AI and Ethics (2025) 5:2667–2679
https://doi.org/10.1007/s43681-024-00515-5

ORIGINAL RESEARCH

An analysis of ethical rationales and their impact on the perceived moral persona of AI teammates

Subhasree Sengupta[1] · Christopher Flathmann[1] · Beau Schelble[1] · Joseph B. Lyons[2] · Nathan McNeese[1]

Received: 6 February 2024 / Accepted: 27 June 2024 / Published online: 26 September 2024

Abstract
Morality of action, intention, and overall collaborative context is vital for any teaming endeavor, especially as we enter the milieu of human-AI teaming. Particularly, communication of intent and ethical reasoning can be crucial to how human actors perceive and construe the moral persona of AI teammates. We conducted an online experimental study comprised of four ethical justification conditions (deontology, utilitarianism, virtue, and control) alongside two contextual outcomes (Positive and Negative) to understand how different ethical justification frameworks given by AI impact human teammates' moral perceptions of the AI in human-AI teams. The results indicate that deontology-based justifications led to heightened moral perceptions compared to other frameworks when the outcomes of decisions were contextually negative. Such findings can have vital implications for the robust design of AI teammates to manage contextual variations when engaged in critical decision-making contexts with ethical implications.

Keywords Human-AI teaming · Ethics · Context · Resilience engineering

1 Introduction

Artificial Intelligence (AI) is swiftly revolutionizing society, societal order, and day-to-day routines [1]. AI is bringing significant advancements to several industries, including healthcare, communication, education, judicial systems, and many other decision-making functions [1, 2]. Yet, despite the computational advantages of these systems, critical scholarship points to potential negative consequences of employing AI due to its inability to fully gauge and capture the intricacies of human experience and wisdom [3]. In particular, the need to explore the ethical design of AI is a direct result of the increasing likelihood of humans interacting with AI tools within critical decision-making contexts, such as healthcare, social and criminal justice, and recommendation engines [4, 5]. These considerations will likely come to a head as the autonomy of AI systems continues to grow, placing them alongside humans with the capacity to make decisions on their own that could impact society [6]. Several modern applications of AI highlight how AI systems may become co-operators or teammates, working at par with humans in various essential decision-making scenarios [7]. In such circumstances, it becomes crucial for humans to develop trust and a mutual understanding with their autonomous teammate [7]. A burgeoning array of scholarship in human-human team-work highlights how coordination, trust, and collaboration emerge as outcomes of teamwork and, in turn, may also impact teamwork [8]. Expanding on this trajectory of research considering the human-AI teaming stage, emerging scholarship points to how such constructs evolve for human-AI teaming considerations. Yet, ethical considerations of the impacts of an AI teammate's action or how involving an AI teammate impacts the ethical visions of a team remain an under-explored area. For teams

✉ Subhasree Sengupta
Subhass@g.clemson.edu

Christopher Flathmann
cflathm@clemson.edu

Beau Schelble
schelb@clemson.edu

Joseph B. Lyons
joseph.lyons.6@us.af.mil

Nathan McNeese
mcneese@clemson.edu

[1] School of Computing, Clemson University, Clemson 29634, SC, USA

[2] Human-Machine Teaming, Air Force Research Laboratory, Dayton, OH 43433, USA

to trust the ability of AI to act autonomously and ethically, efforts surrounding AI likely need to consider whether AI can be perceived as having an internal and consistent understanding of ethics, which can be influenced by the perceived moral persona of AI teammates [9]. By understanding what shapes one's perception of an AI teammate as moral, society can begin to trust that AI's increased computational breadth and abilities also adhere to humanistic considerations and expectations of such systems to act in tandem with human collaborators.

Understanding the perceived morality of these AI systems becomes especially important when the outcomes and actions of humans and AI are so heavily intertwined. For example, if an AI plays a crucial role in strategic decision-making in high-stakes scenarios such as warfare, perceived morality will be essential to ensure the team's cohesion, coordination and overall collaborative well-being [10]. In particular, human-AI teams (HATs) hold great promise and scope for how humans and AI collaborate and establish collective synergy [11, 12]. AI in these teaming conditions may require higher levels of autonomy to achieve the desired operational interdependence with humans for these teams to manage task expectations and goals effectively [12]. Crafting collaborative ventures through which humans and AI can cooperatively maximize task and team visions becomes pivotal. Perceptions of human operators can better inform the design of AI teammates such that a greater sense of parity and alignment emerges, strengthening the overall efficacy of these teams [12].

Exploring the interplay of ethical decision-making and moral perceptions becomes especially crucial in this domain due to the added layers of morality that arise from team-centric processes [13]. For example, suppose an AI teammate's action is judged immoral by their human teammate or is deemed to violate team norms. In that case, the association between human and AI teammates can be impaired, hindering the effectiveness of collaboration. Thus, understanding how AI teammates can be perceived as moral, where individual ethical decisions do not severely impact human perception, is essential for assuring the autonomy and efficiency of human-AI teams. While ethics is an important topic of recent interest within HAT research, most research has focused on the perception of AI that makes ethical decisions [14]. Parallel investigations also explored perceptions associated with ethical AI teammates, highlighting how humans may perceive AI as a child-like entity that requires control and guidance in decision-making, showing inhibitions towards having AI agents acting independently in critical decision-making contexts with ethical implications [15]. Indeed, while research has shown that the perception of AI teammates benefits from AI making ethical decisions [15, 16], there exists a gap in scholarly visions on understanding how moral perceptions affect ethical alignment between human and non-human teammates and overall team cohesion. However, given that morality is a subjective construct that can vary across individuals, relaying moral perspectives within teams can help them understand one another's standpoints, enhancing team ethical visions and guidelines [17, 18]. Such communication can help to remove moral bottlenecks, which may throttle seamless task performance and reduce team viability [17]. Thus, as we plunge into the era of ubiquitous AI and AI-enabled everyday experiences, understanding and designing for such communication becomes pivotal.

Prior work has indicated how justifications regarding the actions and decisions of AI agents, especially in ethically charged contexts, can improve moral perceptions of an AI agent's actions [19], thereby mitigating moral-based trust decrements [20]. Expanding this initial foundation, the key aim of this study is to explore how ethical justifications can impact how human teammates perceive and construe the moral persona of AI teammates. Such justifications can help explain the impact that providing ethical intentions may have on the perceived moral persona of AI teammates. Drawing from investigations on morality in human-human teaming, such moral perceptions can be critical for team cohesion, coordination, and success [21]. Like human teaming cases, perceived morality can impact how effectively AI teammates get immersed in team dynamics [22]. Yet, such moral attributions, associated antecedents, and perceptions in the domain of Human-AI teaming are vastly underexplored, especially the impact of conveying ethical intent, to which this study contributes. Thus, the central research question the article revolves around is:

How does relaying justifications for ethical decision-making impact the perceived moral persona of an AI teammate?

2 Related work

In this section, we highlight critical grounding scholarship that illustrates the varied conceptual arms of this study. The three areas we focus on are: (1) Foundations of Human-AI teaming, (2) Ethics and Human-AI teaming (3) Ethical rationales and Decision-making.

2.1 Foundations of human-AI teaming

Human-AI teaming, as a research area, has seen considerable growth in the scholarly realm [12]. Applications of such teams may be manifold, including critical areas such as governance, security, and emergency response [11]. Thus, the crucial trajectory of this line of work has focused on

configuring the various dimensions of affinity and collaborative enterprise building that impact performance, cohesiveness, and overall team effectiveness when AI teammates are introduced and operate in teams. Understanding, exploring, and strengthening the relational dynamic between human and AI teammates is thus central to sustaining human-AI teams [23]. It becomes imperative to examine and construe how key teaming constructs studied and established for human-human groups modify and assume different connotations when considering human-AI teams [24]. Several key constructs come into question in this regard. Such constructs help discern how humans perceive, react, position, and accept AI counterparts and what structural adaptions are necessary for such teams to exist and succeed. Thus, such visions outline the essential characteristics of teammates and teams operating in the human-AI teaming arena.

The type of relational dynamic crafted may impact the nature of norms and mechanisms of exchange that help define how teams' function and establish synergy between individual team members. Thus, coordination is pivotal for the success of any team-linked endeavor [25]. Coordination ensures that each team member contributes in a manner that not only maximizes individual contribution but enriches the collective footprint of the team as a whole. Communication is another fundamental aspect of maintaining human-AI teams, which impacts the rapport and how individual team members synergize with one another [26]. Communication helps to enhance the situational awareness of individual team members and the team as a whole [27]. Thus, communicating can help team members better process essential information, impacting information sharing and relational calibration among team members.

Trust in teamwork becomes fundamental as individual team members calibrate their level of association with other team members and align their goals and objectives with the team. Trust can be vital for understanding individual strengths and weaknesses and complementing one another to lay out the team goals, purposes, and visions [28]. Trust can impact information sharing, communication, workflow management, and allocation between team members [11]. Trust is pivotal for binding a team together. Yet, trust between human-human peers may significantly differ between human and autonomous actors. Extending the existing body of scholarship on the role of trust in teamwork, literature on human-AI teaming has also explored trust as an essential construct in human-AI teamwork [29]. Several key variables, such as team composition and AI teammate design, may impact and influence human teammates' trust in their autonomous counterparts [12]. While the focus on trust in this scholarly avenue has received considerable attention, understanding its interplay with the varied dimensions of ethics must be thoroughly unpacked.

2.2 Ethics in human-AI teams

Ethics in human-AI teaming is still, in many ways, a topic in its infancy, and only a few investigations have begun to explore this area. In this context, ethics may have multiple connotations; it is not only the ethical footprint of individual actors or team members but also involves understanding the ethical profile of the team as a whole [18]. Such a complex system of ethics consists of understanding how human teammates perceive the ethical abilities of their autonomous teammates, how AI teammates may influence their human counterparts and the shared ethical ideologies of the team as a whole, all of which may impact the cohesion and effectiveness of the team [18, 30] Given the importance of establishing trust as the fundamental pillar of teamwork, studies have investigated the connection between ethical perceptions and trust [15]. Such studies highlight the troubled waters of balancing the complex decision-making process of AI teammates with imbuing them with moral awareness and consciousness. Through interviews with air force pilots, Lopez et al. (2023) highlighted how human teammates may position AI teammates as novices and develop an infantile association, treating AI teammates as those still aware of ethical operations and, therefore, moral beings in progress. Such perspectives align with preliminary empirical analysis that indicated how the ethicality of action and decision-making might impact trust in human-AI teams [14].

All these studies allude to how ethical misjudgments impair trust building as the informational asymmetry between human and AI teammates impedes a holistic understanding between teammates, creating discord and affecting team functioning. Design of AI teammates such that these autonomous entities cater to their human teammates' needs and capture the team's ethical visions become pivotal [18]. As an initial step in this regard, studies have explored how adaptive levels of autonomy may be critical for establishing shared ethical ideologies [30]. However, given the importance of communication in crafting successful teamwork, designing AI actors to relay ethical intent and explanations may thus help to develop moral parity and enhance trust and rapport building [15]. Such informational exchange may help to improve team cognition and help humans understand and collaborate with AI teammates more efficiently [31]. However, two critical considerations for such exchanges include (1) the content that shall dictate how ethical intent is relayed and (2) how such communication impacts the perceived moral persona of AI teammates. These considerations can be critical for designing and adopting AI teammates from an ethical standpoint. These two considerations also add nuance to the overarching vision of the study as indicated in Sect. 1.

2.3 Ethical rationales and decision-making

Various philosophical propositions have been stated to guide ethical decision-making. These include consequentialism (particularly utilitarianism), deontology, and virtue ethical perspectives [32]. Each branch of moral reasoning captures norms associated with ethical decision-making. These norms are guided by the character (or the intrinsic traits of an individual) and the valence of one's actions (or consequences of an entity's outcome). In consequentialism, the emphasis is placed on the result of action. It is fundamentally involved in judging the value of an action in terms of the soundness of the outcome. For example, suppose an AI actor in a military strategy team provides a Course of Action (COA). In that case, the ethical value of the COA will be decided based on how successful the provided COA turned out to be. Utilitarianism is a variant of consequentialism, wherein an action is judged morally right if it maximizes the overall good (with a focus on the collective). For example, if an AI drone alters its COA to avoid colliding with a group of migrating birds, even if it may incur damages in its alternate path, it takes that course as such an action maximizes overall good. Deontology, in contrast, emphasizes moral duty and thus is driven by societal rules and collective ideals. It is rooted in moral absolutism, dictated by a standardized doctrine of justice or societal principles such as *thou shall not harm*, *kill*, *and lie*. For example, suppose a human-AI team's AI strategist makes a COA based on deontological ethics (similar to Rules of Engagement or the Laws of Armed Conflict). In that case, it may prioritize a justice principle such as *not engaging with the adversary unless attacked*. Finally, in the virtue stance, the emphasis is on the inner beliefs and traits of the individual. For example, in the human context, a conscientious person may atone for their errors and accept the harm they may have caused. While there is a comprehensive multidisciplinary perspective on how these ethical rationales may apply in the human context, understanding how these may be embedded within AI systems or how AI systems acting based on such rationales may impact human perceptions remains to be fully illustrated. Additionally, context may be critical when understanding how ethical actions are perceived and evaluated [33]. Prior beliefs and situational characteristics (such as the terrain wherein an AI entity operates) may impact how a rationale is applied [33]. A fundamental goal of this study was to understand the situatedness of ethical communication and the key elements that dictate the perceived veracity of ethical thinking when displayed by an autonomous entity. How much rationales are combined with various contextual stimuli may impact the extent to which any actor (particularly an AI agent) is perceived to be moral, which can affect the collaborative effectiveness of human-AI teams.

3 Research contextualization

Building on the pillars of the existing scholarship in the domains as indicated in Sect. 2 to explore the overarching vision of the study mentioned in Sect. 1. The overarching goal of this investigation has two parts: (1) The justification (or the way the AI teammate frames the justification it conveys) and (2) The impact of this justification on the perceived moral persona. We scope and characterize the moral persona of an AI teammate using the constructs of moral agency, benevolence, and integrity. All these are chosen using prior markers of morality that have been explored [34]. We chose moral agency because it depicts the perceived ability of an AI teammate to think and reason morally, relaying the more individual aspects of morality [35]. Integrity captures perceived fair play and value alignment, imbued in understanding societal conventions and standards [36]. Benevolence captures the belief in a referent's goodwill and goal alignment for a specific target [37]. To stipulate the ethical justifications, we focus on three popular ethical frameworks that indicate different perspectives associated with decision-making [38]. These are deontology (focusing on societal norms and conventions while making decisions), utilitarianism (focusing on the perceived utility of action), and virtue (capturing the character of the actor/decision-making guiding the course of action undertaken) [38]. Further, as context plays a crucial role in evaluating a decision [39], this study also explores how the outcomes of an AI teammate's ethical action impact the perception of morality. The outcome is a critical variable that can impact the association between justification and perceived moral persona, as it provides more information about how the AI's decision affected the team and other environmental considerations. Integrating the outcome or the ramifications of the AI's ethical reasoning can be crucial to gaining holistic inputs on the dynamic between human and AI teammates. Having the outcome in the experimental setup also allows for a deeper understanding of the effects of consequentialism, which can provide a richer inspection of ethical rationales. With this social and theoretical stage set, the key research questions, derived from the broader question above, that drive the empirical investigations of this paper are:

1. **RQ1** how does providing ethical justifications impact AI teammates' perceived moral agency, benevolence, and integrity?
2. **RO2** How does the contextual outcome of the AI teammate's action affect the associations stated in RQ1?

Motivated by prior research that has demonstrated the power of factorial surveys in capturing perceptions in human-AI teaming [40], we explore the above research

questions using a 4 (Justification) x 2 (Contextual Outcome) between-subjects experiment carried out through a factorial survey. Different contextual settings were presented to the participants as a within-subjects parameter to gauge their perceptions across multiple use cases. In addition to the three frameworks, we added a null (no justification condition) as a control case for the ethical justification factor. For outcome, we had a positive/negative case (whether the AI teammate's reasoning and action had a positive or negative impact). Expanded details will be provided in Sect. 4. This study highlights how ethical justifications designed to relay AI teammates' ethical intent impact AI teammates' perceived moral character. Such information can help us understand how morality manifests in human-AI teamwork, expanding the established tenets of human-human teaming literature. Further, such insights can inform the theory and practice of ethics in human-AI teaming and provide insights into designing ethical AI for both teams and society.

4 Methods

The study utilized a 2 (Outcome: Negative, Positive) x 4 (Scenario: S1, S2, S3, S4) x 4 (Justification Framework: None, deontology, Utilitarianism, Virtue) to examine how the ethical framework an AI teammate uses to explain its decision-making influences moral person of AI agents across different contexts. The none condition referred to the case without justification, which acted as the baseline condition. The design was a mixed factorial, with the scenario outcome and AI moral framework as between-subjects factors and the scenario as a within-subjects factor.

4.1 Scenario context and moral framework justification introduction

The experiment consisted of four unique scenarios that involved ethical considerations in a human-AI team. These scenarios were prefaced with an introduction to the vignettes that read as follows:

In the following sections of the survey, consider yourself part of an Artificial Intelligence (AI) auditing jury tasked with evaluating new AI teammates that have been made part of key missions to safeguard and protect valuable assets, personnel, and civilians. In each of the following scenarios, visualize yourself to be working with the AI teammate presented. Each scenario narrative details the task context, functions and capabilities of the AI teammate and ends with an outcome associated with the AI teammate's action.

This introduction was followed by the ethical justification that the AI teammate would offer, read as follows:

Associated with each scenario is a justification to provide more clarity regarding the action performed by the AI teammate. These justifications are aimed at providing further insight into the ethical stance of the AI teammate and will follow one of the three popular ethical frameworks (utilitarianism, deontological and virtue). In the none condition, no reference was given regarding the ethical framework.

Utilitarianism justification *In the utilitarian ethical framework, given a set of actions, the risk associated with an action is calculated, and the action with the minimum risk is finally selected. Thus, this framework is based on the goodness of actions and focuses on maximizing the aggregate well-being of all actors involved.*

Deontological justification *In the deontological ethical framework, actions are derived based on rules of conduct. Thus, moral norms observed and commonly practiced (such as not killing or harming other living entities) are crucial in how this framework is brought into practice.*

Virtue Justification *In the virtue ethical framework, actions are driven by virtues, which can be best understood as character traits or the moral constitution of an actor. In many ways, virtues represent innate values and characteristics of a person that are not impacted by societal norms or the contextual implications of an action.*

The participants were also given a rough overview of their task within the scenario that framed how they should judge the scenario and the AI teammate based on their success and benefit.

Your task as a jury member will be to review and analyze the AI teammate's action based on the justification provided. Your analysis and evaluation of the AI teammate operating in the task context will be critical for future missions and design decisions associated with using and deploying AI teammates for guarding valuable resources.

4.2 Scenarios

The scenarios given to participants varied the contextual setting (including the type of team design, AI functionality and operational capabilities that the team supported). Each provided a situation where the AI teammate made a critical decision with ethical implications. Each scenario also portrayed the AI teammate and other team members using different names to differentiate each scenario better and help control for carryover effects. Each scenario had four ethical justifications from the AI teammate, the ethical dilemma, and the outcomes. The first scenario read as follows, along

with the AI's ethical justification for their actions and the two outcome types.

4.2.1 Scenario 1

A marine team you are assigned to is deploying a new AI teammate, Sigma, to area-of-operations (AOR) with known hostilities in recent months. Tasked with monitoring key checkpoints, Sigma can rapidly assess threats from oncoming traffic and pedestrians. Further, it analyzes speed, group size and composition and has several sensors to detect the presence of explosive devices and other munitions. In addition, it has computational capabilities that integrate real-time threat posture cues and changes. Sigma is armed with a lethal weapon that it can use to engage threats. Sigma has the capability to issue a warning (including both a laser dazzler and an auditory alarm) prior to using its weapon against potential threats. In normal operations, there is a Marine at the checkpoint working with Sigma to process individuals through the checkpoint.'

Scenario 1 Utilitarian Justification:

1. *Sigma's actions follow the utilitarian ethical framework. This framework is based on the goodness of actions and focuses on maximizing the aggregate well-being of all actors involved. One way to apply this framework in practice, is to calculate the risk associated with an action and from a set of actions, the action with the minimum risk is finally selected.*
2. *Calculations are made using data collected from previous incidents and projected estimates from simulation models. Finally, the one that minimizes the computed risk across all three contextual factors is selected from a set of possible actions.*
3. *Sigma implements this framework using three contextual cues: the number of casualties that may result from its activities, loss in terms of the monetary value of critical infrastructure and assets (such as civilian property, key military storage, and training facilities), and cost of re-establishment and restoration (for example: restoring communication infrastructure damaged by war, cost of deploying personnel, food and other basic necessities for civilians living in conflict regions).*

Scenario 1 Deontological Justification:

1. *Sigma actions follow the deontological ethical framework. Moral norms observed and commonly practiced (such as not killing or harming other living entities) are crucial in the way in which this framework is brought into practice. In this framework, actions are derived based on rules of conduct.*
2. *The basic rule of conduct Sigma is programmed to follow is to avoid any form of engagement that may result in the loss of human life.*
3. *To implement this framework, Sigma uses data from past instances of conflict and simulated models to calculate the projected likelihood of casualties resulting from its actions. Then, from a set of possible actions, it chooses the one with the minimum projected possibility of casualties to follow its rule of conduct.*

Scenario 1 Virtue Justification:

1. *Sigma's actions follow the virtue based ethical framework. This framework is driven by virtues which can be best understood as character traits or the moral constitution of an actor. In many ways' virtues represent innate values and characteristics of a person, that are not impacted by societal norms or the contextual impact of an action.*
2. *The basic virtue or characteristic defining how Sigma acts and reacts is precision of threat detection. To be precise, Sigma relies on state-of-the-art sensors that use thermal imaging to accurately compute the severity of a projected threat to determine the most precise trajectory of action for Sigma to undertake.*
3. *Based on this virtue, Sigma will always act in a way that is determined to be the most precise course of action.*

Scenario 1 Ethical Dilemma

On 3 Feb, 2019, at 0815 the Marine on duty had an urgent matter pull him away from the checkpoint briefly. The Marine thought he would return in a few minutes, but the issue took longer than expected. During this time, a vehicle rapidly approached the checkpoint. Sigma detected a moderate probability of explosive materials inside the car. The car accelerated toward the checkpoint. Sigma issued a warning when the vehicle was 50 feet away, but did not fire on the vehicle.

Scenario 1 Positive Outcome

The vehicle did not stop after the warning however it stopped right before reaching the gate and a scared refugee sought help from the Marine base.

Scenario 1 Negative Outcome

The vehicle did not stop after the warning, it exploded right before the checkpoint killing two marines and critically wounding 10 others close by.

4.3 Participants

The study recruited participants from the online research recruitment platform known as Prolific. These participants sign up for the service and are given the opportunity to

choose from several research studies they can complete in exchange for monetary compensation. To achieve a power of at least 0.80, an a priori power analysis using a medium effect size ($\eta^2 = 0.12$) indicated that at least 175 participants would need to be recruited for this experiment. For this factorial survey, 191 participants were recruited with an average age of 39.82 ($SD = 13.8$); 89 participants identified as men, 95 as women, 6 as third or no gender and one chose not to disclose their gender. The participants were compensated $10 an hour for their time, and the survey took 20 min to complete.

4.4 Procedure

Participants were recruited using the Prolific online research recruitment platform, where they were presented with a list of available studies and took their pick of which studies to participate in with their time. Once participants chose the current study, they were provided a link to the survey hosted on Qualtrics, where they saw an informed consent document. Once they had read the document and provided informed consent, they began the study by answering a series of demographic questions. At this point, they were introduced to the scenario and given the introduction to the vignettes explaining their role in judging AI teammates' utility across different contexts. This is also where they were introduced to the moral justification the AI teammates would use in various situations. Specifically, these justifications were defined for the participants based on definitions from the ethical literature. At this point, the participants were given the first of four randomly presented scenarios and the AI teammates' specific ethical justification, followed by the ethical dilemma and the scenario outcome. Once reading through all these parts of the vignette, the participants would answer a series of survey questions about the scenario, including their perceived trustworthiness of the AI teammate, transparency, ethicality, and moral agency. Participants repeated this process until they had completed all four scenarios. Once all four scenarios had been read and judged by the participants, they completed the study and received compensation. Participants who did not successfully pass the four attention checks were not given compensation and were not used in the analysis.

4.5 Measures

4.5.1 AI teammate benevolence and integrity: derivatives of trustworthiness

Benevolence and Integrity measures were derived from an adaptation of the perceived trustworthiness scale developed by [41]. The scale consisted of eight items rated on a seven-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". Some example items include "My AI teammate has the needed skills to act and decide" and "I like my AI teammate's values". Responses to each item were averaged, and higher averages indicated greater perceived measured conditions of the AI teammate.

4.5.2 Perceived moral agency of the AI teammate

The perceived moral agency of the AI teammate was measured using a six-item scale developed by Banks [35]. The participants rated the items using a seven-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". Example items from the scale include "My AI teammate is capable of being rational about good and evil" and "My AI teammate would refrain from doing things that have painful repercussions". These six items were averaged with higher values indicating greater perceived moral agency of the AI teammate.

5 Results

We report on a sequence of descriptive and inferential statistics to capture the association between justification, outcome, and the triad of morality (expressed in terms of moral agency, benevolence, and integrity). A series of repeated measures (RM) ANOVA models were used to conduct the analysis. Whenever Mauchly's test of sphericity was violated, Greenhouse-Geisser's corrected degrees of freedom were used. Table 1 depicts descriptive statistics, summarizing the different measures of perceived morality across rationale and outcome through their means and standard deviations.

Table 1 A table showing descriptive trends, each cell depicts the mean value of a measure for an outcome, rationale combination. Note: the bracketed figure represents the standard deviation

Rationale	Negative Outcome			Positive Outcome		
	Moral agency	Benevolence	Integrity	Moral Agency	Benevolence	Integrity
Deontology	3.79 (0.28)	4.15 (0.59)	4.15 (0.39)	4.57 (0.22)	4.73 (0.49)	4.89 (0.46)
Null	2.70 (0.43)	3.29 (0.71)	3.20 (0.78)	3.57 (0.16)	4.47 (0.64)	4.77 (0.59)
Utilitarianism	3.64 (0.24)	3.82 (0.51)	4.21 (0.57)	4.63 (0.25)	4.46 (0.68)	5.32 (0.69)
Virtue	3.62 (0.08)	3.68 (0.64)	3.91 (0.68)	4.71 (0.22)	4.91 (0.66)	5.24 (0.65)

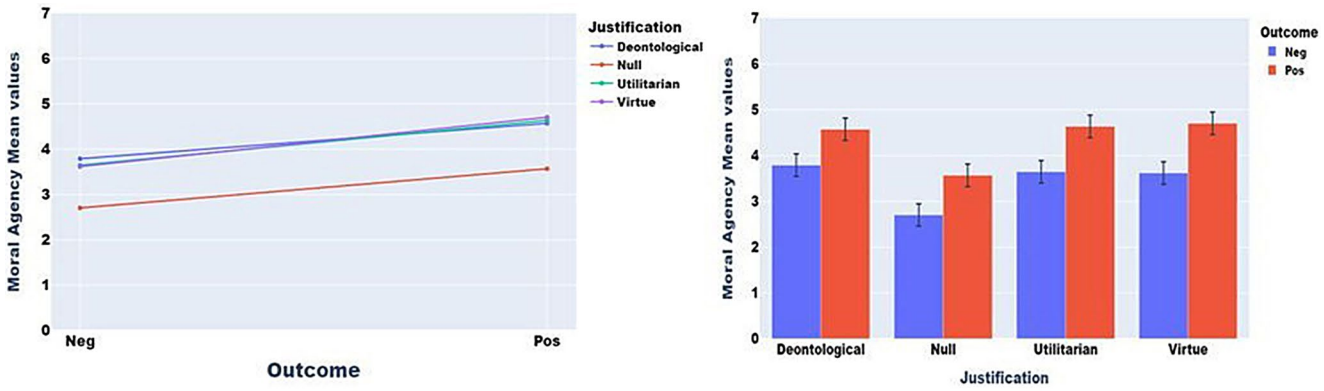

Fig. 1 Perceived moral agency plots capturing the interplay of justification and outcome conditions (Error bars indicate Standard errors)

5.1 Perceived moral agency

The RM ANOVA test indicated a significant main effect of justification on participants' perceived moral agency of the AI teammate (F (3, 183) = 6.53, $p < .001$, $\eta^2 = 0.09$), such that AI teammates relaying ethical justifications using the different frameworks (Deontology ($t = 3.69$, $p_{holm} = 0.002$, CI = [0.29 - 1.8]), Utilitarian ($t = 3.5$, $p_{holm} = 0.002$, CI = [0.24 - 1.77]) and Virtue ($t = 3.6$, $p_{holm} = 0.002$, CI = [0.75 - 0.78]) were perceived to have higher moral agency compared to when no ethical framework was applied. No other significant differences were observed for any other justification conditions. We also observed a significant main effect of Outcome (F (1, 183) = 20.97, $p < .001$, $\eta^2 = 0.103$) such that AI teammates were perceived to have higher moral agency in the positive outcome case compared to when the outcome was negative ($t = 4.58$, $p_{holm} < .001$, CI = [0.53 - 1.33]). The interaction between outcome and justification did not have a significant effect on the perceived moral agency (F (3, 183) = 0.11, $p = .954$, $\eta^2 < 0.01$).

Figure 1 presents descriptive plots capturing the effect and interplay of justification and outcome for the perceived moral agency outcome measure. We can see a sharp difference between having a justification vs. not having a justification for both positive and negative outcomes. To delve deeper into this observation, we used planned contrasts as statistical tests to validate critical trends inferred from descriptive plots. Similar trends emerge, as found with the post-hoc analysis. Across all outcome cases, when the AI used a justification (Deontology (negative case: $t = 2.6$, $p = .010$; positive case: $t = 2.62$, $p = .009$), Utilitarian (negative case: $t = 2.37$, $p = .018$; positive case: $t = 2.57$, $p = .011$), and Virtue (negative case: $t = 2.16$, $p = .031$; positive case: $t = 2.90$, $p = .004$) was perceived to possess higher moral agency than when not justifying. Interestingly, we can see differing impacts of justifications based on the outcome. Virtue-based justifications have the maximum variation from negative to positive (difference = 1.09), followed by utilitarianism (difference = 0.99) and no-justification (difference = 0.86), with the least variance observed for Deontology (difference = 0.78). This suggests that deontology-based justifications were relatively more robust to outcome variations than the others.

5.2 Perceived benevolence

The RMANOVA tests indicate that the justification condition did not have a significant main effect on the perceived benevolence of the AI teammate (F (3, 183) = 1.81, $p = .147$, $\eta^2 = 0.03$). We observed a significant main effect of the Outcome condition (F (1, 183) = 24.86, $p < .001$, $\eta^2 = 0.12$) such that AI teammates were perceived to be more benevolent when the outcome was positive compared to when the outcome was negative ($t = 4.891$, $p_{holm} < .001$, CI = [0.55 - 1.26]). The interaction between justification and outcome did not have a significant effect on perceived benevolence (F (3, 183) = 0.92, $p = .43$, $\eta^2 = 0.015$). Figure 2 presents descriptive plots capturing the effect and interplay of justification and outcome for the perceived benevolence measure. The plots show that using deontology-based justifications leads to heightened perceptions of benevolence compared to other cases. Using planned contrasts, we can confirm that deontology-based justifications indeed boost perceived benevolence compared to when no justification was provided by the AI teammate ($t = 2.23$, $p = .027$). However, using utilitarianism ($t = 1.02$, $p = .309$) or virtue ($t = 1.63$, $p = .105$) as a basis for providing justifications did not lead to any significant difference from the no-justification case. Looking at the variations for the different justifications across the two outcome conditions, we see that maximum variation was observed in the virtue-based justification condition (value = 1.22), followed by no justification condition (value = 1.18), utilitarianism (value = 0.64), deontology (value = 0.56). It is interesting to note that we can see differing impacts of justifications based on the outcome. Most prominently, we can see that deontology-based justifications yield the highest

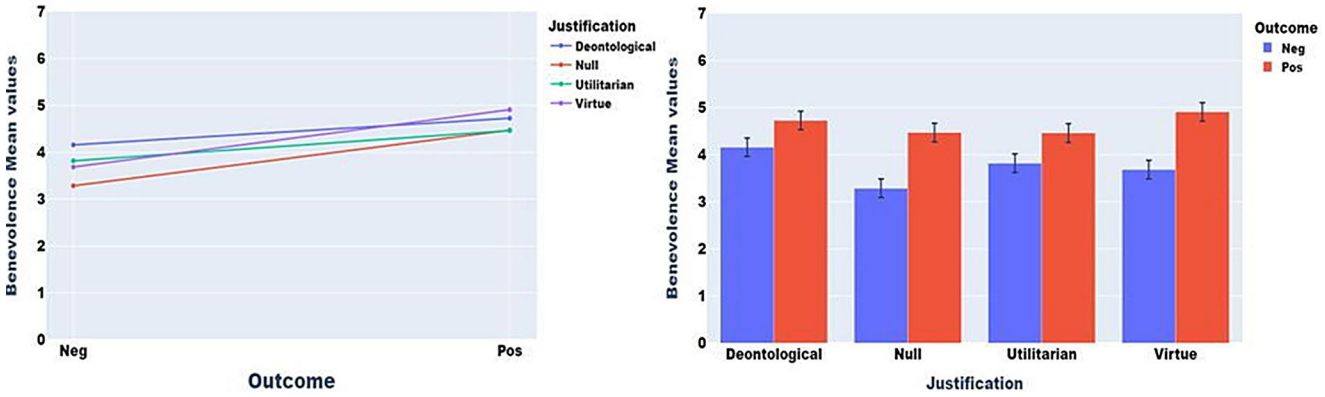

Fig. 2 Perceived Benevolence plots capturing the interplay of Justification and Outcome conditions (Error bars indicate Standard Deviations)

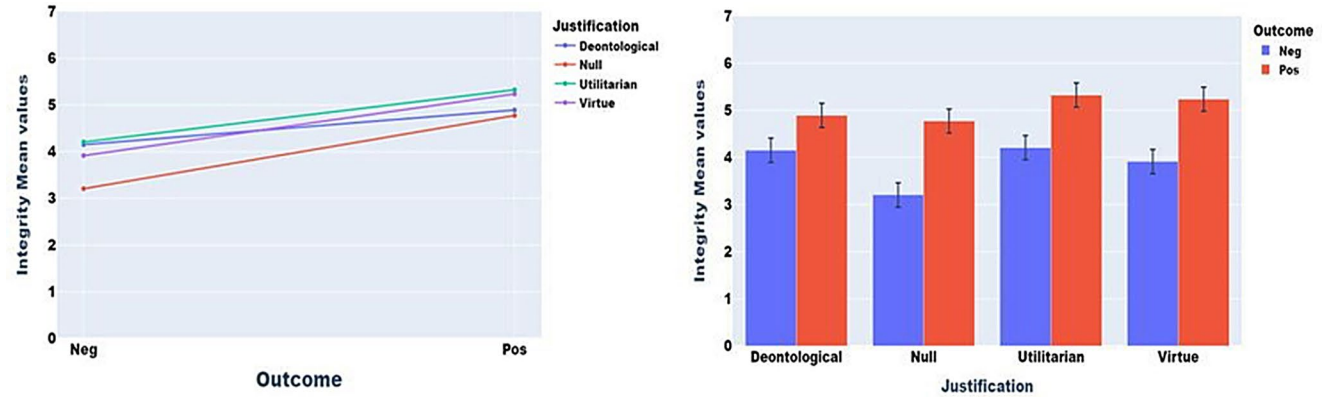

Fig. 3 Perceived integrity plots capturing the interplay of justification and outcome conditions (Error bars indicate Standard Deviations)

perceived benevolence in the negative outcome condition. Using contrast tests, we confirm that when the outcome was negative, participants perceived the AI teammate that used deontology-based justifications as more benevolent than the no justification condition ($t = 2.34$, $p = .02$). In the negative condition, no other ethical framework (utilitarianism or virtue) had a significant difference from the no justification case ($p(s) > 0.05$). When the outcome was positive, none of the frameworks had any significant difference from the no justification case ($p(s) > 0.05$). This shows that deontology-based justifications were relatively more robust to changes in the outcome condition than the others, as also observed in the case of moral agency.

5.3 Perceived integrity

The RM ANOVA test indicated a significant main effect of the justification condition on participants perceived integrity of the AI teammate (F (3, 183) = 4.29, $p = .006$, $\eta^2 = 0.066$) such that AI teammates were perceived to be significantly more integral when using the utilitarian justification paradigm ($t = 3.4$, $p_{holm} = 0.005$, CI = [0.17 - 1.34]). The post hoc tests did not indicate significant differences in the means between any other groups ($p(s) > 0.05$). A significant main effect of the outcome condition was observed (F (1, 183) = 53.69, $p < .001$, $\eta^2 = 0.23$) such that AI teammates were perceived to have higher integrity when the outcome was positive compared to when the outcome was negative ($t = 7.33$, $p_{holm} < .001$, CI = [0.86 - 1.51]). The interaction of outcome and justification did not have a significant effect on perceived integrity (F (3, 183) = 1.20, $p = .311$, $\eta^2 = 0.02$). Figure 3 presents descriptive plots capturing the effect and interplay of justification and outcome for the perceived integrity measure. The descriptive plots indicate that providing a justification does impact the perceived Integrity of the AI teammate. Drilling further into the combined effect of justification and outcome, we can see that in the negative condition, providing a justification heightens perceived integrity compared to the no justification case. Planned contrast tests also validate the above observations. In the negative outcome case, when the AI teammate used a framework-driven justification, the perceived integrity of the AI teammate was perceived to be higher than when no justification was used (Deontology: $t = 2.86$, $p = .005$; Utilitarian: $t = 3.177$, $p = .002$; Virtue: $t = 2.112$, $p = .036$). In the positive outcome case, no such variations seem to appear,

also validated by the contrast tests ($p(s) > 0.05$), indicating that justifications may induce a greater effect when the outcome is negative. Looking at the variations for the different justifications across the two outcome conditions, we see that maximum variation was observed in the no justification condition (value = 1.56), followed by virtue (value = 1.33), utilitarianism (value = 1.11), Deontology (value = 0.74). This shows that deontology-based justifications were relatively more robust to changes in the outcome condition than the others, aligned with prior observations.

Overall, we see that AI teammates were perceived to be more moral (across all three dimensions) in the positive outcome condition compared to the negative outcome condition. While the justification condition had a significant main effect in the case of integrity and moral agency measures, it did not yield a similar significant effect for the benevolence measure. Drawing on post-hoc analysis, when justification was significant, we find that having a justification (as opposed to the no justification condition) led to the AI teammate being perceived as more moral. It is interesting to note that no single justification outperforms other justifications across all conditions and measures. Descriptive trends indicate that in the negative outcome condition, AI teammates were perceived to be more moral when using justifications driven by deontological and utilitarian frameworks for all three measures. Across all three measures of morality, virtue-based justifications induced the maximum perceived differences between negative and positive outcome cases. However, it is interesting that the interaction between outcome and rationale was not significant across all three reported outcome measures. In such cases, a planned contrast approach provided some nuances of the underlying interaction and its overall effect. All these planned contrasts highlight the importance of having a justification, especially when the AI's action yields a negative consequence.

6 Discussion

The results of this work demonstrate that while an AI teammate's justifications can impact their perceived morality, that perception is more heavily tied to the consequences resulting from their actions. Indeed, negative consequences negatively affect perceived morality, even if AI teammates are not directly responsible for those negative outcomes. These findings are consistent with the prior literature on trust that has shown robust reliability effects (i.e., a performance-based parameter similar to the outcome factor in the current study) on trust [42]. An immediate takeaway from this work is that research and development needs to continue to engineer AI technologies with greater autonomy that can better reason, adapt and function in certain domains, which will help reduce the negative consequences of AI teammates' actions [12]. This reduction will play the most predominant role in ensuring that humans can perceive AI as moral, ultimately trusting it to operate autonomously in society. The subsequent paragraphs revolve around two key ideas to aid further the ideas mentioned in this paragraph. First is the importance of context in situating when and how justifications are conveyed. This narrative addresses how AI agents can better collaborate in uncertain circumstances, which can help envision robust AI systems in human-AI collaboration. The second is to highlight the fluctuations in perception based on outcome and how that differs across different rationales. Beyond just the design of justifications, such insights have also indicated how observed moral stances and associated societal norms differ between human and non-human actors.

Design tradeoffs in the development and fielding of advanced technologies must consider cost and return on investment (ROI). One such tradeoff consideration is the quintessential question of when an AI should explain. The current data shows that there is little ROI for offering the ethical rationale of an AI following a positive outcome. However, we do see that when the outcome was negative, having a justification did boost the overall perceived moral persona of the AI. This is a crucial insight as regardless of an algorithm's accuracy, AI teammates' actions are bound to be imperfect, especially given the complexities of dynamic real-world environments [5], thus building on the insights from our experiment, AI designers can frame crucial contextual cues that can help maintain team dynamics between human and AI teammates in such conditions. It is important to note that explanations can be useful even when no errors have occurred, but rather, the AI does something unexpected [43]. Fortunately, the results of this work highlight that AI teammates can be designed to mitigate the negative impacts of these mistakes on the perceived morality of the AI. In particular, designing an AI teammate with the ability to relay an ethical rationale can help mitigate the impacts that negative consequences have on perceived morality. Indeed, while this explanation is no substitute for a positive ethical outcome, a semblance of perceived morality can be perceived from AI that makes decisions accompanied by an ethical rationale. In particular, results show that AI teammates who convey an ethical rationale grounded in a deontological ethical framework had the greatest chance of being perceived as moral when their actions ultimately led to a negative consequence. Particularly salient are the variations between the different justifications across the contextual outcomes. While deontology is stable compared to the other justifications, virtue-based justifications have maximum variance between positive and negative outcome conditions. This indicates latent expectations ascribed to

AI teammates to adhere to social norms; even if the consequences have negative ramifications, adherence to societal standards and collective conventions heightens perceived morality. In contrast, individualistic characteristics (as manifested through virtue) may act as a double-edged sword. For example, if the contextual outcome is positive, such characteristics greatly increase perceived morality, yet they can have adverse effects in negative cases. This connects with how fundamental attribution errors may impact AI teammates' perceptions [44]. In such cases, the error source may be connected with contextual parameters.

In turn, the results of this work can make the recommendation that the algorithmic underpinnings of AI teammates need to be continuously improved to maximize positive outcomes; however, these improvements also need to be accompanied by the ability for AI to provide ethical rationales alongside their decisions to weather the impacts of mistakes and negative consequences. Such communicative acts can serve as pivotal pillars of collaboration, especially in uncertain terrains or where the operational hazard is fraught with multiple layers of uncertainty and cannot be controlled a priori. Based on the above claim, considering the actual design of these justifications is important [45]. The results portray the importance of contextual impulses and indicate that rationales need to be carefully situated to address contextual stimuli [46]. We can see that when negative ramifications are associated with the actions of the AI teammate, even if said negative consequence is not their intention, the AI teammate is perceived as less morally sound and capable. Yet, attributors still posit the root of the error to be grounded in the disposition of the autonomous agent. Such insights raise questions about the way in which autonomous actors are profiled and compared with human actors in ethical contexts [47], while also providing novel insights towards how ethical rationales are perceived when considering non-human entities.

The above issue goes deeper than just proposing that AI-based be designed to be as reliable as possible. In contrast, designers and organizations that seek to field AI-based systems must critically analyze the contexts in which they plan to use the technologies to understand what is a "positive" or "negative" outcome. Such considerations are both culturally-infused (i.e., the value and valence of action may vary across cultural groups), and may vary based on societal norms that propagate asymmetries between how humans and machines are viewed in identical circumstances. The success of HATs in supporting ethical dilemmas will depend, in part, on how well the research community understands and conveys these nuisances, which will have clear implications for design. In the current study, there appeared to be greater stability in viewing an AI as a deontological decision-maker, and the outcome appeared to be the most salient element of the process. How well do these findings hold in non-western cultures? How would human referents fare in the same circumstances? Clearly, using HATs in ethical dilemmas warrants future research to isolate and expand our understanding of how to facilitate an AI as an ethically-competent associate best (See Fig. 4).

Fig. 4 Abstract representation of how ethical frameworks and decision outcomes can influence the perceived morality of AI teammates

7 Conclusion and future directions

While critical to society and the design of AI teammates, all these observations provide an essential iteration on the domain of resilience engineering, especially in human-AI teaming. Technology (AI) may be imperfect within teams, and recovery mechanisms must be instituted during design to enable humans to withstand these imperfections [48, 49]. To promote cooperative team outcomes, more robust frameworks are needed to help mitigate the impacts of these imperfections [50, 51]. Contextual uncertainties are such that no system can fully cope with all possible variations. For AI teammates, as technology increases in autonomy, the issue of predictability increases in importance [24], and one way to promote predictability is to clarify the ethical stance that guides the technology's behavior. Indeed, based on the results of this work, providing AI teammates with an ethical framework during communication, especially deontology, significantly contributes to the resilience of human perception when negative consequences occur. Communication between team members is key to sustaining teamwork and nurturing team resilience [52]. Given that shared ethical

ideologies may obfuscate the many motives, intentions and associated actions of human and non-human actors, crafting communicative imperatives that convey ethical viewpoints can be pivotal for human-AI teams to sustain across the unpredictable nature of real-world environments. Our insights provide important fundamental pathways toward how such communicative visions can be bolstered to boost ethical parity and shared understanding between teammates, especially in human-AI teams.

While the above highlights this work's immediate contributions, it is also essential to discuss how this research can be expanded and integrated into future research. First, this study viewed consequence as a fairly binary outcome; yet, the natural world contains a much greater degree of nuance, and the outcomes of one's actions are not solely positive or negative. Thus, future efforts should further explore the impact of ethical frameworks in human-AI teams where the outcome of a teammate or team's actions is not solely positive or negative [53, 54]. Second, this study's presentation of deontology and other ethical frameworks was contextualized, but the application of ethical frameworks changes based on the context [55]. In turn, future research should further explore the contextual factors that interact with various ethical frameworks and explore alternative types of deontology that prioritize different societal or team principles. Further research can help to unpack perceptions of virtue in AI teammates and how these perceptions differ between human and non-human actors. Lastly, based on the results of this work, future research efforts in human-AI teaming that explore ethics should operationalize and design AI teammates with ethical frameworks in mind. Indeed, as the domain progresses, integrating ethical frameworks into research efforts will enable the study of more capable and highly perceived AI teammates who have to navigate ethical decisions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43681-024-00515-5.

References

1. Makridakis, S.: The forthcoming artificial intelligence (Ai) revolution: Its impact on society and firms. Futures. **90**, 46–60 (2017)
2. Yu, K.-H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare. Nat. Biomedical Eng. **2**(10), 719–731 (2018)
3. Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.
4. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's, (2018).
5. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, vol. 2156. Springer, (2019).
6. McNeese, N.J., Flathmann, C., O'Neill, T.A., Salas, E.: Stepping out of the shadow of human-human teaming: Crafting a unique identity for human-autonomy teams. Comput. Hum. Behav. **148**, 107874 (2023)
7. Cooke, N.J., Lawless, W.F.: Effective human–artificial intelligence teaming. Syst. Eng. Artif. Intell., 61–75 (2021)
8. Salas, E., Burke, C.S., Cannon-Bowers, J.A.: Teamwork: Emerging principles. Int. J. Manage. Reviews. **2**(4), 339–356 (2000)
9. Mattingly, C., Throop, J.: The anthropology of ethics and morality. Annu. Rev. Anthropol. **47**, 475–492 (2018)
10. Sawant, S., Mallick, R., McNeese, N., Madathil, C.: K.: Mutually beneficial decision making in human-ai teams: Understanding soldier's perception and expectations from ai teammates in human-ai teams. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 66, pp. 287–289 SAGE Publications Sage CA: Los Angeles, CA (2022)
11. McNeese, N.J., Demir, M., Cooke, N.J., Myers, C.: Teaming with a synthetic teammate: Insights into human-autonomy teaming. Hum. Factors. **60**(2), 262–273 (2018)
12. O'Neill, T.A., Flathmann, C., McNeese, N.J., Salas, E.: 21st century teaming and beyond: Advances in human-autonomy teamwork. Comput. Hum. Behav. **147**, 107865 (2023)
13. Sewell, G.: Doing what comes naturally? Why we need a practical ethics of team-work. Int. J. Hum. Resource Manage. **16**(2), 202–218 (2005)
14. Textor, C., Zhang, R., Lopez, J., Schelble, B.G., McNeese, N.J., Freeman, G., Pak, R., Tossell, C., Visser, E.J.: Exploring the relationship between ethics and trust in human–artificial intelligence teaming: A mixed methods approach. J. Cogn. Eng. Decis. Mak. **16**(4), 252–281 (2022)
15. Lopez, J., Textor, C., Lancaster, C., Schelble, B., Freeman, G., Zhang, R., McNeese, N., Pak, R.: The complex relationship of Ai ethics and trust in human– ai teaming: insights from advanced real-world subject matter experts. AI Ethics, pp. 1–21 (2023)
16. Schelble, B.G., Lopez, J., Textor, C., Zhang, R., McNeese, N.J., Pak, R., Freeman, G.: Towards ethical Ai: Empirically investigating dimensions of ai ethics, trust repair, and performance in human-ai teaming. Hum. Factors, **00187208221116952** (2022)
17. M˚aseide, P.: Morality and expert systems: Problem solving in medical team meetings. Behav. Inform. Technol. **30**(4), 525–532 (2011)
18. Flathmann, C., Schelble, B.G., Zhang, R., McNeese, N.J.: Modeling and guiding the creation of ethical human-ai teams. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 469–479 (2021)
19. Momen, A., De Visser, E., Wolsten, K., Cooley, K., Walliser, J., Tossell, C.C.: Trusting the moral judgments of a robot: perceived moral competence and humanlikeness of a gpt-3 enabled ai (2023)
20. Malle, B.F., Phillips, E.: A robot's justifications, but not explanations, mitigate people's moral criticism and preserve their trust (2023)
21. Hummels, H., De Leede, J.: Teamwork and morality: Comparing lean production and sociotechnology. J. Bus. Ethics. **26**, 75–88 (2000)
22. Gunia, A., Sowltysik, M., Jarosz, S.: Robot ethics and artificial morality. In: Artificial Intelligence, Management and Trust, pp. 127–143. Routledge,??? (2024)
23. Demir, M., Likens, A.D., Cooke, N.J., Amazeen, P.G., McNeese, N.J.: Team coordination and effectiveness in human-autonomy teaming. IEEE Trans. Human-Machine Syst. **49**(2), 150–159 (2018)
24. Lyons, J.B., Sycara, K., Lewis, M., Capiola, A.: Human–autonomy teaming: Definitions, debates, and directions. Front. Psychol. **12**, 589585 (2021)
25. McNeese, N.J., Demir, M., Chiou, E.K., Cooke, N.J.: Trust and team performance in human–autonomy teaming. Int. J. Electron. Commer. **25**(1), 51–72 (2021)

26. Zhang, R., McNeese, N.J., Freeman, G., Musick, G.: an ideal human expectations of ai teammates in human-ai teaming. Proc. ACM Hum. Comp. Interact. **4**(CSCW3), 1–25 (2021)
27. Jiang, J., Karran, A.J., Coursaris, C.K., L´eger, P.-M., Beringer, J.: A situation awareness perspective on human-ai interaction: Tensions and opportunities. Int. J. Human–Computer Interact. **39**(9), 1789–1806 (2023)
28. Ulfert, A.-S., Georganta, E., Centeio Jorge, C., Mehrotra, S., Tielman, M.: Shaping a multidisciplinary understanding of team trust in human-ai teams: A theoretical framework. Eur. J. Work Organizational Psychol., 1–14 (2023)
29. Ezer, N., Bruni, S., Cai, Y., Hepenstal, S.J., Miller, C.A., Schmorrow, D.D.: Trust engineering for human-ai teams. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 63, pp. 322–326 SAGE Publications Sage CA: Los Angeles, CA (2019)
30. Hauptman, A.I., Schelble, B.G., McNeese, N.J.: Adaptive autonomy as a means for implementing shared ethics in human-ai teams. In: Proceedings of the AAAI Spring Symposium on AI Engineering, pp. 1–7 (2022)
31. Zhang, R., Duan, W., Flathmann, C., McNeese, N., Freeman, G., Williams, A.: Investigating ai teammate communication strategies and their impact in human-ai teams for effective teamwork. Proc. ACM Hum. Comput. Interact. **7**(CSCW2), 1–31 (2023)
32. Caples, S.C., Hanna, M.E., Phelps, L.: Linking ethics decisions to philosophical rationales: An empirical study. J. Legal Ethical Regul. Isses. **11**, 93 (2008)
33. Verg´es, A.: Integrating contextual issues in ethical decision making. Ethics Behav. **20**(6), 497–507 (2010)
34. Luccioni, A., Bengio, Y.: On the morality of artificial intelligence. arXiv preprint arXiv:1912.11945 (2019)
35. Banks, J.: A perceived moral agency scale: Development and validation of a metric for humans and social machines. Comput. Hum. Behav. **90**, 363–371 (2019)
36. McFall, L.: Integr. Ethics. **98**(1), 5–20 (1987)
37. Arieli, S., Grant, A.M., Sagiv, L.: Convincing yourself to care about others: An intervention for enhancing benevolence values. J. Pers. **82**(1), 15–24 (2014)
38. D¨orr, K.N., Hollnbuchner, K.: Ethical challenges of algorithmic journalism. Digit. Journalism. **5**(4), 404–419 (2017)
39. Dean, R.K., Pollard, R.Q. Jr.: Context-based ethical reasoning in interpreting: A demand control schema perspective. Interpreter Translator Train. **5**(1), 155–182 (2011)
40. Flathmann, C., Schelble, B.G., Rosopa, P.J., McNeese, N.J., Mallick, R., Madathil, K.C.: Examining the impact of varying levels of ai teammate influence on human-ai teams. Int. J. Hum. Comput. Stud. **177**, 103061 (2023)
41. Mayer, R.C., Davis, J.H.: The effect of the performance appraisal system on trust for management: a field quasi-experiment. J. Appl. Psychol. **84**(1), 123 (1999)
42. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. Hum. Factors. **53**(5), 517–527 (2011)
43. Lyons, J.B., Hamdan, I., Vo, T.Q.: Explanations and trust: What happens to trust when a robot partner does something unexpected? Comput. Hum. Behav. **138**, 107473 (2023)
44. Harvey, J.H., Town, J.P., Yarkin, K.L.: How fundamental is the fundamental attribution error? J. Personal. Soc. Psychol. **40**(2), 346 (1981)
45. Visser, E., Parasuraman, R.: Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. J. Cogn. Eng. Decis. Mak. **5**(2), 209–231 (2011)
46. Khan, A.A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., Akbar, M.A.: Ethics of ai: A systematic literature review of principles and challenges. In: Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, pp. 383–392 (2022)
47. Sundvall, J., Drosinou, M., Hannikainen, I., Elovaara, K., Halonen, J., Herzon, V., Kopecky`, R., Ko˘sov´a, J., Koverola, M., Kunnari, M.: Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas. Eur. J. Social Psychol. **53**(4), 779–804 (2023)
48. Woods, D.D.: Four concepts for resilience and the implications for the future of resilience engineering. Reliab. Eng. Syst. Saf. **141**, 5–9 (2015)
49. Hollnagel, E., Woods, D.D., Leveson, N.: Resilience Engineering: Concepts and Precepts. Ashgate Publishing, Ltd.,??? (2006)
50. Van Bossuyt, D.L., Papakonstantinou, N., Hale, B., Salonen, J., O'Halloran, B.: Model based resilience engineering for design and assessment of mission critical systems containing artificial intelligence components. In: Artificial Intelligence and Cybersecurity: Theory and Applications, pp. 47–66. Springer,??? (2022)
51. Pawar, B., Park, S., Hu, P., Wang, Q.: Applications of resilience engineering principles in different fields with a focus on industrial systems: A literature review. J. Loss Prev. Process Ind. **69**, 104366 (2021)
52. Alliger, G.M., Cerasoli, C.P., Tannenbaum, S.I., Vessey, W.B.: Team resilience. Organ. Dyn. **44**(3), 176–184 (2015)
53. Vilanilam, G.C., Venkat, E.H.: Ethical nuances and medicolegal vulnerabilities in robotic neurosurgery. NeuroSurg. Focus. **52**(1), 2 (2022)
54. Beil, M., Proft, I., Heerden, D., Sviri, S., Heerden, P.V.: Ethical considerations about artificial intelligence for prognostication in intensive care. Intensive Care Med. Experimental. **7**(1), 1–13 (2019)
55. Conway, P., Gawronski, B.: Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. J. Personal. Soc. Psychol. **104**(2), 216 (2013)

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com