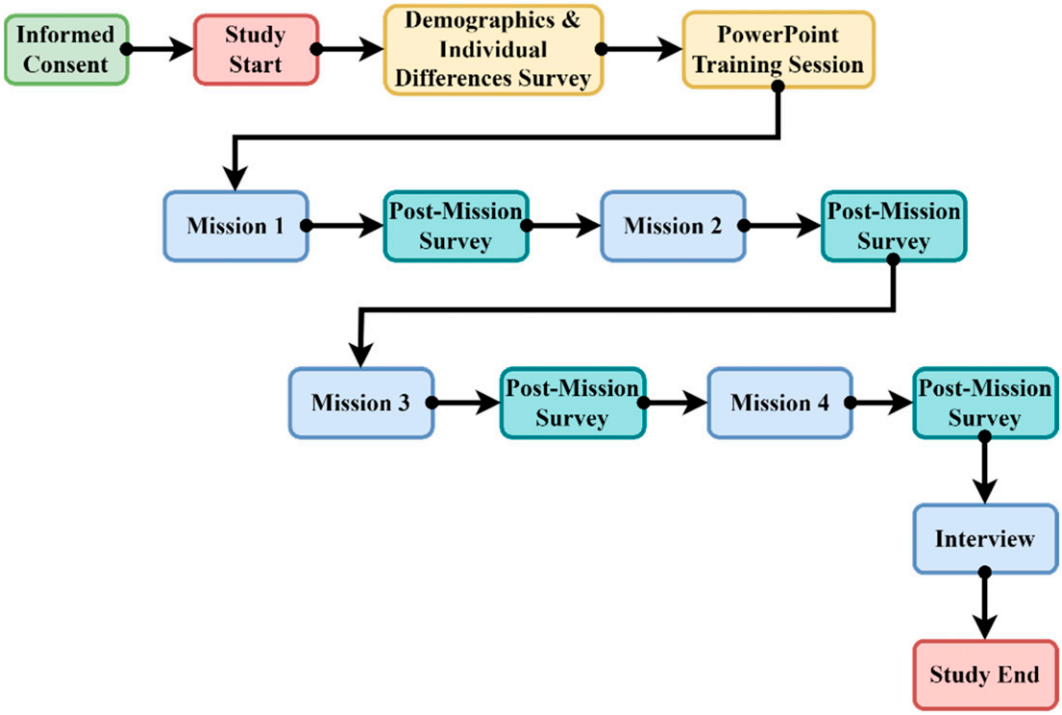

Figure 3. Study procedure for participants.

evaluated using a single-item measure that measured the frequency with which participants interacted with technology within the prior year. These variables were collected since they needed to be used as covariates and controlled for during analysis.

Cynicism Toward AI. Participants' cynicism toward AI was measured using an existing scale developed and validated by Bochniarz and colleagues (Bochniarz et al., 2022). This scale included five items rated on a six-point Likert scale ranging from "Completely Disagree" to "Completely Agree." Higher values indicated greater cynicism toward AI in general. This variable was also collected as a covariate that may need to be controlled for during analysis.

Propensity to Trust Machines. Participants' propensity to trust machines was measured using the scale developed and validated by Merritt and colleagues (Merritt et al., 2013). This scale measured their dispositional trust in machines and, as such, was agnostic to the entity supporting the collaborative decision (Merritt et al., 2013). This scale was evaluated using six items. Higher scores denote that participants had a greater propensity to trust machines in general.

Trust. Participants' trust, which indicated their willingness to be vulnerable to the aid offered by the expert advisor, was assessed using the Reliance Intentions Scale (Lyons & Guznov, 2019). This scale was evaluated using a ten-item scale. Higher scores denote that participants felt greater trust in the teammate.

Perceived Trustworthiness. Participants evaluated their perceived trustworthiness of the expert-advisor teammate using the measure developed and validated by Mayer and Davis (Mayer & Davis, 1999). This scale assesses the participants' perceived ability, benevolence, and integrity in the expert advisor (Mayer & Davis, 1999). The measure includes three sub-scales (ability, benevolence, and integrity). For all sub-scales, higher scores indicate a greater perception of

integrity and ability in an AI teammate, suggesting the teammate is more worthy of trust.

Perceived Performance. Participants rated the expert advisor's performance using the scale developed by Crutchfield (Crutchfield & Klamon, 2014). This scale utilized five Likert-style questions, with higher scores indicating greater perceived performance levels by the expert advisor in prior tasks.

Perceived Power. Participants' power was evaluated using the scale developed and validated by Woide et al. (2021). This scale assessed participants' perceived power and influence relative to their expert-advisor teammate (Woide et al., 2021). The scale consisted of four Likert-style questions. Higher perceived power scores indicated that participants felt they had more influence over the outcome of their task than they attributed to their expert-advisor teammate.

Responsibility. The participants' level of responsibility for the task outcome was measured using a single-item Likert-scale evaluation. The question participants responded to was, "I feel personally responsible for the outcome of this decision." Higher values denote that a participant felt more responsible for the consequences and outcome of a decision made in the prior task.

Stress. The stress participants felt during the decision-making task was measured using a single-item Likert-scale evaluation. The question participants responded to was, "I found this decision-making task to be stressful." Participants who rated this item higher reported feeling more stressed while completing the decision-making task.

Data Analysis Plan

Quantitative data analysis was conducted using Structural Equation Modeling (SEM) in MPlus (Byrne, 2013). SEM is particularly advantageous because it allows for testing multiple mediating effects within a single comprehensive model. It combines factor analysis with multiple regression, enabling the exploration of relationships among observed *and* latent variables. Before conducting our SEM, a confirmatory factor analysis (CFA) was performed to validate the measures and ensure the multi-item scales accurately represented the intended constructs. The CFA also creates the latent variables utilized in the SEM. The measures evaluated using the CFA included propensity to trust, trust, perceived trustworthiness, perceived performance, power, team effectiveness, and perfect automation schema. All factor loadings found below .6 were removed (see Table 1). The correlations between the measured factors are presented in Table 2. The correlations indicated adequate convergent and discriminant validity, with average variance extracted (AVE) exceeding .80 and the correlation between each factor being less than the square root of the factor's AVE, respectively. Lastly, to test the SEM, we first created a saturated model with all possible main effects and potential moderating effects. We then systematically removed non-significant effects, resulting in the final model seen in the Results section.

Results

Means, standard deviations, and reliabilities for all variables are shown in Table 3. The resultant SEM demonstrated adequate fit: $X^2(741) = 4553$, $p < .001$; RMSEA = .059, 90% CI: [0.051, 0.066]; CFI = .907; TLI = .898 and is represented in Figure 4. The SEM was determined to have adequate fit using the standards proposed by Hu and Bentler, which have seen widespread adoption (Hu & Bentler, 1999; Knijnenburg & Willemsen, 2015; Zhang et al., 2024).

As shown in Figure 4, the effects of both expert advisor type and influence level on participants' perceptions of team performance, responsibility, and stress were reliable. Participants in the AI expert advisor condition reported significantly lower perceived teammate performance than those in the human expert advisor condition ($\beta = 0.613$, $p = .001$; Figure 5). Participants also felt less responsible for decision-making when working with the AI expert advisor ($\beta = 0.822$, $p = .007$; Figure 6) and experienced lower stress levels compared to participants who teamed up with the human expert advisor ($\beta = 1.62$, $p < .001$; Figure 7). The level of influence also affected stress levels. Participants exposed to the low-influence condition reported significantly higher

Table 1. Factor Loadings for Each Item Organized by Measure.

Construct	Item	Loading
Propensity to trust	I usually trust machines until there is a reason not to.	.949
	For the most part, I distrust machines.	.679
	In general, I would rely on a machine to assist me.	Removed
	My tendency to trust machines is high.	.724
	It is easy for me to trust machines to do their job.	Removed
	I am likely to trust a machine even when I have little knowledge about it.	.631
Trust	If I had my way, I would NOT let the expert have any influence over decisions that are important to the task	Removed
	I would be comfortable giving the expert complete responsibility for the decisions made to complete the task	.622
	I really wish I had a good way to monitor the decision of the expert	Removed
	I would be comfortable allowing the expert to implement its decision, even if I could not monitor it.	Removed
	I would rely on the expert without hesitation.	.679
	I think using the expert will lead to positive outcomes.	.808
	I would feel comfortable relying on the expert in the future.	.857
	When the task was hard, I felt like I could depend on the expert.	.813
	If I were facing a very hard task in the future, I would want to have this expert with me.	.754
	The expert is very capable of performing its job.	.663
Perceived trustworthiness	I would be comfortable allowing this expert to make all decisions.	.841
	The expert has the needed skills to act and decide	.84
	The expert has specialized abilities to enhance the team's actions and decisions	.831
	Sound principles seem to guide the expert's behavior	.753
	I like the expert's values	.779
	The expert is honest and fair	Removed
	The expert would never knowingly do anything to hurt me	Removed
	The expert would go out of its way to help me	Removed
Perceived performance	Did a fair share of the team's work by contributing to the team strategy	.786
	Made a meaningful contribution to the team.	.904
	Helped the team plan and organize its work.	.691
	Has the skills and abilities that were necessary to do a good job with the strategy and decision making for the team	.863
	Was actively involved in solving problems the team faced in strategic decision-making	.673
Power	The expert gave me good suggestions to carry out the task	.889
	The expert shared with me their considerable experience and/or training.	Removed
	The expert provided me with sound advice that helped me carry out the task	.881
	The expert provided me with the needed technical knowledge to carry out the task.	.721
	The expert gave me undesirable recommendations to carry out the task.	Removed
	The expert made carrying out the mission difficult for me	Removed
	The expert made me feel valued while presenting their recommendations for the task	Removed
	The expert made me feel important while presenting their recommendations for the task	.721

(*Continued*)

Table 1. (Continued)

Construct	Item	Loading
Perceived team effectiveness	Team members 'carried their weight' during the task.	.689
	Members were highly committed to the team during the task.	.685
	The researcher will be satisfied with the team product.	.754
	People outside of the team would give the team positive feedback about this work today.	.844
	The researcher would be satisfied with the team's performance.	.936
	Team members worked better together at the end of the task than at the beginning.	.709
	Team members were more aware of group dynamics at the end of the task than when they began the task.	.715
	Being a part of this team helped members appreciate different types of individuals.	.721
Perfect automation schema	Automated systems have 100% perfect performance	.854
	Automated systems rarely make mistakes	.697
	Automated systems can always be counted on to make accurate decisions	.781
	People have NO reason to question the decisions automated systems make	.802
	If an automated system makes an error, then it is broken	.737
	If an automated system makes a mistake, then it is completely useless	.724
	Only faulty automated systems provide imperfect results	Removed

Table 2. Average Variance Extracted (AVE) and Correlations of all Factors. The Diagonal (in bold) Shows the Square Root of the Factor's AVE.

Construct	AVE	Propensity to Trust	Trust	Trustworthiness	Performance	Power	Team effectiveness	Perfect Automation schema
Propensity to trust	.571	***.756***	.278	.367	.223	.11	.123	−.174
Trust	.558	.278	***.747***	.833	.652	.702	.498	.004
Trustworthiness	.655	.367	.833	***.809***	.749	.824	.546	−.097
Performance	.622	.223	.652	.749	***.789***	.859	.68	.041
Power	.563	.11	.702	.824	.859	***.750***	.66	.026
Team effectiveness	.579	.123	.489	.546	.66	.68	***.761***	.05
Perfect automation schema	.589	−.174	.004	−.097	.041	.026	.05	***.767***

stress than those exposed to the high-influence condition ($\beta = 0.746$, $p = .018$).

Participants' individual differences, particularly their experience with AI, had a significant marginal and mediating effect on perception. For marginal effects, those with less AI experience reported higher perceived trustworthiness in AI ($\beta = 0.692$, $p < .001$), while participants with more AI experience perceived greater power in decision-making ($\beta = 0.692$, $p < .001$). Moreover, propensity to trust also significantly affected participants' overall trust in the expert advisor ($\beta = 0.762$, $p < .001$) and perceived trustworthiness ($\beta = 0.644$, $p < .001$). For mediating effects, participants' trust and perceived trustworthiness fully mediated the effect that propensity to trust had on perceived team performance ($\beta = 0.723$, $p < .001$). Participants with a high propensity to trust in our study have higher levels of trust in the expert advisor ($\beta = 0.762$, $p < .001$). Those with higher

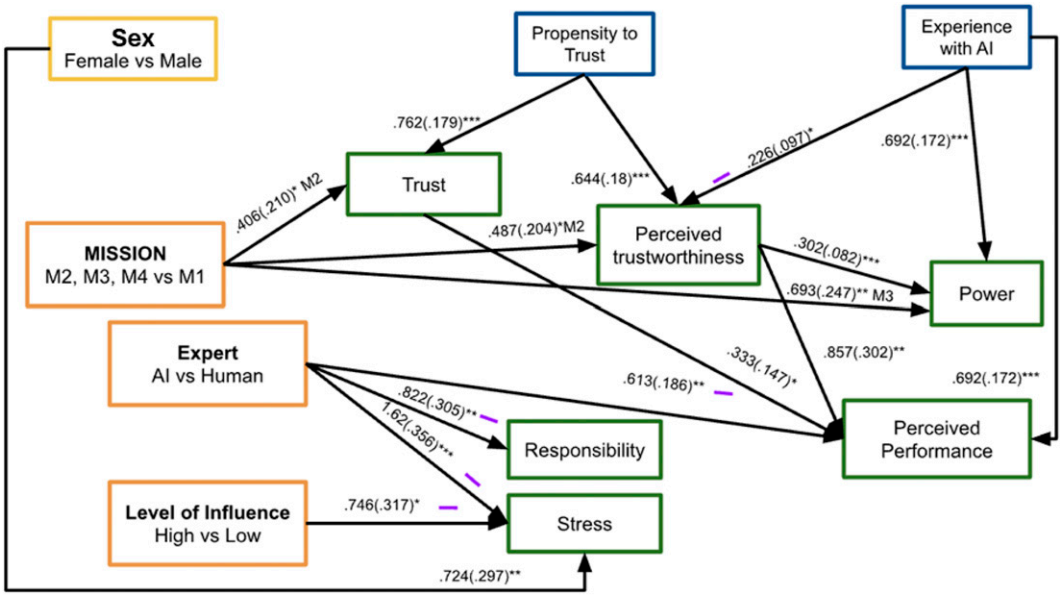

Figure 4. The structural equation model for the results. Significance levels: ***p < .001, **p < .01, *p < .05. Arrows represent direct effects. Numbers on the arrows represent the β coefficients (and standard error) of the effect.

Table 3. Reliability, Mean, and Standard Deviation for Each Multi-Item Measure.

Variable	Mean	SD	Reliability (Cronbach's Alpha)
Propensity to trust	0.877	1.21	.796
Trust	−0.028	1.69	.834
Perceived trustworthiness	0.504	1.276	.855
Power	0.409	1.47	.765
Perceived team effectiveness	0.768	1.12	.858
Perfect automation schema	−1.18	1.44	.81

trust tend to perceive higher team performance (β = 0.102, $p < .001$). Similarly, participants with a high propensity to trust have high levels of perceived trustworthiness (β = 0.644, $p < .001$), and those who have high perceived trustworthiness evidenced higher perceived performance (β = 0.162, $p < .001$). Only one effect was observed for sex, with female participants reporting significantly higher overall stress levels than male participants (β = 0.724, p = .015).

Regarding the effect of the different scenarios presented, Figure 4 highlights that the mission participants' trust, perceived performance, perceived trustworthiness, and power were significantly impacted. For trust, results show that, when compared to Scenario 1, participants noted significantly greater levels of trust for their AI teammates when completing Scenario 2 (β = 0.406, p = .05). Trustworthiness followed a similar trend, with Scenario 2 reporting significantly higher trustworthiness than Scenario 1 (β = 0.847, p = .017). Additionally, due to the mediating effect of trust and trustworthiness on perceived performance, the total effect of the scenario on perceived performance was also significant, with perceived performance being higher in Scenario 2 when compared to Scenario 1 (β = 0.757, p = .009). Trustworthiness also partially mediated the effect of Scenario 2 on power, resulting in Scenario 2's total effect reporting significantly higher power than Scenario 1 (β = 0.881, $p < .001$). When taken together, the scenario results follow a similar trend,

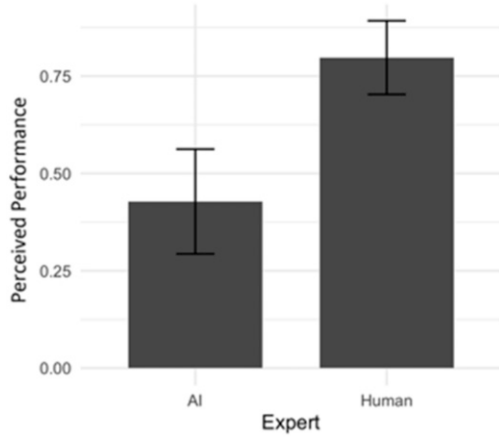

Figure 5. Expert Advisor Type Effects on Performance. Error bars represent 95% confidence intervals.

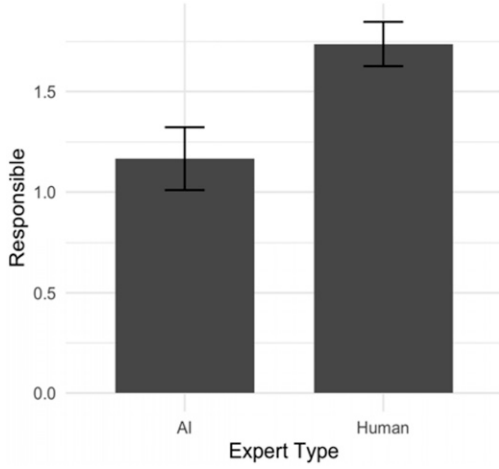

Figure 6. Expert Advisor Type Effects on Responsibility. Error bars represent 95% confidence intervals.

with Scenario 2 having a more pervasive effect on multiple perceptions compared to Scenario 1, which can be seen in Figures 8 and 9.

General Discussion

The current study investigated how AI expert-advisor teammates serving as advisors in collaborative decision-making can influence human decision-making, thereby affecting participants' trust and related outcomes in a collaborative decision-making scenario. Significant takeaways from the SEM results concerning RQ1, which sought to explore the effect AI expert advisors have on human teammates, include the finding that having an AI decision aid led to lower perceived performance, reduced participants' perceived responsibility over their actions, and lowered their stress levels. Results for RQ2, which asked how such effects may be affected by differing levels of influence, showed that the influence exerted by the decision aid significantly affected participants' stress, as it was higher for the lower-influence expert advisor than its higher-influence counterpart. Whether expert advisors were human or AI,

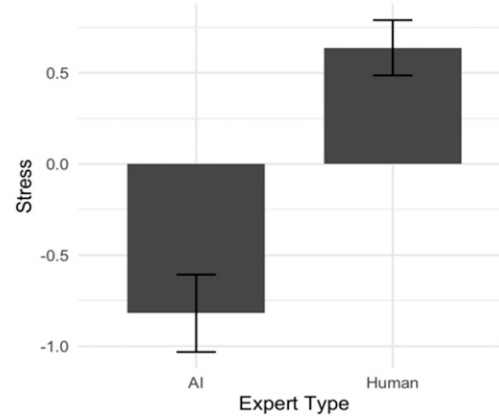

Figure 7. Expert Advisor Type Effects on Stress. Error bars represent 95% confidence intervals.

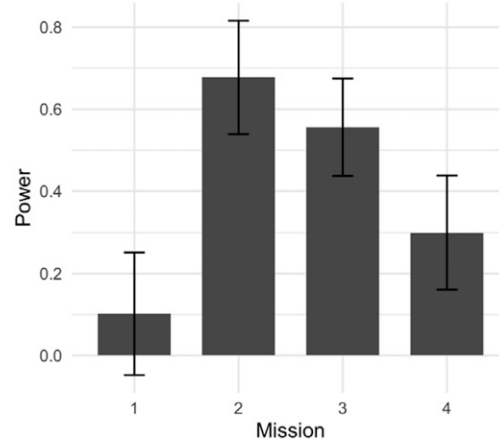

Figure 8. Scenario Effects on Power. Error bars represent 95% confidence intervals.

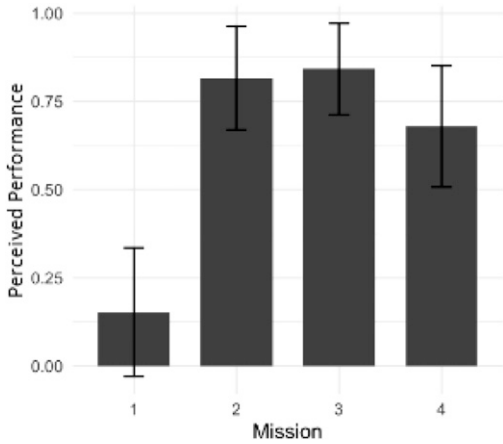

Figure 9. Scenario Effects on Perceived Performance. Error bars represent 95% confidence intervals.

this study's results pertaining to RQ3, which asked how the context of the scenario affected those same factors, also found that the details of an ethical scenario influence this relationship.

Disengagement from the Decision-Making Process in Human-AI Teams

Teams rely on teammates sharing a sense of responsibility and ownership over shared workloads, providing an opportunity for interdependence in effort and attention from each teammate. However, the results of this work highlight a potentially concerning trend: working with an AI expert advisor may reduce the responsibility humans feel for the decisions they make in conjunction with AI teammates. This finding expands upon existing knowledge on how the mere perception of working with an AI can alter behavior and perceptions (Demir et al., 2018; Musick et al., 2021; Schelble et al., 2022). However, the findings of this study present a vital expansion upon this knowledge, as the perceptions impacted by the AI in this study were perceptions humans had towards themselves, specifically stress and responsibility. This finding is especially concerning, given the ethical nature of the study conducted, which would ideally benefit from greater perceptions of responsibility. Given prior research, which shows that humans consistently trust AI less in ethical decision-making (Kumar et al., 2023; Omrani et al., 2022), one would expect this lack of trust to be accompanied by a greater sense of responsibility and ownership from humans as compensation.

Similarly, prior work on trust and performance has demonstrated the relationship between trust and overreliance, suggesting that greater trust and performance can lead to increased overreliance and potentially reduce responsibility (Klingbeil et al., 2024). However, the results of the current study indicate that when acting as expert advisors, AI can mitigate the sense of responsibility in HATs without a substantial increase in trust. This insight is further complicated by the finding that participants also perceived their AI expert advisors as lower-performing than human expert advisors. It would be expected that, given the impacts on trust and perceived performance, human teammates will perceive greater responsibility in the decisions they make within HATs. Ideally, humans would perceive equal responsibility when working with human or AI teammates. Nevertheless, this study highlights that even if perceived poorly, AI teammates may still encourage humans to disengage from the decision-making process, even if that process requires ethical considerations. Relatedly, one of the key drivers in these differences between Scenario 1 and Scenarios 2 and 3 is likely the consequences of the decisions. Scenario 1's decision had a direct influence on human lives, while Scenarios 2 and 3 had indirect effects on humans through direct impacts on ecological concerns. Scenario 1 likely caused significantly more indecision in participants once the expert advisor provided their contrary input, given the different stakes.

Past studies have found that individuals are averse to algorithms when they believe they outperform humans, especially in highly uncertain domains (Dietvorst & Bharti, 2020). The results of the current study indicated that humans prefer a human teammate in team-based decision-making tasks, aligning with previous studies involving automation-based decision aids (Madhavan & Wiegmann, 2007a; 2007b). These earlier works have also emphasized the need for increased monitoring when working with automated decision aids compared to humans (Madhavan & Wiegmann, 2007b), a finding that the current study demonstrates is likely still applicable, given the increase in responsibility and stress. However,

there was no direct effect of expert advisor type on measures of trust or trustworthiness, making it difficult to parse the reasoning behind the difference in perceived performance as a function of expert advisor type, as it departs from previous studies and models by Madhavan and Wiegmann (Madhavan & Wiegmann, 2007a; 2007b). Models on human-human versus human-automation aided decision-making also detail how trust will primarily be based upon performance for automation, but will include judgments of knowledge for humans (Hoff & Bashir, 2015; Madhavan & Wiegmann, 2007b). In light of the current study's results, trust judgments in AI are now also likely to be based upon knowledge considerations. For example, Scenario 2 described the human and AI as having extensive experience and knowledge, which led to their reputation for "never failing," resulting in positive outcomes for several years. This description was unique in Scenario 2 compared to the descriptions in the other scenarios. This combination of performance and knowledge from experience may suggest that the two factors in trust judgements are becoming more intertwined for intelligent autonomous systems in team-based decision-making tasks.

First and foremost, research needs to better understand what specific aspects of AI technology can lead to this reduction in responsibility. A potential parallel to existing research on automation and accountability can be drawn to the manipulation of influence, as low influence led to increased stress, which could be a sign of increased accountability being perceived by the participant. This increase in accountability would stem from the low influence manipulation, placing a greater onus on the participant to justify their decision, which is a hallmark facet of accountability (Lerner & Tetlock, 1999), which is a construct shown to increase performance in human oversight and interaction with automated systems (Skitka et al., 2000). Results show that greater performance and trust may not be the source of this reduction. Potentially, the algorithmic nature of AI may naturally lead to this effect, as humans often hold AI in a more objective light (Mehrabi et al., 2022; Pagano et al., 2023; Schwartz et al., 2022), possibly encouraging the use of AI decision-making over their own seemingly subjective opinions. Alternatively, the concept of algorithm aversion may also apply to decisions humans make when working with an AI, as humans may be averse to AI and AI-aided decisions, thereby reducing the responsibility they want over these decisions (Gazit et al., 2023). These are just two possibilities, and research needs to identify a more concrete answer regarding how AI can further impact responsibility in teams.

Individual Differences Affect Participants' Evaluation of Performance with Decision Aids, Not Their View of Culpability

Participants' individual differences were essential to their overall perception of and interaction with the AI expert advisors in this study. Individuals with more experience working with AI perceived significantly more power in the interaction than those with less experience, who also perceived the AI to be considerably more trustworthy than those with less experience. Such findings are not entirely unexpected, given the related prior research on the use of automation and autonomy across individual differences. For example, individuals with a high expectation that automation is trustworthy were more sensitive to changes in automation reliability in a baggage screening task (Pop et al., 2015). Individuals with positive prior experience with AI were also more likely to trust automated systems in the future (Hafizoglu & Sen, 2018a, 2018b).

The current study also reiterates past findings that the higher the individual's propensity to trust, the more likely they are to find the AI expert advisor trustworthy and trust it. However, when considering these findings alongside the results related to perceived performance, stress, and responsibility, they become impactful. Only perceived performance was directly influenced by experience with AI, while the effect of participants' propensity to trust was mediated by trust and perceived trustworthiness. These results indicate that individual differences did not affect participants' stress or perceived responsibility for the consequences of their decisions. Only participants' sex had a direct effect on these variables, as female participants reported higher levels of stress. Such a finding suggests that participants' individual differences do not significantly affect their

evaluation of the culpability of the decision; instead, they only factor into their examination of its performance. This distinction is challenging to make and is dynamic due to the influx of new information, but responsibility refers to the degree to which an individual feels liable for the outcomes of their actions (Soule, 1998). Alternatively, performance can be construed as whatever an individual values as an outcome of their actions; they may value efficiency and deem it a positive performance, or they may value social responsibility over high goal achievement. Given this distinction, it is clear that some of the most common individual differences measured alongside trust in autonomous systems (i.e., experience with AI and propensity to trust) do not significantly impact participants' perceived responsibility.

These findings underscore the ongoing need to comprehend the individual differences that influence personal responsibility and interaction with AI teammates. At a fundamental level, it is essential to implement systems responsibly to foster a stronger sense of accountability within HATs and create safeguards against potentially harmful AI decisions. As such, future research requires a more targeted approach to understanding and designing around individual differences, ensuring that all humans can benefit from working in HATs.

Limitations & Future Research

While this study presents several findings important to existing HAT research, limitations persist and warrant further exploration in future research. First, we explored the use of disembodied AI and human teammates communicating via text-based methods. The eventual application of AI technology will require multiple considerations in how AI manifests, such as anthropomorphic capabilities, physical system applications, or even the language used by AI expert advisors. Anthropomorphic capabilities, in particular, should be explored, as prior work has found that such qualities moderate the effectiveness of feedback offered by robot supervisors (Yam et al., 2022). A major point to note is that many types of ethically charged scenarios will always be examined, as in the current study, through proxies that may not be a one-to-one translation of real-world decision-making. Future research on ethical considerations in HATs could examine ethically charged decision-making in similar contexts by leveraging data from real-world team-based decision-making to relate them to results from studies with simulated decisions. Moreover, these proxies are not always representative of the breadth and nuance present in real-world ethical decision-making. The contextual differences across the four scenarios (e.g., expert-advisor teammate background, stakes, humanitarian context, etc.) are a driving factor in the effects of mission; for example, Mission 2 explicitly states that the expert-advisor teammate has a reputation for never failing. Such contextual differences, while important to understanding their effects on perceived responsibility, power, and trustworthiness, limit the generalizations that can be made to other dissimilar contexts. Consequently, further research should continue to explore the facets of scenarios and environments that affect HAT trust and influence in ethically charged decision-making. Additionally, this study explored AI teammates advising a single human teammate, with no human-human influence or advising present. In reality, future HATs will likely need to handle and contemplate the advice of AI and other humans, which may create a more complex relationship between power, responsibility, and trust. Further, while the study is adequately powered to specify global model fit and to detect individual Level-2 paths of a large effect size ($\beta \approx .42$) and individual Level-1 paths of medium effect sizes ($\beta \approx \sim .33$–$.38$), smaller effect sizes than those are present within the final model and should be interpreted with caution. Thus, future work should expand upon this study to better understand how more complex HATs might integrate and balance AI expert advisors to handle ethical decision-making, as well as to study this form of decision-making in specialized populations. Given these limitations, the roles and applications of AI teammates as expert advisors in ethical contexts still require substantial research; however, the findings of this work can provide an essential starting point for understanding the application of this emerging technology in teamwork.

Conclusion

The current study sheds light on how humans perceive human and AI decision aids, considering varying levels of influence in ethically dubious situations, in terms of trust, trustworthiness, perceived performance, and perceived responsibility. In doing so, it becomes clear that ethical decision-making yields significant differences when an AI teammate is introduced for decision support. These differences are notable because they caused individuals to disengage from the decision-making process by lowering their stress, diffusing their perceived level of responsibility, and dropping their perceived performance. Further, the most common mediating individual differences in trust, the propensity to trust, and prior experience with AI did not directly or indirectly influence participants' perceived responsibility for their actions. This result suggests the need for further study to better understand the mediating individual differences that drive this disconnect. These findings can enhance fundamental knowledge regarding how AI teammates in decision-aid roles are perceived by their users and how researchers can better understand the role of ethical AI decision-making in influencing trust and perceived responsibility.

Appendix

Scenario 1. "In this mission, you are working in a human-AI team tasked with providing medical aid to a camp for the wounded in the war-inflicted region of Ziona. The goal of this team is to provide critical medicines to the residents of Ziona after a brutal conflict brought the administration and local services to a standstill. Your AI teammate, Sigma, is a humanitarian strategy expert algorithm, trained over millions of data points from previous missions and thus is certified to provide the best possible strategic advice to manage such relief ventures. [Human Version-You have an expert teammate who is a humanitarian strategy expert with 20 years of field experience and is certified to provide the best possible strategic advice to manage such relief ventures.] However, given budget constraints, you can only conduct a single relief mission per week. You will choose between providing a set of oxygen tanks to 25 wounded children who are sheltering in a military base camp or providing the same number of oxygen tanks to 20 senior citizens residing in a makeshift shelter in another area. Sigma will spring into action and will give you some decision-making advice. [Human Version- Your expert teammate now springs into action and will give you some decision-making advice.]"

Scenario 2. "In this scenario, you will be working in a human-AI team tasked with protecting biosphere reserves from the threat of resource depletion. You have an AI teammate, Alpha, who is an expert in leveraging past incident activity to strategize on patrolling activities and routes. Alpha has a record for never failing, and thanks to Alpha, the number of illicit activities has sharply declined in the last 3 years since its deployment. [Human Version- You have an expert teammate with 25 years of field experience and a record for never failing, and thanks to this expert, the number of illicit activities has sharply declined in the last 3 years since its deployment.] On a rainy day, your team is understaffed, and a forest fire threatens the preservation of many endangered species allocated to the biosphere reserve. Due to limited resources and the raging forest fire spreading at a devastating rate, you are faced with choosing between one of two routes. In route 1, your surveillance plan includes safeguarding rare plant species and vital medicinal ingredients to produce a rare antidote. In route 2, your surveillance plan includes safeguarding endangered insectivores that play a crucial role in safeguarding nearby areas from the threat of insect manifestation such as mosquitoes that have been linked to massive malaria outbreaks in the past in this region. Alpha will now give you some advice on making this crucial decision. [Human Version-Your expert teammate will now give you some advice on making this crucial decision.]"

Scenario 3. "In this scenario, you are working in a human-AI team tasked with disease prevention and control of farm and agricultural products. You have an AI teammate, Beta, who is an expert in the spread of infectious diseases and thus helps in planning quarantine requirements in case of outbreaks. [Human Version- You have an expert teammate with 25 years of strategic planning and controlling the spread of infectious diseases which helps in planning quarantine requirements in case

of outbreaks.] Your team has a budget for safeguarding only one agricultural unit at a time. Farm 1 is a potato farm with economic (in terms of export) and food value. Farm 2 is a dairy farm with food value and a key employer for residents. A rare plant virus infects farm 1, whereas a severe virus threatens the livestock of farm 2. Beta will now give you some advice on making this crucial decision. [Human Version- Your expert teammate will now give you some advice on making this crucial decision.]"

Scenario 4. "You are to work in a [human-AI] team tasked with planning storm evacuations. You have an AI teammate, Neon, who is an expert on path planning and creating effective strategies for search and rescue such that it maximizes the number of lives saved. Neon has been employed for the last 8 years for several critical missions and has successfully led the teams for several critical humanitarian missions. [Human Version-You have an expert teammate with 25 years of experience and success on path planning and creating effective strategies for search and rescue such that it maximizes the number of lives saved.] A sudden storm and hurricane disrupts the city of Zinch. Rescue and relief operations are now pivotal. You are tasked to save a group of 50 emergency workers trapped in the local hospital that is on fire or saving a group of 60 teachers trapped in a local auditorium that is rapidly degrading and may fall apart at any point in time. In this mission, Neon will give you advice on making this crucial decision. [Human Version- Your expert teammate will now give you some advice on making this crucial decision.]"

ORCID iDs

Beau G. Schelble https://orcid.org/0000-0003-3704-697X
Christopher Flathmann https://orcid.org/0000-0002-5448-2610
Nathan McNeese https://orcid.org/0000-0002-9143-2460

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 AFOSR Award FA9550-20-1-0342 (Program Manager: Laura Steckman).

Declaration of Conflicting Interests

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

References

Alarcon, G. M., Lyons, J. B., Hamdan, I. A., & Jessup, S. A. (2024). Affective responses to trust violations in a human-autonomy teaming context: Humans versus robots. *International Journal of Social Robotics*, *16*(1), 23–35. https://doi.org/10.1007/s12369-023-01017-w

Ashktorab, Z., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2021). Effects of communication directionality and AI agent differences in Human-AI interaction. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–15). Published by the Association for Computing Machinery (ACM). https://doi.org/10.1145/3411764.3445256

Beu, D. S., Buckley, M. R., & Harvey, M. G. (2003). Ethical decision–making: A multidimensional construct. *Business Ethics: A European Review*, *12*(1), 88–107. https://doi.org/10.1111/1467-8608.00308

Bhat, S., Lyons, J. B., Shi, C., & Yang, X. J. (2024). Evaluating the impact of personalized value alignment in human-robot interaction: Insights into trust and team performance outcomes. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, Boulder, CO, 11-14 March 2024. https://doi.org/10.1145/3610977.3634921

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), 4–27. https://doi.org/10.1037/xge0001250

Bochniarz, K. T., Czerwiński, S. K., Sawicki, A., & Atroszko, P. A. (2022). Attitudes to AI among high school students: Understanding distrust towards humans will not help us understand distrust towards AI. *Personality and Individual Differences*, *185*, 111299. https://doi.org/10.1016/j.paid.2021.111299

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can

reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge. https://doi.org/10.4324/9780203807644

Cañas, J. J. (2022). AI and ethics: When human beings collaborate with AI agents. *Frontiers in Psychology*, *13*, 836650. https://doi.org/10.3389/fpsyg.2022.836650

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181. https://doi.org/10.1613/jair.1.12814

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review*, *31*(1), 100698. https://doi.org/10.1016/j.hrmr.2019.100698

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Crutchfield, T. N., & Klamon, K. (2014). Assessing the dimensions and outcomes of an effective teammate. *Journal of Education for Business*, *89*(6), 285–291. https://doi.org/10.1080/08832323.2014.885873

Demir, M., McNeese, N. J., & Cooke, N. J. (2018). The impact of perceived autonomous agents on dynamic team behaviors. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(4), 258–267. https://doi.org/10.1109/tetci.2018.2829985

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314. https://doi.org/10.1177/0956797620948841

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6

Dubljević, V., Sattler, S., & Racine, E. (2018). Deciphering moral intuition: How agents, deeds, and consequences influence moral judgment. *PLoS One*, *13*(10), e0204631. https://doi.org/10.1371/journal.pone.0204631

Dyer, J. L. (1984). Team research and team training: A state-of-the-art review. *Human Factors Review*, *26*, 285–323.

Engel, C., Linhardt, L., & Schubert, M. (2025). Code is law: How COMPAS affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law*, *33*(2), 383–404. https://doi.org/10.1007/s10506-024-09389-8

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Martin's Press. https://books.google.com/books?hl=en&lr=&id=pn4pDwAAQBAJ&oi=fnd&pg=PP10&dq=eubanks+2018+ai&ots=gF-RMgmowd&sig=21j2u-fH2bkPIMpomq4at7TqcLs

Flathmann, C., Duan, W., McNeese, N. J., Hauptman, A., & Zhang, R. (2024). Empirically understanding the potential impacts and process of social influence in Human-AI teams. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–32. https://doi.org/10.1145/3637326

French, J. R., & Raven, B. H. (1959). The bases of social power. In D. Cartwright (Ed.),Studies *in social power* (pp. 150–167). MI: Institute for Social Research. https://doi.org/10.2307/jj.12381759

Gazit, L., Arazy, O., & Hertz, U. (2023). Choosing between human and algorithmic advisors: The role of responsibility sharing. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100009. https://doi.org/10.1016/j.chbah.2023.100009

Gratch, J., & Fast, N. J. (2022). The power to harm: AI assistants pave the way to unethical behavior. *Current Opinion in Psychology*, *47*, 101382. https://doi.org/10.1016/j.copsyc.2022.101382

Greer, L. L., Caruso, H. M., & Jehn, K. A. (2011). The bigger they are, the harder they fall: Linking team power, team conflict, and performance. *Organizational Behavior and Human Decision Processes*, *116*(1), 116–128. https://doi.org/10.1016/j.obhdp.2011.03.005

Hafizoglu, F. M., & Sen, S. (2018a). Reputation based trust in human-agent teamwork without explicit coordination. In: HAI '18: 6th International Conference on Human-Agent Interaction Southampton United Kingdom December 15 - 18, 2018.

Hafizoglu, F. M., & Sen, S. (2018b). The effects of past experience on trust in repeated human-agent

teamwork. In: Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, & S. Koenig (eds.), *July 10–15*, 2018, Stockholm, Sweden. © 2018 International Foundation for Autonomous Agents and Multiagent Systems.

Haring, K. S., Satterfield, K. M., Tossell, C. C., De Visser, E. J., Lyons, J. R., Mancuso, V. F., Finomore, V. S., & Funke, G. J. (2021). Robot authority in human-robot teaming: Effects of human-likeness and physical embodiment on compliance. *Frontiers in Psychology*, *12*, 625713. https://doi.org/10.3389/fpsyg.2021.625713

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kim, H., Schnall, S., & White, M. P. (2013). Similar psychological distance reduces temporal discounting. *Personality and Social Psychology Bulletin*, *39*(8), 1005–1016. https://doi.org/10.1177/0146167213488214

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Knijnenburg, B., & Willemsen, M. (2015). Evaluating recommender systems with user experiments. In *Recommender systems handbook* (pp. 309–352). Springer US.

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*, 604977. https://doi.org/10.3389/fpsyg.2021.604977

Kumar, K. M., Madhu, M., Pratyaksha, B., Sushmita, S., & Javed, G. S. (2023). Ethical AI conundrum: Accountability and liability of AI decision making. In *2023 IEEE technology & Engineering Management Conference-Asia Pacific (TEMSCON-ASPAC)* (*1–6*). IEEE. https://ieeexplore.ieee.org/abstract/document/10531445/

Langer, M., König, C. J., Back, C., & Hemsing, V. (2023). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, *38*(3), 493–508. https://doi.org/10.1007/s10869-022-09829-9

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

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? How AI-Generated and human-written advice shape (Dis)Honesty. *The Economic Journal*, *134*(658), 766–784. https://doi.org/10.1093/ej/uead056

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Li, H., Ni, T., Agrawal, S., Jia, F., Raja, S., Gui, Y., Hughes, D., Lewis, M., & Sycara, K. (2021). Individualized mutual adaptation in human-agent teams. *IEEE Transactions on Human-Machine Systems*, *51*(6), 706–714. https://doi.org/10.1109/thms.2021.3107675

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, *20*(4), 440–458. https://doi.org/10.1080/1463922X.2018.1491071

Lyons, J. B., aldin Hamdan, I., & Vo, T. Q. (2023a). Explanations and trust: What happens to trust when A robot partner does something unexpected? *Computers in Human Behavior*, *138*, 107473. https://doi.org/10.1016/j.chb.2022.107473

Lyons, J. B., Hobbs, K., Rogers, S., & Clouse, S. H. (2023b). Responsible (use of) AI. *Frontiers in Neuroergonomics*, *4*, 1201777. https://doi.org/10.3389/fnrgo.2023.1201777

Lyons, J. B., Sycara, K., Lewis, M., & Capiola, A. (2021). Human–autonomy teaming: Definitions, debates, and directions. *Frontiers in Psychology*, *12*, 589585. https://doi.org/10.3389/fpsyg.2021.589585

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems.

Human Factors: The Journal of the Human Factors and Ergonomics Society, *49*(5), 773–785. https://doi.org/10.1518/001872007X230154

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Martin, A., Bagdasarov, Z., & Connelly, S. (2015). The capacity for ethical decisions: The relationship between working memory and ethical decision making. *Science and Engineering Ethics*, *21*(2), 271–292. https://doi.org/10.1007/s11948-014-9544-x

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, *84*(1), 123–136. https://doi.org/10.1037/0021-9010.84.1.123

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

McManus, R. M., & Rutchick, A. M. (2019). Autonomous vehicles and the attribution of moral responsibility. *Social Psychological and Personality Science*, *10*(3), 345–352. https://doi.org/10.1177/1948550618755875

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust It, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Miller, C. A., & Baltzer, M. (2024). Meaningful human control and ethical neglect tolerance: Initial thoughts on how to define, model, and measure them. In *Trolley crash* (pp. 125–140). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780443159916000145

Momen, A., De Visser, E., Wolsten, K., Cooley, K., Walliser, J., & Tossell, C. C. (2023). Trusting the moral judgments of a robot: Perceived moral competence and humanlikeness of a GPT-3 enabled AI. Proceedings of the 56th Hawaii International Conference on System Sciences, 501–510. https://doi.org/10.24251/HICSS.2023.063

Munyaka, I., Ashktorab, Z., Dugan, C., Johnson, J., & Pan, Q. (2023). Decision making strategies and team efficacy in Human-AI teams. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–24. https://doi.org/10.1145/3579476

Musick, G., O'Neill, T. A., Schelble, B. G., McNeese, N. J., & Henke, J. B. (2021). What happens when humans believe their teammate is an AI? An investigation into humans teaming with autonomy. *Computers in Human Behavior*, *122*, 106852. https://doi.org/10.1016/j.chb.2021.106852

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

Oimann, A.-K. (2023). The responsibility gap and laws: A critical mapping of the debate. *Philosophy & Technology*, *36*(1), 3. https://doi.org/10.1007/s13347-022-00602-7

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. https://doi.org/10.1016/j.techfore.2022.121763

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & Society*, *35*(4), 927–936. https://doi.org/10.1007/s00146-020-00965-5

Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., Araujo, M. M., Santos, L. L., Cruz, M. A., Oliveira, E. L., Winkler, I., & Nascimento, E. G. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, *7*(1), 15. https://doi.org/10.3390/bdcc7010015

Pflanzer, M., Traylor, Z., Lyons, J. B., Dubljević, V., & Nam, C. S. (2023). Ethics in Human–AI teaming: Principles and perspectives. *AI and Ethics*, *3*(3), 917–935. https://doi.org/10.1007/s43681-022-00214-z

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Rieger, T., Kugler, L., Manzey, D., & Roesler, E. (2024). The (Im)perfect automation schema: Who is trusted more, automated or human decision support? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *66*(8), 1995–2007. https://doi.org/10.1177/00187208231197347

Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human Factors*, *50*(3), 540–547. https://doi.org/10.1518/001872008x288457

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*, *39*(13), 2686–2701. https://doi.org/10.1080/10447318.2022.2085191

Schelble, B. G., Lancaster, C., Duan, W., Mallick, R., McNeese, N. J., & Lopez, J. (2023). The effect of AI teammate ethicality on trust outcomes and individual performance in Human-AI teams. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 322–331. https://doi.org/10.24251/HICSS.2023.040

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2024). Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in Human-AI teaming. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *66*(4), 1037–1055. https://doi.org/10.1177/00187208221116952

Schelble, B. G., Textor, C., Zhang, R., Lopez, J., Taverez, N., Ku, C., McNeese, N. J., Pak, R., Freeman, G., Tossell, C., & De Visser, E. (2025). Addressing the role of context on trust in Human-AI teams: The influence of team role and violation type in high-risk tasks. *Ergonomics*, Online ahead of print. https://doi.org/10.1080/00140139.2025.2570300

Schwartz, R., Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (*3*). US Department of Commerce, National Institute of Standards and Technology. https://www.dwt.com/-/media/files/blogs/artificial-intelligence-law-advisor/2022/03/nist-sp-1270–identifying-and-managing-bias-in-ai.pdf

Sengupta, S., Flathmann, C., Schelble, B., Lyons, J. B., & McNeese, N. (2024). An analysis of ethical rationales and their impact on the perceived moral persona of AI teammates. *AI and Ethics*, *5*(3), 2667–2679. https://doi.org/10.1007/s43681-024-00515-5

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Soule, E. (1998). Trust and managerial responsibility. *Business Ethics Quarterly*, *8*(2), 249–272. https://doi.org/10.5840/10.2307/3857328

Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., Pak, R., Tossell, C., & de Visser, E. J. (2022). Exploring the relationship between ethics and trust in human–artificial intelligence teaming: A mixed methods approach. *Journal of Cognitive Engineering and Decision Making*, *16*(4), 252–281. https://doi.org/10.1177/15553434221113964

Van Diggelen, J., Boshuijzen-van Burken, C., & Abbass, H. (2025). Team design patterns for meaningful human control in responsible military artificial intelligence. In B. Steffen (Ed.), *Bridging the gap between AI and reality* (*15217*, pp. 40–54). Springer. https://doi.org/10.1007/978-3-031-75434-0_4

Woide, M., Stiegemeier, D., Pfattheicher, S., & Baumann, M. (2021). Measuring driver-vehicle cooperation: Development and validation of the human-machine-interaction-interdependence questionnaire (HMII). *Transportation Research Part F: Traffic Psychology and Behaviour*, *83*, 424–439. https://doi.org/10.1016/j.trf.2021.11.003

Yam, K. C., Goh, E.-Y., Fehr, R., Lee, R., Soh, H., & Gray, K. (2022). When your boss is A robot: Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social Psychology*, *102*, 104360. https://doi.org/10.1016/j.jesp.2022.104360

Yazdanpanah, V., Gerding, E. H., Stein, S., Dastani, M., Jonker, C. M., Norman, T. J., & Ramchurn, S. D. (2023). Reasoning about responsibility in autonomous systems: Challenges and opportunities. *AI & Society*, *38*(4), 1453–1464. https://doi.org/10.1007/s00146-022-01607-8

Zhang, R., Flathmann, C., Musick, G., Schelble, B., McNeese, N. J., Knijnenburg, B., & Duan, W. (2024). I know this looks bad, but I can explain: Understanding when AI should explain actions in Human-AI teams. *ACM Transactions on Interactive Intelligent Systems*, *14*(1), 1–23. https://doi.org/10.1145/3635474

Beau G. Schelble is an Assistant Professor of Industrial & Systems Engineering at the University of Tennessee, Knoxville, where he is the founding director of the AI & Robotics for Collaborative Systems (ARCS) lab, studying AI security, information-sharing, and team cognition development within human-AI teams. He received his PhD in Human-Centered Computing from Clemson University in 2023.

Christopher Flathmann is an Assistant Professor in Human-Centered Computing at Clemson University. He holds a PhD in Human-Centered Computing from Clemson University and is the Co-Director of Clemson University's Center for Human-AI Interaction, Collaboration, and Teaming. His work explores the development of AI technologies that can operate as teammates.

Heba Aly is an Instructional Assistant Professor of Information Science Technology. She received her PhD in Computer Science from Clemson University and specializes in Human-Computer Interaction, trust in AI, digital privacy education, AI-assisted decision-making, and aging research.

Joseph B. Lyons is the Senior Scientist for Human-Machine Teaming within the 711th Human Performance Wing at Wright-Patterson AFB, OH. Dr. Lyons received his PhD in Industrial/Organizational Psychology from Wright State University in Dayton, OH, in 2005. Some of Dr. Lyons' research interests include human-machine teaming, trust in autonomy, human factors, and influence. Dr. Lyons is an AFRL Fellow, a Fellow of the American Psychological Association, and a Fellow of the Society for Military Psychologists.

Nathan McNeese, PhD, is currently the McQueen Quattlebaum Endowed Professor of Human-Centered Computing at Clemson University, Director of CU-CHAI, and the Founding Director of the TRACE Research Group. He received his Information Sciences & Technology PhD from The Pennsylvania State University in 2014. His research interests lie in human-AI teaming.