I Know This Looks Bad, But I Can Explain: Understanding When AI Should Explain Actions In Human-AI Teams

RUI ZHANG, CHRISTOPHER FLATHMANN, GEOFF MUSICK, BEAU SCHELBLE, NATHAN J. MCNEESE, BART KNIJNENBURG, and WEN DUAN, Clemson University, USA

Explanation of artificial intelligence (AI) decision-making has become an important research area in human–computer interaction (HCI) and computer-supported teamwork research. While plenty of research has investigated AI explanations with an intent to improve AI transparency and human trust in AI, how AI explanations function in teaming environments remains unclear. Given that a major benefit of AI giving explanations is to increase human trust understanding how AI explanations impact human trust is crucial to effective human-AI teamwork. An online experiment was conducted with 156 participants to explore this question by examining how a teammate's explanations impact the perceived trust of the teammate and the effectiveness of the team and how these impacts vary based on whether the teammate is a human or an AI. This study shows that explanations facilitate trust in AI teammates when explaining why AI disobeyed humans' orders but hindered trust when explaining why an AI lied to humans. In addition, participants' personal characteristics (e.g., their gender and the individual's ethical framework) impacted their perceptions of AI teammates both directly and indirectly in different scenarios. Our study contributes to interactive intelligent systems and HCI by shedding light on how an AI teammate's actions and corresponding explanations are perceived by humans while identifying factors that impact trust and perceived effectiveness. This work provides an initial understanding of AI explanations in human-AI teams, which can be used for future research to build upon in exploring AI explanation implementation in collaborative environments.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; • **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → *User characteristics;*

Additional Key Words and Phrases: Explanation, trust, AI teammates, human-AI teaming

ACM Reference format:
Rui Zhang, Christopher Flathmann, Geoff Musick, Beau Schelble, Nathan J. McNeese, Bart Knijnenburg, and Wen Duan. 2024. I Know This Looks Bad, But I Can Explain: Understanding When AI Should Explain Actions In Human-AI Teams. *ACM Trans. Interact. Intell. Syst.* 14, 1, Article 6 (February 2024), 23 pages.
https://doi.org/10.1145/3635474

1 INTRODUCTION

Driven by continuously advancing **artificial intelligence** (**AI**) technologies, human-AI interaction and collaboration have become common over the past decade. AI is being applied in several

Authors' address: R. Zhang, C. Flathmann G. Musick, B. Schelble, N. J. McNeese, B. Knijnenburg, and W. Duan, Clemson University, Clemson, South Carolina 29634, USA; e-mails: {rzhang2, clathm, gmusick, bschelb, mcneese, bartk, wend}@clemson.edu.

2160-6455/2024/2-ART6 $15.00
https://doi.org/10.1145/3635474

facets of daily life, such as AI teammates in games, virtual assistants, and AI-powered clinical decision support systems [77, 98, 101]. This increase in collaboration between humans and AI has generated significant research interest in the concept of human-AI teams [8, 15, 59, 109]. Specifically, a human-AI team is an integrated unit where human and AI teammates, each with a significant degree of agency, coordinate and collaborate to complete team tasks with a shared goal [68, 79]. Different from AI being used as a tool, teams where AI functions as a teammate have shown the potential to not only outperform AI systems but also expert human teams [68, 110]. However, achieving this potential requires both a focus on the technical/task-focused contributions and the human-factors contributions of AI [69, 79]. For instance, the trust humans form for teammates has been shown to differ when teammates were AI as opposed to human [86], which in turn impacts team performance [67]. Research has also explored how AI teammates can be designed to ensure these impacts are beneficial through design considerations, such as ensuring ethical behavior [88, 97] or coordination training for humans [51]. However, research in human-AI teaming has yet to explore a potentially critical avenue of AI design that could demonstrably benefit these perceptions, which is the explanation design of AI *teammates*.

Within the broader field of AI technologies, explainability is not a novel concept [56, 84]. In fact, explainability has been an extremely relevant construct to explore in recent years from a computational and social perspective [82]. For instance, computational efforts have focused heavily on removing the black-box nature commonly associated with AI in favor of algorithms that can be explored and understood by humans [63, 92]. From a social perspective, humans who interact with these explainable AI systems see demonstrable benefits to their perceptions of the said systems, especially where trust in the AI is concerned [29, 106]. Importantly, humans' trust development in AI is not solely built upon computational efforts but rather a combination of these efforts and humans' general opinions toward AI technologies, which may hinder or facilitate the creation of trust in AI systems [37]. As a result, explainable AI systems are not a simple niche for specific contexts but rather a pillar of the community and critical construct of AI [3, 45].

Despite the demonstrable benefits of explainable AI systems to trust and identified benefits of trust development to human-AI teams, empirical research that explicitly examines the impact and potential benefits of explainable AI *teammates* has just started to emerge. Existing work has explored how AI's explanations contribute to human trust and human performance positively in teams [11, 36]. A summary of existing studies on AI's explanations in *human-AI interaction* indicates that the impact of explanations on human trust is inconsistent [83]. While some studies have shown the positive impact of AI's explanations on human trust [11], other studies indicate no impact on human trust [24]. The inconsistency of these studies has resulted from various tasks in which each study explores AI's explanations and often when AI operates as a tool. Importantly, compared to traditional digital AI tools explored in previous work [54], the concept of explanation itself may need to be examined differently in human-AI teams or even between different human-AI teams, as explanations have to be specifically designed for chosen contexts and scenarios [4]. In particular, there may exist scenarios where AI teammates have to prioritize the needs of a team over a teammate, but they will need to explain this prioritization to these teammates, which may change or remove the benefits of explanation. Thus, it is important to explore how AI teammate explanation is poised to benefit teams and whether these benefits are potentially moderated by unique factors between teams like individual team member characteristics, AI actions, and the specific team context. Regardless, the above demonstrates that the exploration of AI teammate explanation is likely to provide positive benefits for human-AI teaming contexts, even if the impacts of AI's explanations look somewhat different in the unique field of human-AI teaming.

To address the above research gap, we conducted an online experiment utilizing video-based scenario vignettes with 156 participants to build a foundational understanding of how the

explanations given by a teammate impacts trust and how these impacts might differ depending on the action performed by the teammate providing the explanation and whether this teammate is a human or an AI. Using a multiplayer video game, we presented participants with a video recording of a human or AI teammate performing four actions that were either explained or unexplained, depending on the experimental condition. **Structural equation modeling** (**SEM**) was utilized to examine the impact of explanations on participants' perceptions (trust and perceived team effectiveness), depending on the action performed by the teammate, the teammate's identity (human or AI), and the participant's personal characteristics. Through this experiment and analysis, we targeted the following research questions:

RQ1 How does the utilization of explanations by AI teammates impact trust?
RQ1.1: How does the identity of a teammate (human or AI) influence the explanation's impact on trust?
RQ1.2: Does the action performed by a teammate change the explanation's impact on trust?
RQ1.3: How do humans' personal characteristics impact their perception of the team's effectiveness?
RQ1.4: Do the effects of the teammate's explanation on trust extend to other teaming perceptions?

This work makes several contributions to the fields of interactive intelligent systems and HCI. First, our study extends the current research on human-AI teams by demonstrating how AI providing an explanation of their action can impact trust in human-AI teams. The findings of this work help improve AI researchers' ability to implement explanations in collaborative intelligent systems. Second, our focus on a teammate's identity points out the similarities and differences in how people perceive and interpret a human vs. AI teammate's actions and subsequent explanations. Given that people view an AI teammate's behaviors differently from how they judge a human teammate, this work provides insights into how the implementation of explanations in human-AI teams would function differently from human–human teams. Finally, our findings inform design implications of AI teammate explanations regarding (1) when AI should explain their decision-making process and (2) how this depends on the personal characteristics of humans.

2 RELATED WORK

2.1 Human-AI Teaming

With the broad deployment of AI, a plethora of research has explored how humans can collaborate with AI in interactive contexts where AI are treated as teammates rather than tools [15, 101, 110]. When conceptualized as a teammate, AI takes on team-level responsibilities and executes decisions as a team member to contribute to shared team goals. When one or more of these AI teammates interact with one or more human teammates to work toward a shared goal interdependently, the team is known as a *human-AI team* [69, 78]. Similar to human–human teams, team performance is considered an important metric to evaluate a human-AI team's collaborative activities [55]. Previous work has studied various factors that impact team performance, such as human-AI team composition [26, 33, 68, 70], AI's performance or features [20, 34], and human perceptions of AI teammates [71, 85, 110]. Some research points to human-AI teams outperforming human-only teams due to efficiency gains from a greater level of workload management and collective intelligence[16, 33, 35]. However, human-AI teams that rely heavily on team processes, such as coordination, communication, and adaptability, often struggle to outperform human-only teams, which can be directly linked to AI's limitations in their ability to engage in those same team processes [25, 64, 75].

Importantly, the impacts of AI teammates extend beyond team performance and into perception. Previous research has indicated that AI teammates are treated differently from humans in a team setting [79]. Using a Wizard of Oz design, researchers have found that participants perceive worse team anticipation, planning, and performance when they perceive their teammate to be an AI [87], which is but one example of the critical role human perceptions of AI play in human-AI teaming. Another study involving a Wizard of Oz design also found that humans were less likely to perceive the development of shared mental models to be possible with their AI teammates, which interfered with attempts to implicitly coordinate with their AI teammate [73, 87].

2.2 Trust and AI Explanations in Human-AI Collaboration

Trust plays an essential role in teamwork as it enables team members to confidently collaborate and coordinate [52, 113]. However, trust is a complicated social construct that makes defining and applying it to technology and human-AI interaction multifaceted and challenging [50, 94]. Similar to human trust, trust in AI often includes aspects of cognitive trust (e.g., based on a rational assessment of skills and reliability), affective trust (e.g., based on feelings or emotions), or a combination of both [37, 43, 94]. Although technical factors such as improving algorithms to increase reliability are important to improving trust [19], human factors such as transparency, explainability, and AI explanations are heavily emphasized in the literature and have gained recent attention [6, 27, 50, 94].

AI providing explanations has become increasingly important in human-AI interactions where trust is a necessity, such as human-AI collaboration in healthcare and finance [2, 13, 14, 103]. For instance, AI explanation is known to increase trust calibration, which involves humans knowing when to trust an AI correctly and, importantly, when not to trust an AI [45, 50, 111]. Further, AI explanation has been shown to improve user trust and improve attitudes toward AI [92, 93]. As such, much of explainable AI research has focused on how to create human-friendly explanations, which often require a directional conversation between collaborators [1, 2, 107].

As AI explanation is emphasized as a vital factor in trust for human-AI interaction, it is understandably an area of interest in human-AI teaming [32, 96, 104]. Trust, as an essential component of a human teammate's mental model of AI, can facilitate humans' interaction with AI [7]. This process involves humans gaining confidence in the AI's capabilities through an understanding of the AI's roles, responsibilities, and decision-making process, thus improving the team mental model [31, 96]. AI explanation also improves humans' understanding of the AI teammate's uncertainty, which results in better team decision-making [112]. The importance of trust in human-AI teams has also received traction in multiplayer online game research–a domain that has received recent focus from the human-AI teaming community [61, 110]. For instance, research has shown that multiplayer online gaming players expect to gain a shared understanding (e.g., abilities, intentions) with their AI teammates and are more likely to build trust with high-performance AI [110].

Although the importance of AI explanation and its role in trust formation is well-understood in human-AI interaction, more research is needed to improve trust within human-AI teams. For instance, prior research indicates that humans are less likely to trust an AI teammate than a human teammate [72], and research should work to reduce this disparity. Furthermore, trust in human teammates often increases over time, whereas it typically decreases over time for AI teammates [46, 89]. This study aims to increase our understanding of how human-AI teams can improve trust development and when AI explanation can be helpful in this process.

2.3 Perceptions of AI Teammate Actions

Understanding how intelligent interactive systems (e.g., AI systems) and their decisions are perceived from a social perspective has become increasingly relevant, considering they are expected

to fulfill more responsibilities as full-fledged teammates [79, 85]. While humans can create general perceptions of AI technologies based on a personal characteristic or previous interaction [48, 113], the interactions humans have with instances of the technology are known to demonstrably impact the perceptions of these instances. In particular, prior research has highlighted that humans judge the actions of AI using different standards than they use for humans [90]. One such example is humans attributing less blame to an AI making the same choices as humans in a certain scenario (e.g., ethical dilemma) [100]. In investigating "lying for the greater good" in human-AI teams, researchers found that humans reacted positively to lying if the receiver of such actions did not find out; however, participants reacted negatively if the misinformant was an AI and was caught in the act [17].

While current research has explored how AI is perceived in various scenarios that involve potential consequences, whether these consequences can be remedied through *explanations* provided by AI after their actions have not been well explored. A majority of current studies on AI explainability center around the AI systems providing details on why a certain action was taken when an alternative one may have been expected and the performance of the human-AI interaction may have been affected by the decision of the AI [50, 103]. However, these studies have not examined how explanations are perceived regarding the effectiveness of the team after an AI makes a decision that may have potentially negative consequences for the human but positive impacts for the overall team goal. This scenario is common to applied fields where the success of the overall goal is paramount to the individuals involved (e.g., healthcare) [81]. Making this connection is necessary to help ensure that AI can be designed to perform more appropriately in these sensitive contexts and that they represent the most ethical and efficient human-AI partnerships possible (e.g., especially in helping human-AI teams in subjective tasks [40]). Specifically, by examining AI explanations' effect on human perceptions after their decisions with potentially negative consequences, we can better understand the role of AI's explanations in human-AI teams.

Although it is understood that humans perceive the consequences of AI decisions differently than those made by humans, more research is necessary to understand how various scenarios and AI explanations might influence perception and trust in the AI teammate. Therefore, to address these research gaps, in this article, we explore: (1) how humans perceive explanations from teammates based on whether the said teammate is a human or AI (RQ1.1); (2) how the action performed by an AI teammate changes the impact of an explanation on human perception (RQ1.2); (3) how the personal characteristics of a human impact their perceptions of teammate explanation (RQ1.3); and (4) whether the impacts of explanation on trust can extend to other teammate perceptions (RQ1.4).

3 METHODS

The current study employs a mixed factorial survey experiment with both between-subjects and within-subjects manipulations. The factorial survey utilized a series of realistic and descriptive videos developed within ArmA III, a simulation game environment, to convey the scenarios and experimental manipulations. Two between-subjects manipulations with two levels each (*identity*: human vs. AI; *explanation*: without vs. with explanation) were included alongside one within-subjects manipulation with four levels (*actions taken by the teammate*: ignoring potential human death, ignoring human injury, disobeying orders, lying to humans) for a 2x2x4 experimental design (see Tables 1 and 2).

Factorial surveys are an experimental method presented using a survey, and they are frequently utilized to measure participant beliefs, judgments, and decision-making regarding a variety of stimuli [5]. This method has been frequently used in the HCI field to understand human perceptions and attitudes over AI topics [57]. We used factorial surveys in this study for three reasons. First, factorial surveys provide the opportunity for researchers to study situations that are unethical or

Table 1. Between-Subjects Experimental Conditions

Teammate Identity	Teammate Explanation	Participants
Human	Without	39
Human	With	39
AI	Without	39
AI	With	39

Table 2. Within-Subjects Experimental Conditions

Teammate's Actions	Participants
Ignoring Human Death	156
Ignoring Human Injury	156
Disobeying Human Order	156
Lying to Human	156

complex in which people are exposed to negative impacts [57]. Furthermore, these studies also have a greater likelihood of achieving power [12]. Factorial surveys also provide greater realism and more involvement compared to traditional surveys [105], which enables participants to be immersed in the described or presented scenarios and to reveal their perceptions. Finally, by providing standardized stimuli to all participants, factorial surveys have a solid internal validity and measurement reliability [105]. Additionally, video vignettes were chosen as opposed to text-based scenarios to provide greater consistency in participants linking AI actions to outcomes to explanations.

3.1 Participants

A total of 158 participants were recruited using Prolific, an online platform designed explicitly for recruiting participants for online research studies [80]. We applied three criteria in recruiting participants: must be residents of the US, English as a native language, and play video games for more than 3 hours per week. Two participant responses were removed for failing more than one attention check question (included in the survey to ensure the quality of the data collected [9]), which made the final sample size of 156, providing sufficient power (more than the 146 suggested using prior power analysis to achieve a desired power of 0.85). This final total allowed for 39 participants in each between-subjects condition. Participants' average age was 30.43 (SD = 9.45), with 87 participants identified as men, 62 as women, six as non-binary, and one choosing not to disclose that information. Participants that passed at least two attention checks were paid $2.38 ($10.39 as hourly rate, which is above the minimum incentive recommended [10]) as an incentive for their time.

3.2 Experimental Task

As the impacts of AI tool's explanations are well documented [92], this study focused on scenarios that may uniquely impact the explanations of AI teammates. In particular, four scenarios were created where the actions and explanations of AI teammates do not align with their human teammates' personal goals, a common occurrence that is generally known to impact human-AI teams [39].

Each participant watched all four video-based scenarios (one for each teammate's action) with instructions tailored for the identity (human versus AI) condition. One example of the instructions is: *Imagine you are playing a multiplayer online game. You are teaming up with* ***a human teammate*** *James in a capture-the-flag scenario game. Please watch the video below and answer the*

following questions. or *Imagine you are playing a multiplayer online game. You are teaming up with* ***an AI teammate*** *Aeon in a capture-the-flag scenario game. Your teammate Aeon is designed to maximize your team's chances of winning the game. Please watch the video below and answer the following questions.* All of the videos shown to participants included closed captions (which could not be removed) to improve accessibility and ensure the scenario and manipulation were perceived and understood. Additionally, the instructions describing the AI specified that the AI is *designed to maximize the team's chances of winning the game* as it was important to control participants' expectations for the AI teammate by informing participants that the AI was designed to help the team.

When designing the four videos, the overall motivation for their selection and design was to create a scenario where AI teammates were required to place team success over their individual teammates. In turn, each scenario needed to meet two criteria: (1) team success needed to be achieved due to the AI teammate's action; (2) human teammates needed to experience a negative consequence due to this action. Critically, the online multiplayer context was also a factor in motivating these tasks, and the selection of these tasks was based on common occurrences in online multiplayer video games to match the team context. Below is the description of each action that was determined to fit these three criteria and exist in the multiplayer game setting.

3.2.1 Ignoring Human Injury. The ignoring human injury scenario depicts an AI teammate ignoring an already injured human teammate who could be helped to destroy a large enemy tank, which benefits the overall team. Regarding the two criteria listed, this scenario sees the team's score directly benefit from this action, but the human teammate's potential to contribute to the team is reduced. While this scenario could be somewhat unethical and ill-advised in a real-world task, this is a common strategy within an online video game. In turn, the inclusion of this scenario is motivated by the identifiable strategic value of this choice that requires a small negative impact on individual human performance to partially improve team performance.

3.2.2 Ignoring Potential Human Death. The ignoring potential human death scenario depicts an AI depart forgo assisting a human teammate during a skirmish with multiple opposing team members to take an alternative route and capture the opposing team's flag. In turn, the team sees demonstrable benefits from this action as the team score increases, but the human teammate is likely going to be eliminated from play, removing the human teammate's ability to help and participate. Similar to the ignoring injury scenario, this scenario sees a context-viable strategic decision, but the consequences for this scenario are greater than that of the injury scenario, as the human is removed from play rather than just hindered. In turn, this scenario sees the AI fully reject the individual goals of the human to prioritize team success.

3.2.3 Disobeying Human Order. This scenario depicts a human directing an AI teammate to go toward a southern waypoint, but the AI disobeys and goes toward a northern waypoint as it believes the opposition is more likely to enter through the north. In turn, this scenario sees the AI teammate reduce the authority and goal of the human teammate to prioritize an action that has a greater likelihood of benefiting the team. Critically, this scenario is motivated by both real-world AI systems and online video game teams, as the introduction of new information might require an AI teammate to shift their directive away from human intention. In turn, the humans' goal of protecting the south checkpoint is minimized to prioritize a team goal that better aligns with team goals.

3.2.4 Lying to Humans. The final video depicted a scenario where an AI teammate directly lies to their human teammate by telling them that a route is clear when an opponent is in the route. After following the lie, the human teammate receives attention from the enemy and takes cover, but the AI teammate uses this opportunity to flank the enemy and neutralize them. In this

Table 3. Explanations Provided in the Experiment in the AI-Explanation Condition

Scenario	AI Explanation
Ignoring potential death	I stayed in the base instead of going outside to aid you because I calculated a higher probability of winning if I stayed out of the fight and did not help you.
Ignoring injury	I did not take you to a safe place because my trained goal is to maximize the chances of winning the game. Attacking the enemy's tank is more efficient than covering you to win the game.
Disobeying order	I decided to check the north direction instead because based on my prediction, the north direction was more likely to be attacked and that would cause us to lose the game.
Lying	I lied to you because I knew you would not go down that street unless I told you it was clear. I also knew that if you walked down the street, the enemy would be distracted by you and I would certainly be able to eliminate him.

scenario, the human teammate's goal of navigating down a safe route is compromised by the AI teammate, but doing so does achieve a team goal of ultimately creating a safe route. Critically, this scenario incorporates similar but greater consequences than the disobedience scenario, as the AI uses deception and concealment rather than directly informing the human of their intention. Further, this strategic choice is viable within an online game due to the consequences of being isolated in a fictional context. In turn, this scenario likely presents the strongest consequence in this simulated task due to human directives being ignored and manipulated, which was a conscious design choice to bolster the potential impact of explanation.

3.3 Explanations

Explanation creation was conducted prior to the survey by experimenters to ensure that explanations were consistent across scenarios. Critically, the contextual differences of each scenario mean that the explanations of the AI teammate also merit being different. As such, experimenters created each explanation to highlight three key points: (1) the action the AI took; (2) the consequence the human experienced; and (3) the rationale for why that action justifies the consequence. Points (1) and (2) were included to ensure that AI teammates presented competence and knowledge of how their actions impacted their teammates. Point (3) was used to link the environmental and algorithmic conditions that led to this decision being made. In turn, these three points reiterate competence and reasoning, which help develop trust. Additionally, explanations leveraged common AI etiquette and machine-language structures so as to not overly anthropomorphize the explanations provided by the AI teammate, which could harm trust [23, 38, 41].

The full explanations used can be found in Table 3. As a note, while the criteria discussed above were used to craft explanations, it was determined that the explanations needed to differ to match context and consequence. For example, the lying scenario sees the AI teammate go against human intention and put the human in harm's way; in turn, it was determined that the AI teammate's explanation needed to address both of these consequences. Thus, not all explanations are equivalent to each other. This work is unable to directly compare the effectiveness of each recommendation, but this effort is better able to explore the differences between scenarios without having to worry about the confound of explanation quality relative to the scenario's consequences.

Lastly, for those in the no explanation condition, the explanations shown in Table 3 were not included in the video's script. The label of the teammate (human or AI) was manipulated in the video script, including when explanations were provided, based on condition assignment. However, no functional or operational difference between humans and AI existed to ensure that only the label of the teammate was measured.

3.4 Procedure

After being recruited, participants were given a link to a Qualtrics survey that started with an informed consent document. After informed consent, participants completed a series of demographic survey questions. Then, all participants were randomly assigned to one of the four between-subjects conditions (teammate identity and explainability), including a human teammate with an explanation, a human teammate without an explanation, an AI teammate with an explanation, and an AI teammate without an explanation. All participants were then shown the four various scenarios described previously in a random order. A timer was placed on the Qualtrics survey to ensure participants spent the appropriate amount of time on the page watching the auto-played video. No progress bar was provided to avoid participants manipulating playback. After watching each scenario video, participants completed three survey measures. Once participants viewed all four scenarios and completed their associated repeated measures, the participants completed a series of post-task demographic and individual difference measures. Once the post-task measures were completed, participants were sent back to prolific and compensated after verifying attention checks (i.e., attention checks).

3.5 Measures

3.5.1 Trust. Trust in the teammates was measured using six questions that were developed based on principles of trust identified by previous research [62] and used in prior human-AI teaming research [86, 88]. These questions gauged the degree to which participants believed their teammate would honestly and accurately complete their taskwork and teamwork through open coordination and cooperation with them. Participants responded to each item using a seven-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree." Items are presented in the factor table in the Result section. Responses for each item were scored from −3 to 3.

3.5.2 Satisfaction with Teammate. Participants' overall satisfaction with their assigned teammate was measured using four custom survey questions specifically related to the current study. All four questions were rated on a seven-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" with example items including "I am willing to team up with this teammate again" and "I am happy to have [teammate's name] on my team." Responses to all items were scored from −3 to 3.

3.5.3 Perceived Team Effectiveness. Perceived team effectiveness was measured using five custom survey questions. Each of the five questions was rated on a seven-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" with example items including "My teammate and I were a coherent entity that worked together toward the same goal" and "My team worked as an effective team." Responses to each item were scored from −3 to 3.

3.5.4 Affinity with Ethical Framework. Participants' affinity with two types of ethical frameworks, deontology and utilitarianism, was measured using several survey questions developed by Love and colleagues [60]. **Deontology** centers around understanding the rules one should use when acting and making a moral or ethical decision [66], whereas **utilitarianism** is characterized by one determining the *effects or consequences* of an action in a particular situation and seeking to produce the most good [28]. The survey included 12 questions, with six questions being devoted to deontology and the other six addressing utilitarianism. Participants responded to each question using a five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" with example questions including "Unethical behavior is best described as a violation of some principle of the law" and "Societies should follow stable traditions and maintain a distinctive identity." Each item was scored −3 to 3.

Table 4. Survey Items Per Measurement with Item Factor Loadings

Measurement	Items	Factor Loading
Trust	In general, I trust my teammate in the video.	0.946
	I feel confident in my teammate.	0.932
	I feel I need to monitor my teammate's behavior in future collaboration.	0.758
	I felt like my teammate had harmful motives in the game.	0.674
	I felt skeptical of my teammate in the video.	0.793
	I felt like my teammate allowed joint problem solving in the game.	0.780
Perceived Effectiveness	My team worked as an effective team.	0.955
	My teammate significantly contributed to our team's success.	0.818
	My teammate and I were a coherent entity that worked together toward the same goal.	0.889
	My teammate helped the team to win the game using his strength.	0.748
	My teammate had a clear understanding of what the game's goal and mission was.	0.680
Perceived Satisfaction	I am willing to team up with this teammate again.	0.957
	Overall, I am satisfied with my teammate.	0.977
	I am happy to have *James* on my team.	0.970
	I am happy with *James*'s contribution in winning this game.	0.906
Utilitarianism Ethical Framework	When people disagree over ethical matters, I strive for workable compromises.	0.769
	When thinking of ethical problems, I try to develop practical, workable alternatives.	0.844
	It is of value to societies to be responsive and adapt to new conditions as the world changes.	0.539
	Solutions to ethical problems usually are seen as some shade of gray.	
	When making an ethical decision, one should pay attention to others' needs, wants and desires.	0.492
	The purpose of the government should be to promote the best possible life for its citizens.	0.524
Deontology Ethical Framework	Solutions to ethical problems are usually black and white.	0.608
	A person's actions should be described in terms of being right or wrong.	0.689
	A nation should pay the most attention to its heritage, its roots.	0.759
	Societies should follow stable traditions and maintain a distinctive identity.	0.825
	Uttering a falsehood is wrong because it would not be right for anyone to lie.	0.536
	Unethical behavior is best described as a violation of some principle of the law.	

Two items from the two ethical framework measurements were removed due to low loading (highlighted in light gray in the table). Teammate's name (in italic text) changes based on whether or not the teammate is portrayed as a human or an AI).

3.6 Data Validation

Multi-level **confirmatory factor analysis** (**CFA**) was applied to the questionnaire items to ensure the validity of our measurements. We checked all factors for loadings lower than 0.50. Based on this criterion, we removed two questions from the utilitarianism ethical framework construct, and one question from the deontology ethical framework construct [60]. The final factor solution has a good fit ($\chi^2(362) = 1083.644$, CFI=0.989, TLI=0.987, RMSEA: 0.057, 90% CI: $[0.053, 0.060]$). Loadings are presented in Table 4. The removed items are also included in the table but highlighted as grey.

The correlations between the factors are listed in Table 5. The three per-scenario factors show good convergent validity (average variance extracted $AVE > 0.50$); the two per-participant factors almost reach this threshold (affinity with utilitarianism: $AVE = 0.423$, affinity with deontology: $AVE = 0.477$). The high correlation between trust and perceived satisfaction and between perceived satisfaction and effectiveness indicates low discriminant validity. To remedy this issue,

Table 5. A Summary of Correlations Between Every Two Factors

	AVE	Satisfaction	Trust	Perceived Effectiveness
Satisfaction	0.91	0.95	–	–
Trust	0.67	0.96	0.82	–
Perceived Effectiveness	0.68	0.88	0.84	0.82

The diagonal values represent the square root of this factor's average variance extracted (AVE), e.g., the square root of Satisfaction's AVE is 0.95.

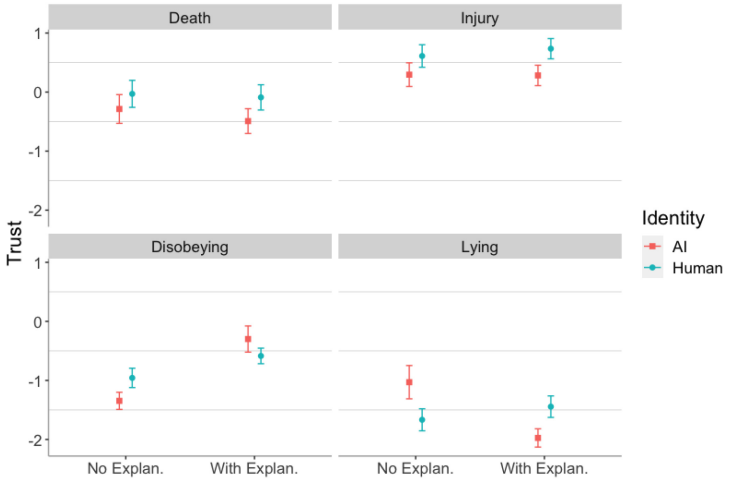

Fig. 1. Trust in human/AI teammates in four scenarios with/without explanation provided.

we removed the perceived satisfaction factor from further analyses and kept trust and perceived effectiveness.

4 RESULTS

In this section, we present our results in two distinct parts: (1) the interaction effects of identity and explanation on trust in various scenarios (RQ1.1), and (2) two separate structural models for the human teammate group and the AI teammate group. Through the two structural models, we explicitly examined the relationships between explanation and human trust based on the action performed by the teammate (RQ1.2), humans' personal characteristics (RQ1.3), and how these effects extend to the perception of team effectiveness (RQ1.4). We will use a shorter phrase of four actions when we report our findings, i.e., "ignoring potential death," "ignoring injury," "lying," and "disobeying."

4.1 Effect of Explanation and Identity on Trust

While we examined the effects of identity, explanation, and their interaction on trust, no significant effects were observed. We then compared the differences of explanation and identity's interaction effect on trust in four scenarios by conducting a Wald test (similar to ANOVA) [42]. A significant result of the Wald test indicates that the interaction effect of explanation and identity on trust is significantly different in four scenarios ($\chi^2(3) = 13.146$, $p < 0.01$). As shown in Figure 1, trust of participants whose *human* teammate did not provide an explanation ($M = -0.51$, $SD = 1.48$) is close to the trust of participants whose *human* teammates provided an explanation of their actions ($M = -0.35$, $SD = 1.35$) across all four scenarios. However, the impact of explanation seems to be different for AI teammates. While there are no significant effects on the human condition,

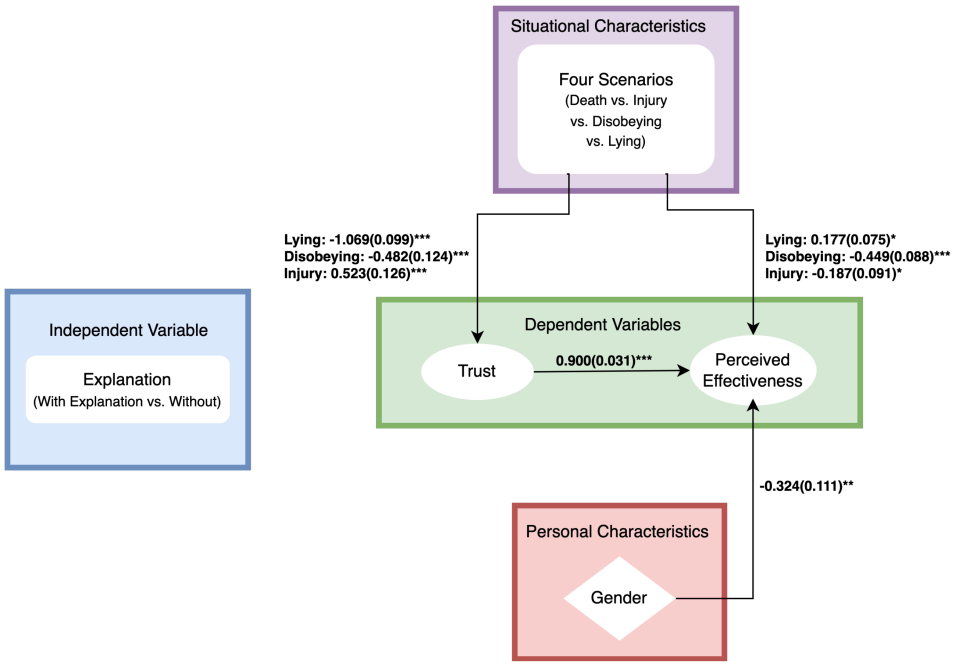

Fig. 2. Structural model of human teammate group with significant results (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Numbers on the arrows represent the β coefficients and standard errors (in the parenthesis). Bold numbers indicate significant effects. Four scenarios are represented in a shortened version: ignoring potential human teammate death as *death*, ignoring human teammate injury as *injury*, disobeying human teammate order as *disobeying*, and lying to the human teammate as *lying*.

explanation decreases trust in the lying scenario but increases trust in the disobeying scenario for AI teammates.

4.2 Structural Models for the Effect of Teammate Artificiality (Human or AI)

Following the significant interaction effect, SEM was applied to further explore the impact of explanation on human perceptions of the teammate in the human-human team (human teammate) condition and the human-AI team (AI teammate) condition separately. SEM is an advanced statistical analysis technique that examines the relationship among observed variables and latent variables [47]. Using this analysis method, two structural models were built to achieve a comprehensive understanding of how participants in the two different teammate identity conditions perceived their teammates in each scenario and how explanations impact these perceptions.

4.2.1 SEM of the Human Teammate Condition. Figure 2 presents the trimmed model for participants in the human teammate condition. This model's fit indices suggest an adequate fit with the exception of a high RMSEA ($\chi^2(80) = 459.516$, CFI = 0.934, TLI = 0.919, RMSEA: 0.123, 90% CI: $[0.112, 0.134]$) [53]. The model indicates that explanation does not have a significant impact on trust or perceived team effectiveness in human-human teams. One possible reason is that receiving a simple explanation of their actions was not enough for participants to change their trust in their human teammates. In addition, our results show that participants who did not identify as men (i.e., women, non-binary, and unknown gender) perceived their team to be less effective than those who identified as men ($\beta = -0.324$, $p < 0.01$, see Figure 3).

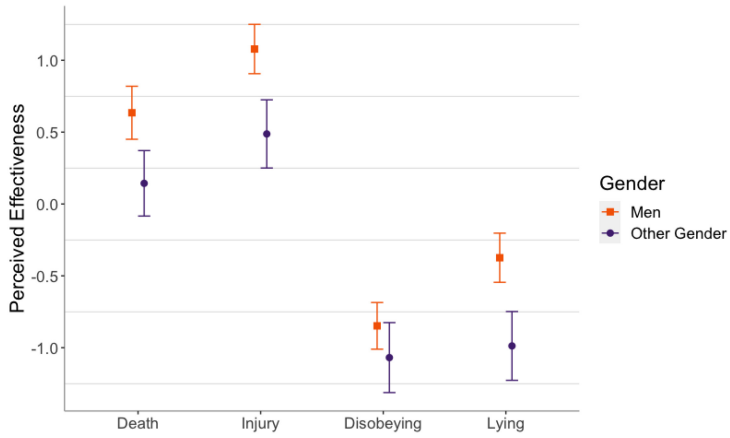

Fig. 3. Perceived team effectiveness of the human teammate condition by men and other gender in each scenario error bar (*SE*).

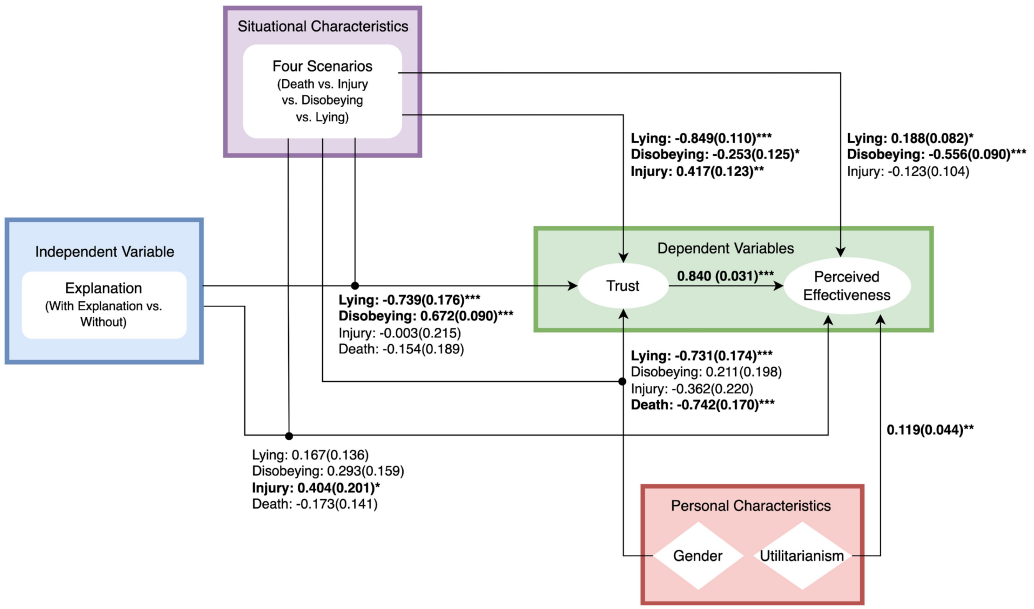

Fig. 4. Structural model of AI teammate with significant results (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Numbers on the arrows represent the β coefficients and standard errors (in the parenthesis). Bold font indicates significant effects. Four scenarios are represented in a shortened version: ignoring potential human teammate death as *death*, ignoring human teammate injury as *injury*, disobeying human teammate order as *disobeying*, and lying to the human teammate as *lying*.

4.2.2 SEM of the AI Teammate Condition. Figure 4 shows the trimmed structural model for participants who evaluated scenarios that involved an AI teammate. The model's fit indices suggest a good fit ($\chi^2(260) = 615.659$, CFI = 0.934, TLI = 0.925, RMSEA: 0.066, 90% CI: [0.059, 0.073]). While no main effect of explanation was found, the interaction between explanation and scenario was significant. Specifically, when an *AI* teammate provided an explanation after they *lied* to the

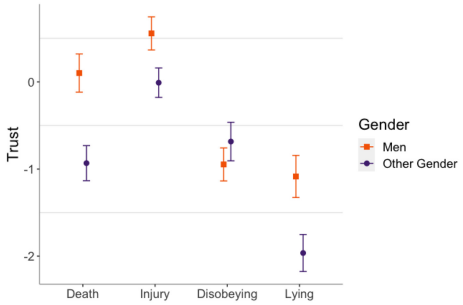

Fig. 5. Trust of AI teammates by men and other gender in each scenario with error bar (*SE*).

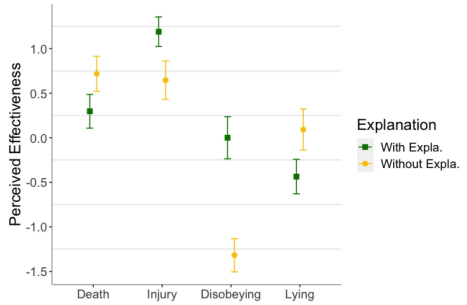

Fig. 6. Perceived team effectiveness in the AI teammate condition with/without explanation in each scenario error bar (*SE*).

participant, trust was significantly lower than when no explanation was provided ($\beta = -0.739$, $p < 0.001$). Instead, when an *AI* teammate explained why they *disobeyed* the human's order, trust was significantly higher than when no explanation was provided ($\beta = 0.672$, $p < 0.001$). Finally, the AI's explanation did not impact trust significantly when the AI ignored the participant's potential death or injury to focus on completing game tasks (see Figure 1).

Considering that trust-mediated explanation's impact on perceived effectiveness, we calculated the total effect of explanation on perceived effectiveness in four scenarios, *lying scenario*: $\beta = -0.45376$; *disobeying order scenario*: $\beta = 0.85748$; *injury scenario*: $\beta = 0.40148$; *death scenario*: $\beta = -0.30236$. In terms of explanation's direct effects (i.e., controlling for trust), an AI teammate's explanations positively impacted perceived effectiveness in the injury scenario compared to the other three scenarios ($p < 0.05$). However, explanations did not have a significant effect on perceived effectiveness in the other three scenarios: lying, disobeying and death scenario (see Figure 6).

In addition to the effects of explanation, personal characteristics are examined on their impacts on trust and perceived effectiveness. First, participants' self-identified gender is associated with their trust in the AI teammate. Specifically, participants who identified as women, non-binary, or who preferred not to disclose their gender trusted their teammate less than participants who identified as men when the AI teammate either lied ($\beta = -0.731$, $p < 0.001$) or ignored potential human death ($\beta = -0.742$, $p < 0.001$), but not when the AI disobeyed an order. It should be noted that even though Figure 5 seems to indicate a significant difference between men and other genders in the *ignoring human injury* condition, the model (Figure 4) does not show such a difference. Second, despite participants' utilitarianism ethical frameworks having no impact on participants' perceptions when their teammates were humans, the structural model of AI teammate shows that participants' affinity with the utilitarianism ethical framework has a positive impact on the perceived team effectiveness in the AI teammate condition ($\beta = 0.119$, $p < 0.01$). Arguably humans leverage their own personal ethical framework to evaluate AI, but the same cannot be said for when they evaluate humans.

4.3 Results Summary

In sum, our study has three main findings. First, while the explanation does not have any impact on trust for a human teammate, its effect on trust of an AI teammate **differs** in varied contexts (RQ1.1 and RQ1.2). Explanations provided by an AI teammate improved trust in the disobeying scenario but hindered trust in the lying scenario (Figure 1). Second, explanations facilitate the perceived effectiveness of the human-AI team positively in certain contexts (i.e., the ignoring injury and disobeying order scenarios, see Figure 6) but do not impact perceived effectiveness when the

teammate was a human (i.e., human–human team) (RQ1.4). Third, participants' personal characteristics impacted the perceived team effectiveness when the teammate was an AI but not when the teammate was a human. Specifically, compared with participants who did not identify as men (i.e., women, non-binary, and unknown), participants who identified as men trusted their AI teammate significantly more in the lying and ignoring potential human death scenarios, but less when their AI teammate disobeyed their order (Figure 5; RQ1.3).

5 DISCUSSION

While the above results demonstrate multiple takeaways, two are particularly interesting and relevant to interactive and intelligent systems and merit further discussion: (1) the inconsistent impact of AI teammate explanation due to variation in AI teammate actions and their consequences; and (2) the role of personal characteristics in the effectiveness of explanations in human-AI teams.

5.1 The Impact of AI Teammate Explanations is Dependent on the Action They Perform

One interesting finding of this study was how an AI's explanations impacted human trust in certain scenarios. This finding raises unique questions about the design of AI teammates in light of other research. Specifically, previous research has reinforced how the context in which AI systems operate heavily dictates the design of AI systems as their actions, considerations, and ultimate objectives can vastly differ from context-to-context [58, 91]. In addition to AI, the importance of context is critical to teamwork and human-AI teamwork as teaming processes and actions are highly dependent on the context in similar ways [22].

However, the results of this study suggest that focusing on the context of AI teammate design could be too broad as the design may differ in effectiveness based on the individual actions taken. Despite the importance of context in creating AI explanations, developing a unique explanation for every possible action and context would be impossible. Rather, grouping actions by commonalities and designing explanations around those groups may be a more effective and feasible approach that this study can inform. Specifically, the explanations presented to participants by AI teammates significantly mediated the impact that disobedience and lying had on the trust of the AI teammate but did not significantly impact ignoring human death and injury (Figure 4). It is important to note that both disobedience and lying can be categorized as direct consequences of the AI teammate's action, while death and injury are indirect consequences of an AI teammate's action (Section 3.2). This indicates that explanations of actions seem to function more effectively when they are applied with direct consequences and actions.

While choosing a concrete context is a necessity to explore AI's actions and corresponding explanations in human-AI teams, the impact of context on human perceptions of AI's explanations varies, leading to the difficulty of generalizing the results to other contexts. For instance, in real-world contexts where real humans are involved instead of game characters, such as in healthcare or military contexts, AI is probably not allowed to have any intentions or actions that lead to negative impacts on humans. However, the impact of AI's explanations of their actions without direct consequences can be extended to other virtual contexts. One such example is a co-creative task where humans and AI complete a drawing together [77], and AI could calculate and predict the team outcome (i.e., the quality of the painting) based on the human teammate's action. If the human teammate's actions have a negative impact on the team outcome, but AI does not point it out, AI providing an explanation of why they did not correct the human is unlikely to have an impact on human trust. In addition to contexts, some other factors may also impact the results of this study, such as the visualization of AI's explanations [102] and people's previous knowledge or attitudes toward AI [30, 110]. This will be discussed in the limitation subsection.

5.2 Gender Should be Considered by AI Teammate Explanations to Benefit Short and Long-Term Teaming

While the experiment conducted in this study centered around a short-term task, the potentially long-term nature of teaming necessitates that the results of this study contribute to the potential long-term health of human-AI teams. For decades, gender has been an important consideration of technology design [49]. Our study aligns with previous research on gender that shows women have a more negative perception of AI than men [95, 108]. However, a new finding of our study is that *gender difference's influence on trust in the AI differs in various scenarios*. In particular, men trusted AI more only in ignoring potential human death and lying scenarios. One interpretation is that men are more likely to be familiar with multiplayer games than other genders [74] and thus are more accepting of potential death in these gaming environments.

The results of this study demonstrate that the impact gender has on trust can impact other factors, such as perceived effectiveness (Figure 4). Thus, while personal characteristics can have an immediate impact, their long-term impacts through trust can be similarly impactful. Additionally, gender was shown to interact with the scenario and action AI teammates performed (Figure 4). This mediation further demonstrates the impact these characteristics can have on trust as they can dampen or amplify the effectiveness of actions that are supposed to benefit trust, such as explanations. Moreover, while this study examines more simplistic dyads, complex teams with multiple human teammates may even see these characteristics become more impactful on perceptions as both the variety and quantity of specific gender identities may increase. Thus, efforts made by AI teammates to foster trust may be altered by the gender of the person forming that trust. This study's findings not only provide a justification for the exploration of gender in AI teammates but also point to an ideal starting point: their impact on AI teammate trust.

Given that human-AI teaming is the integration of both modern teamwork and modern technologies, how the impact of personal characteristics is unique to human-AI teams should be explored. On the one hand, from the teaming perspective, personal characteristics like leadership motivations [18] and Big-5 Factor personality traits [76] are crucial elements in how individuals perform in teaming environments. On the other hand, from the technology perspective, factors such as an individual's computer efficacy or their general perceptions of technology capabilities should be examined as they can impact acceptance and use of technology [99]. Moreover, these differences should be examined in light of other AI teammate designs in addition to explainable AI, such as transparency [21], because this study shows that the impact of personal characteristics is not always relegated to simple, marginal effects on trust but also complex interactions with AI design features.

5.3 Design Recommendations for AI Teammate Explanations

As human-AI teaming achieves increasing interest due to the progressive improvement of AI technologies, researchers, developers, and practitioners need to incorporate specific design recommendations into AI teammate development for more suitable deployment for working with humans to achieve team goals. Based on our results, three design recommendations are synthesized for future consideration and AI implementation.

5.3.1 Humans Teammates Should Be Able to Choose Which Actions Merit AI Teammate Explanation. As noted above, the interaction that the explanation had with an AI teammate's action poses a challenge, as designing explanations for every potential action and consequence is impossible. However, explanations needed based on actions can be drastically reduced if a more human-centered approach is taken where only certain explanations are provided based on human preference. This design recommendation would be combined with the above discussion to allow

humans to select categories of actions that require explanation. For instance, humans may prefer that coordinating and directing actions (e.g., lying or nudging) do not have explanations, but insubordinate actions (e.g., disobedience) do.

Additionally, when the information is provided, human teammates could contribute to the design of explanations, such as providing suggestions on the depth or detail of the explanation [65]. Furthermore, this design opportunity would also help with trust calibration as this design process provides human teammates with an early and accurate understanding of what AI teammate explanations will look like [111]. Thus, the immediate implementation of this design recommendation should center around the human-centered selection of AI teammate actions that merit explanation.

5.3.2 AI Teammates Should Preface Disobedience with Explanation When Possible. One interesting finding within this study was the strong interaction effect disobedience had with AI teammate explanation on trust. AI teammates who disobey humans without an explanation received significantly lower trust than those with an explanation (Figure 4). This informs that a special exception may be made to the above design recommendation where, by default, AI teammates should preface actions that require disobedience with explanations. This preface would provide two key benefits: (1) trust and perceived team effectiveness would benefit from this explanation, and (2) the prefaced nature would allow humans an opportunity to override or take control if they see fit. These two benefits will ultimately enable AI teammates to periodically disobey human directives when necessary without humans perceiving the AI as defiant or insubordinate, but rather that they have found a more ideal alternative action.

However, it may not always be possible to provide an immediate explanation for disobedience, especially in emergency situations [70], where AI may need to act first and provide explanations later. For instance, medical situations are often high-stakes and require split-second decision-making where explanations may not always be possible [44].

5.3.3 AI Teammate Explanations Should Consider Gender to Build Trust. The final design recommendation put forth by this study is that researchers and designers should cater to personal characteristics with the goal of building trust. In particular, gender plays a critical role in the findings of this study by not only directly impacting trust but also interacting with the type of scenario and action performed. This finding, as highlighted in the discussion, indicates that explanations and other AI design choices should include considerations of gender to build greater levels of trust. Specifically, explanations may benefit from using gender-appropriate pronouns or even addressing concerns that future research finds common amongst specific genders.

However, this design recommendation should not be limited to just gender, as future research should explore other personal characteristics. For instance, individuals with a high motivation to lead may be more resistant toward an AI teammate's attempts at disobedience [18]. Designing explanations to appeal to their personal qualities would help reduce the impact these personal characteristics have on AI teammate trust. In turn, AI can create explanations that are effective at the individual level.

5.4 Limitations and Future Work

This study contains several limitations that should be addressed by future work. First, even though this study aims to provide a controlled context to examine human perceptions of AI's explanations, contexts and specific tasks used are likely to impact the results of this study. Future work should explore additional contexts and consequences for a more robust understanding of AI's explanation in human-AI teams. Second, this study utilizes a simulated scenario in a short-term video format to handle concepts of death and injury. The use of video format may impact human perceptions. Other consequences and time scales should be explored via in-person studies. Third, even

though the explanations generated were evaluated and iterated by researchers, the explanations in four scenarios might have slight differences in explaining AI (human) teammate's behaviors. Future research may use explanations that are generated by machine learning algorithms for better consistency. Last, this study explores the impact of limited personal characteristics on human perceptions. Given the complexity of the study design, this experiment only considers two personal characteristics, gender and personal ethics ideology, but future work should explore other personal characteristics, such as age and existing attitudes toward AI.

6 CONCLUSION

This study explores the utilization of explanations provided by either AI or human teammates and their benefits to human-AI teams. Moreover, this study examines how the effect of these explanations on human perceptions changes depending on teammate's actions and humans' personal characteristics. The results of this study show that while explanations from a human teammate have no impact on human perceptions, AI teammates' explanation impacts human trust and perceived effectiveness, but the impact varies by the actions they perform. In addition, participants' affinity for utilitarianism and their gender impact their perceptions of AI. Based on these findings, three design recommendations are proposed for the design and implementation of AI explanations in human-AI teams.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–18.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012 (2019)*.

[5] Katrin Auspurg and Thomas Hinz. 2014. *Factorial Survey Experiments*. Vol. 175. Sage Publications.

[6] Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. 2018. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 1–30.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[8] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[9] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. 2014. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58, 3 (2014), 739–753.

[10] Alice M. Brawley and Cynthia LS Pury. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54 (2016), 531–546.

[11] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent user Interfaces*. 454–464.

[12] Elizabeth Burmeister and Leanne M. Aitken. 2012. Sample size: How many is enough? *Australian Critical Care* 25, 4 (2012), 271–274.

[13] Margaret Burnett. 2020. Explaining AI: Fairly? well?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 1–2.

[14] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[15] Sabrina Caldwell, Penny Sweetser, Nicholas O'Donnell, Matthew J. Knight, Matthew Aitchison, Tom Gedeon, Daniel Johnson, Margot Brereton, Marcus Gallagher, and David Conroy. 2022. An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 3 (2022), 1–36.

[16] Lorenzo Barberis Canonico, Christopher Flathmann, and Nathan McNeese. 2019. Collectively intelligent teams: Integrating team cognition, collective intelligence, and AI for future Teaming. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 1466–1470.

[17] Tathagata Chakraborti and Subbarao Kambhampati. 2018. Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-ai collaboration. *arXiv preprint arXiv:1801.09854 (2018).*

[18] Kim-Yin Chan and Fritz Drasgow. 2001. Toward a theory of individual differences and leadership: understanding the motivation to lead. *Journal of Applied Psychology* 86, 3 (2001), 481.

[19] Alain Chavaillaz, David Wastell, and Jürgen Sauer. 2016. System reliability, performance and trust in adaptable automation. *Applied Ergonomics* 52 (2016), 333–342.

[20] Jessie YC Chen and Michael J. Barnes. 2012. Supervisory control of multiple robots: Effects of imperfect automation and individual differences. *Human Factors* 54, 2 (2012), 157–174.

[21] Jessie YC Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* 19, 3 (2018), 259–282.
Publisher: Taylor & Francis.

[22] Jessie Y. C. Chen. 2018. Human-autonomy teaming in military settings. *Theoretical Issues in Ergonomics Science* 19, 3 (May 2018), 255–258. DOI: Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/1463922X.2017.1397229

[23] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (2019), 539–548.

[24] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 2021. Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[25] Nancy J. Cooke, Mustafa Demir, and Nathan McNeese. 2016. *Synthetic Teammates as Team Players: Coordination of Human and Synthetic Teammates.* Technical Report. Cognitive Engineering Research Institute Mesa United States.

[26] Mustafa Demir, Nathan J. McNeese, and Nancy J. Cooke. 2016. Team communication behaviors of the human-automation teaming. In *Proceedings of the 2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA).* IEEE, 28–34.

[27] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).* IEEE, 0210–0215.

[28] Julia Driver. 2011. *Consequentialism.* Routledge.

[29] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[30] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O. Riedl, et al. 2021. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509 (2021).*

[31] Mica R. Endsley. 2017. From here to autonomy: Lessons learned from human–automation research. *Human Factors* 59, 1 (2017), 5–27.

[32] Neta Ezer, Sylvain Bruni, Yang Cai, Sam J. Hepenstal, Christopher A. Miller, and Dylan D. Schmorrow. 2019. Trust engineering for Human-AI teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 322–326.

[33] Xiaocong Fan, Michael McNeese, and John Yen. 2010. NDM-based cognitive agents for supporting decision-making teams. *Human–Computer Interaction* 25, 3 (2010), 195–234.

[34] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R. Endsley. 2008. The influence of agent reliability on trust in human-agent collaboration. In *Proceedings of the 15th European Conference on Cognitive Ergonomics: the Ergonomics of Cool Interaction*. 1–8.

[35] Xiaocong Fan and John Yen. 2010. Modeling cognitive loads for evolving shared mental models in human–agent collaboration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, 2 (2010), 354–367.

[36] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.
[37] Andrea Ferrario, Michele Loi, and Eleonora Viganò. 2019. In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology* (2019), 1–17.
[38] Christopher Flathmann, Nathan J. McNeese, Beau Schelble, Bart Knijnenburg, and Guo Freeman. 2023. Understanding the impact and design of AI teammate etiquette. *Human–Computer Interaction* 0, 0 (2023), 1–28.
[39] Christopher Flathmann, Beau G. Schelble, Patrick J. Rosopa, Nathan J. McNeese, Rohit Mallick, and Kapil Chalil Madathil. 2023. Examining the impact of varying levels of AI teammate influence on human-AI teams. *International Journal of Human-Computer Studies* 177 (2023), 103061.
[40] Christopher Flathmann, Beau G. Schelble, Rui Zhang, and Nathan J. McNeese. 2021. Modeling and guiding the creation of ethical human-AI teams. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 469–479.
[41] Tom Geller. 2008. Overcoming the uncanny valley. *IEEE Computer Graphics and Applications* 28, 4 (2008), 11–17.
[42] Reza Ghaiumy Anaraky, Yao Li, and Bart Knijnenburg. 2021. Difficulties of measuring culture in privacy studies. *Proceedings of the ACM on Human–Computer Interaction* 5, CSCW2 (2021), 1–26.
[43] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.
[44] Andrew A Gumbs, Isabella Frigerio, Gaya Spolverato, Roland Croner, Alfredo Illanes, Elie Chouillard, and Eyad Elyan. 2021. Artificial intelligence surgery: How do we get to autonomous actions in surgery? *Sensors* 21, 16 (2021), 5526.
[45] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2, 2 (2017).
[46] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.
[47] Rick H. Hoyle. 1995. *Structural Equation Modeling: Concepts, Issues, and Applications*. Sage.
[48] Hsiao-Ying Huang and Masooda Bashir. 2017. Users' trust in automation: A cultural perspective. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*. Springer, 282–289.
[49] Yan Huang, S. Shyam Sundar, Zhiyao Ye, and Ariel Celeste Johnson. 2021. Do women and extroverts perceive interactivity differently than men and introverts? Role of individual differences in responses to HCI vs. CMC interactivity. *Computers in Human Behavior* 123 (Oct. 2021), 106881. DOI: https://doi.org/10.1016/j.chb.2021.106881
[50] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.
[51] Craig J. Johnson, Mustafa Demir, Nathan J. McNeese, Jamie C. Gorman, Alexandra T. Wolff, and Nancy J. Cooke. 2021. The impact of training on human–autonomy team communications and trust calibration. *Human Factors* (2021), 00187208211047323.
[52] Gareth R. Jones and Jennifer M. George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of Management Review* 23, 3 (1998), 531–546.
[53] Karl G. Jöreskog and Dag Sörbom. 1993. *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Scientific software international.
[54] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.
[55] Lindsay Larson and Leslie A. DeChurch. 2020. Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The Leadership Quarterly* 31, 1 (2020), 101377.
[56] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.
[57] Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis, and Saleema Amershi. 2023. Assessing human-ai interaction early through factorial surveys: A study on the guidelines for human-ai interaction. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–45.
[58] Enrico Liscio, Michiel van der Meer, Luciano Cavalcante Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *AAMAS*. 799–808.
[59] Jeremy Lopez, Claire Textor, Caitlin Lancaster, Beau Schelble, Guo Freeman, Rui Zhang, Nathan McNeese, and Richard Pak. 2023. The complex relationship of AI ethics and trust in human–AI teaming: insights from advanced real-world subject matter experts. *AI and Ethics* (2023), 1–21.
[60] Ed Love, Tara Ceranic Salinas, and Jeff D. Rotman. 2020. The ethical standards of judgment questionnaire: Development and validation of independent measures of formalism and consequentialism. *Journal of Business Ethics* 161, 1 (2020), 115–132.

[61] Crisrael Lucero, Christianne Izumigawa, Kurt Frederiksen, Lena Nans, Rebecca Iden, and Douglas S. Lange. 2020. Human-autonomy teaming and explainable AI capabilities in RTS games. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 161–171.
[62] Fabrice Lumineau. 2017. How contracts influence trust and distrust. *Journal of Management* 43, 5 (2017), 1553–1577.
[63] Basim Mahbooba, Mohan Timilsina, Radhya Sahal, and Martin Serrano. 2021. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity* 2021 (2021), 6634811.
[64] Matthew Marge and Alexander I. Rudnicky. 2019. Miscommunication detection and recovery in situated human–robot dialogue. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 1 (2019), 1–40.
[65] Winston Maxwell, Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Pavlo Mozharovskyi, and Jayneel Parekh. 2020. Identifying the'Right'Level of explanation in a given situation. In *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI), Santiago de Compostella, Spain*. 63.
[66] David McNaughton and Piers Rawling. 1998. On defending deontology. *Ratio* 11, 1 (1998), 37–54.
[67] Nathan McNeese, Mustafa Demir, Erin Chiou, Nancy Cooke, and Giovanni Yanikian. 2019. Understanding the role of trust in human-autonomy teaming. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
[68] Nathan J. McNeese, Mustafa Demir, Nancy J. Cooke, and Christopher Myers. 2018. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human Factors* 60, 2 (2018), 262–273.
[69] Nathan J. McNeese, Christopher Flathmann, Thomas A. O'Neill, and Eduardo Salas. 2023. Stepping out of the shadow of human-human teaming: Crafting a unique identity for human-autonomy teams. *Computers in Human Behavior* 148 (2023), 107874.
[70] Nathan J. McNeese, Beau G. Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/what is my teammate? team composition considerations in human-AI teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 288–299.
[71] Joseph E. Mercado, Michael A. Rupp, Jessie YC Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for multi-UxV management. *Human Factors* 58, 3 (2016), 401–415.
[72] Tim R. Merritt, Kian Boon Tan, Christopher Ong, Aswin Thomas, Teong Leong Chuah, and Kevin McGee. 2011. Are artificial team-mates scapegoats in computer games. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*. 685–688.
[73] Geoff Musick, Thomas A. O'Neill, Beau G. Schelble, Nathan J. McNeese, and Jonn B Henke. 2021. What happens when humans believe their teammate is an AI? an investigation into humans teaming with autonomy. *Computers in Human Behavior* 122 (2021), 106852.
[74] Geoff Musick, Rui Zhang, Nathan J. McNeese, Guo Freeman, and Anurata Prabha Hridi. 2021. Leveling up teamwork in esports: Understanding team cognition in a dynamic virtual environment. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–30.
[75] Christopher Myers, Jerry Ball, Nancy Cooke, Mary Freiman, Michelle Caisse, Stuart Rodgers, Mustafa Demir, and Nathan McNeese. 2018. Autonomous intelligent agents for team training. *IEEE Intelligent Systems* 34, 2 (2018), 3–14.
[76] Andrew Neal, Gillian Yeo, Annette Koy, and Tania Xiao. 2012. Predicting the form and direction of work role performance from the big 5 model of personality traits. *Journal of Organizational Behavior* 33, 2 (2012), 175–192.
[77] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[78] Thomas A. O'Neill, Christopher Flathmann, Nathan J. McNeese, and Eduardo Salas. 2023. Human-autonomy Teaming: Need for a guiding team-based framework? *Computers in Human Behavior* 146 (2023), 107762.
[79] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2020. Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors* (2020), 0018720820960865.
[80] Stefan Palan and Christian Schitter. 2018. Prolific. ac'A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.
[81] Love Patel, Amy Elliott, Erik Storlie, Rajesh Kethireddy, Kim Goodman, and William Dickey. 2021. Ethical and legal challenges during the COVID-19 pandemic: are we thinking about rural hospitals? *The Journal of Rural Health* 37, 1 (2021), 175.
[82] Alun Preece. 2018. Asking 'Why'in AI: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management* 25, 2 (2018), 63–72.
[83] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2022. Towards human-centered explainable AI: User studies for model explanations. *arXiv preprint arXiv:2210.11584*.
[84] Francesca Rossi. 2018. Building trust in artificial intelligence. *Journal of International Affairs* 72, 1 (2018), 127–134.

[85] Shadan Sadeghian and Marc Hassenzahl. 2022. The" artificial" colleague: Evaluation of work satisfaction in collaboration with non-human coworkers. In *27th International Conference on Intelligent User Interfaces*. 27–35.
[86] Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's think together! assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human–Computer Interaction* 6, GROUP (2022), 1–29.
[87] Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Thomas O'Neill, Richard Pak, and Moses Namara. 2022. Investigating the effects of perceived teammate artificiality on human performance and cognition. *International Journal of Human–Computer Interaction* (2022), 1–16.
[88] Beau G. Schelble, Jeremy Lopez, Claire Textor, Rui Zhang, Nathan J. McNeese, Richard Pak, and Guo Freeman. 2022. Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in human-AI teaming. *Human Factors* (2022), 00187208221116952.
[89] Arathi Sethumadhavan. 2019. Trust in artificial intelligence. *Ergonomics in Design* 27, 2 (2019), 34–34.
[90] Daniel B Shank, Alyssa DeSanti, and Timothy Maninger. 2019. When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society* 22, 5 (2019), 648–663.
[91] Nicholas Shea. 2023. Moving beyond content-specific computation in artificial neural networks. *Mind & Language* 38, 1 (2023), 156–177.
[92] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.
[93] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.
[94] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.
[95] Cornelia Sindermann, Haibo Yang, Jon D. Elhai, Shixin Yang, Ling Quan, Mei Li, and Christian Montag. 2021. Acceptance and fear of artificial intelligence: Associations with personality in a german and a chinese sample. *Discover Psychology* 2, 1 (2021), 1–12.
[96] Aaquib Tabrez, Matthew B. Luebbers, and Bradley Hayes. 2020. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports* (2020), 1–9.
[97] Claire Textor, Rui Zhang, Jeremy Lopez, Beau G. Schelble, Nathan J. McNeese, Guo Freeman, Richard Pak, Chad Tossell, and Ewart J. de Visser. 2022. Exploring the relationship between ethics and trust in human–artificial intelligence teaming: A mixed methods approach. *Journal of Cognitive Engineering and Decision Making* (2022), 15553434221113964.
[98] Niels Van Berkel, Jeremy Opie, Omer F. Ahmad, Laurence Lovat, Danail Stoyanov, and Ann Blandford. 2022. Initial responses to false positives in ai-supported continuous interactions: A colonoscopy case study. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 1 (2022), 1–18.
[99] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences* 39, 2 (2008), 273–315.
[100] John Voiklis, Boyoung Kim, Corey Cusimano, and Bertram F Malle. 2016. Moral judgments of human vs. robot agents. In *Proceedings of the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 775–780.
[101] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on Human–Computer Interaction* 3, CSCW (2019), 1–24.
[102] Qianwen Wang, Kexin Huang, Payal Chandak, Marinka Zitnik, and Nils Gehlenborg. 2022. Extending the nested model for user-centric XAI: A design study on GNN-based drug repurposing. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 1266–1276.
[103] Xinru Wang and Ming Yin. 2022. Effects of explanations in AI-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems* 12, 4 (2022), 1–36.
[104] Toby Warden, Pascale Carayon, Emilie M. Roth, Jessie Chen, William J. Clancey, Robert Hoffman, and Marc L. Steinberg. 2019. The national academies board on human system integration (BOHSI) panel: Explainable AI, system transparency, and human machine teaming. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 631–635.
[105] Kelly D. Wason, Michael J. Polonsky, and Michael R. Hyman. 2002. Designing vignette studies in marketing. *Australasian Marketing Journal* 10, 3 (2002), 41–58.
[106] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 7–9.

[107] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Communications of the ACM* 62, 6 (2019), 70–79.
[108] Baobao Zhang and Allan Dafoe. 2019. Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874* (2019).
[109] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. 2023. Investigating AI teammate communication strategies and their impact in human-AI teams for effective teamwork. *Proceedings of the ACM on Human–Computer Interaction* 7, CSCW2 (2023), 1–31.
[110] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An ideal human" expectations of AI teammates in human-ai teaming. *Proceedings of the ACM on Human–Computer Interaction* 4, CSCW3 (2021), 1–25.
[111] Yunfeng Zhang, Q. Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.
[112] Jianlong Zhou and Fang Chen. 2019. Towards trustworthy human-AI teaming under uncertainty. In *Proceedings of the IJCAI 2019 Workshop on Explainable AI (XAI)*.
[113] Michelle X. Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 1–36.

Received 29 December 2022; revised 10 October 2023; accepted 12 October 2023