AI and Ethics
https://doi.org/10.1007/s43681-023-00303-7

ORIGINAL RESEARCH

The complex relationship of AI ethics and trust in human–AI teaming: insights from advanced real-world subject matter experts

Jeremy Lopez[1] · Claire Textor[1] · Caitlin Lancaster[2] · Beau Schelble[2] · Guo Freeman[2] · Rui Zhang[2] · Nathan McNeese[2] · Richard Pak[1]

Received: 27 March 2023 / Accepted: 18 May 2023

Abstract
Human-autonomy teams will likely first see use within environments with ethical considerations (e.g., military, healthcare). Therefore, we must consider how to best design an ethical autonomous teammate that can promote trust within teams, an antecedent to team effectiveness. In the current study, we conducted 14 semi-structured interviews with US Air Force pilots on the topics of autonomous teammates, trust, and ethics. A thematic analysis revealed that the pilots see themselves serving a parental role alongside a developing machine teammate. As parents, the pilots would feel responsible for their machine teammate's behavior, and their unethical actions may not lead to a loss of trust. However, once the pilots feel their teammate has matured, their unethical actions would likely lower trust. To repair that trust, the pilots would want to understand their teammate's processing, yet they are concerned about their ability to understand a machine's processing. Additionally, the pilots would expect their teammates to indicate that it is improving or plans to improve. The findings from this study highlight the nuanced relationship between trust and ethics, as well as a duality of infantilized teammates that cannot bear moral weight and advanced machines whose decision-making processes may be incomprehensibly complex. Future investigations should further explore this parent–child paradigm and its relation to trust development and maintenance in human-autonomy teams.

Keywords Human–AI teaming · Artificial intelligence · Ethics · Trust · Applied · Trust repair

1 Introduction

Human-autonomy teams (HATs) are a rapidly growing research agenda in HCI research on human–AI interaction [60, 67, 74, 80]. These HATs are being developed for and applied to various contexts, including fashion, manufacturing, and public safety [107]. In these teams, autonomous agents are unique from automated technologies in that they serve a more independent role with a significant degree of agency [74], making AI technology more of a teammate than merely a tool. Therefore, with the use of HATs expected to rise in future [27], it is essential to investigate how to design autonomy to be a suitable teammate to humans for supporting positive outcomes of such teamwork, for example, how autonomous teammates can participate in HATs whose actions will bear ethical consequences.

We believe that an in-depth investigation of the ethical aspects of HATs is crucial. First, existing organizations are currently working toward developing artificial intelligence (AI) for applications with ethical implications. For example, in the medical field, AI is teaming with medical professionals to support, diagnose and treat humans. While AI is used for seemingly neutral tasks, such as scheduling appointments, it is also involved with potentially consequential tasks like drug dose algorithm creation that places considerable ethical weight on the engineers who create the system, the doctors who use it, and, perhaps, even the system itself [62]. Second, as autonomous teammates (ATs) are placed in roles that require ethical considerations, their decisions and actions may influence human teammates' trust in an AT. Understanding the dynamics between ethics and trust within HATs is essential, given their high levels of agency and independence. Trust is a core component of teamwork and is associated with team performance [66] and teamwork effectiveness [114] within HATs. Despite this early progress,

✉ Beau Schelble
bschelb@clemson.edu

[1] Clemson University Psychology Department, 321 Calhoun Drive, Clemson, SC 29634, USA

[2] Clemson University Division of Human-Centered Computing, 821 McMillan Road, Clemson, SC 29634, USA

Published online: 05 July 2023

there are major gaps and topic areas in the HAT space which have yet to be explored. For example, while some work has investigated ethical judgments of autonomous agents (e.g., [16]), findings have generally been theoretically constructed and do not include the perspectives of persons who have worked with autonomy. It is important to consider the perspectives of these persons to better understand their expectations of an AT and any potential barriers that may impede the development of an effective HAT. The present research is also limited regarding how an autonomous teammate's ethicality influences trust in practice, necessitating an investigation into which factors influence human teammates' ethical judgments of their ATs. Lastly, various methods for repairing trust have been studied and proposed (e.g., [88, 103]), but there have yet to be investigations of how to possibly repair trust damaged by the ethicality of an autonomous teammate's actions.

To address these research gaps, in this paper, we investigate the following research questions by conducting 14 in-depth interviews with Air Force members who pilot the Lockheed Martin F-35A Lightning II:

RQ1: How do human teammates perceive their autonomous teammates' roles in ethical or unethical actions?

RQ2: What are the barriers to human teammates' acceptance of an ethical autonomous teammate, especially when ethical decisions are involved?

RQ3: If trust in an autonomous teammate is influenced by the ethicality of its actions, how can damaged trust be repaired?

We chose to use Air Force members' experiences and perceptions of HATs as our research context to address these research questions for two distinct reasons. First, Air Force pilots, especially F-35A pilots, regularly interact with highly automated technologies on a regular basis. For example, the F-35A is equipped with the Autonomic Logistics Information System, a system containing early forms of artificial intelligence (AI) capable of making decisions on its own, and future plans include creating an entirely autonomous system with the aircraft [75]. Furthermore, advancements in truly autonomous teaming have been primarily applied in military settings where necessary security, resource, and funding support are available, making these individuals primed for discussing these concepts as a tangible reality. Second, Air Force pilots consistently work in high-stakes dynamic environments. As we will detail in our findings, F-35A pilot squads often face unexpected, time-sensitive dilemmas where they must make their own decisions while considering the potentially lethal consequences of their actions. Taken together, these elements ground our findings in real-world applications in which this research is urgently needed. Unlike other studies on HATs that commonly rely on college students and other populations who do not have real-world experience with highly automated systems or environments that required ethical decision-making, this research is informed by experienced individuals and offers recommendations for the future of HATs grounded in the everyday experiences of those faced with ethical decisions involving highly advanced technological systems.

Given the grounded nature of our work, this study provides two primary contributions to existing research on ethical HATs. First, discussions of making ethical AI tend to adopt a highly theoretical deductive approach (e.g., [16, 41, 112]), whereas the current study is one of the first to provide empirical evidence that relies on the unique insights of Air Force pilots to explore how to design an ethical autonomous teammate. Gaining the perspectives of highly experienced professionals will provide support for existing and future theories, resulting in a robust understanding of ethics within HATs while also confirming and extending existing discussions. Second, the current study provides an in-depth exploration of the relationship between ethics and trust from a population that often makes decisions with ethical ambiguity. The findings from our interviews convey the complexity of developing trust in a teammate when the risks of failure have serious implications. For some situations, performing an unethical action may be the best option at a given time. Situations also exist where the ethicality of an action can influence trust, regardless of how team performance is impacted. This disconnect between trust and performance should lead the field to reconsider our conceptualizations of trust, many of which are directly tied to performance outcomes (e.g., [54, 66]).

2 Related work

2.1 Ethical considerations in human-autonomy teams (HATs)

Traditionally, human–machine interaction was primarily discussed within industrial domains such as production and transportation. Machines were originally operated by trained professionals, acting as tools to maximize efficiency and productivity. Over time, the role of humans in these interactions has shifted from being a controller to primarily supervisory in nature. Operators have become less involved in task completion in favor of monitoring system states. The transition from automation to autonomy has further decreased human involvement in task completion. While this role change is associated with a reduction in situation awareness [27], autonomous systems are defined by a higher level of independence and self-governance [43,

103], features that are comparable to human teammates. The development of artificial intelligence has also greatly impacted human–machine interaction, leading to what some refer to as the fourth industrial revolution [24, 72]. Artificial intelligence has given machines the ability to sense the environment and alter their behavior based on changing or novel information. These dynamic capabilities represent a critical change in how machines can be characterized in relation to their human counterparts: a shift from tool to teammate.

Under this new paradigm, humans will work alongside technology that is capable of making its own decisions and accepting the responsibilities of a teammate. Although AI researchers seem confident that technology can act as a teammate, it is unclear if people who will be members of HATs are willing to consider a machine as a teammate. In addition to theorizing about creating the ideal AT, it will also be important for the field to incorporate input from the persons who will likely be the first to be part of a HAT (e.g., oncologists and military personnel). One possibility is that humans will be unwilling to adopt a machine teammate. For example, military officers interviewed on the topic of autonomy stated that a human's ability to override an autonomous system would influence their ability to trust in the system [106]. While it may be important to consider the possibility of overriding a system when a machine is a tool, overriding a teammate may have more impactful consequences. Requesting the ability to override an autonomous teammate may indicate an expectation or preference to continue acting as a supervisor to a machine despite its capabilities of acting as a full teammate, which may have its own influence on the HAT's efficacy. By exploring humans' current expectations of autonomous systems as teammates, we can determine what types of roles human teammates will expect to fulfill and predict potential pitfalls as machines become teammates.

As humans' roles change and technology advances, some important elements with current technologies will likely continue to be relevant, and formerly irrelevant elements may become relevant. For example, trust in automation has been a relevant topic of research since its introduction several decades ago [71]. Trust has remained relevant as technology has advanced from a decision aid for perceptual tasks [25] to personnel selection [52]. In contrast, a factor like a machine's ethicality will become important as machines satisfy tasks with ethical consequences. Importantly, this topic concerns the ethicality of an agent, not the ethicality of humans implementing AI or automation, a topic with extensive literature (e.g., [53, 80, 101]). Machine ethicality will become important not only for the new tasks it will satisfy but also because we are considering the possibility of HA teaming. Therefore, an AT's ethicality could influence human teammates' perception of the machine, which could negatively impact team effectiveness.

In this work, we adopt a conceptualization of ethics as a general framework for understanding and examining moral life [8]. This domain-general definition allows for individual interpretation without subscribing to specific ethical frameworks (e.g., virtue ethics, utilitarianism, deontology; see Three Moral Theories [34]). This definition allows us to consider the multi-faceted nature of ethical judgments, which are "an individual's personal evaluation of the degree to which some behavior or course of action is ethical or unethical" [94]. Ethical judgments are highly personal and influenced by one's characteristics [94], including age [13, 78], education [13, 30], and ethical awareness [5, 99]. Ethical judgments are also influenced by one's ethical environment, which includes pressures from the workplace and organizational climates [59, 87]. These external pressures can influence one's perception of the moral intensity and perspicuity of an ethical problem [4, 51, 96]. However, the organizational climate can also lead organization members to feel pressured to perform unethical actions or ignore unethical actions performed by others [19]. Thus, members within an organization must contend with their personal morals, ethical organizational pressures, and non-ethical organizational pressures as they perform ethical judgments.

The advent and development of AI systems have driven efforts to define ethical principles by which they should be governed. However, AI presents challenges that make it difficult to prescribe a unified set of guidelines. Mittelstadt addresses why AI is set apart from other domains, such as medicine, in setting ethical standards [70]. For example, unlike medicine which is primarily guided by promoting patient well-being, the goals of AI are not unified. AI systems are often developed with organizational interests taking precedence over end-user considerations. Another factor that complicates the development of ethical AI is the challenge of putting high-level principles into practice. While many sets of AI ethical guidelines concur that AI should uphold principles, such as transparency and non-maleficence, there is disagreement on how they should be implemented [41]. Resulting in certain actions being perceived as perfectly ethical by some but unethical by others. These challenges are exacerbated by the fact that AI capabilities and applications are constantly expanding.

As AI systems continue to develop, the focus on their ethical implications is also increasing. AI systems are used to inform decision-making and supplement human capabilities in a variety of domains, such as law enforcement, medical diagnosis, and finance. Often, the focus is placed on the technical aspects of how systems can be implemented with less emphasis on unintended ethical implications. However, these ethical implications can greatly impact how people perceive and interact with AI. Ethics is often cited as a concern when discussing openness to and acceptance of AI systems [26, 49, 76]. Additionally, as AI systems become more

autonomous, the role of the human will likely be altered such that involvement and oversight are reduced. As with automated systems, autonomous systems pose concerns related to issues, such as situation awareness, the out-of-the-loop performance problem, and trust [27]. Autonomy changes human involvement and oversight, raising concerns about the ethical implementation of AI systems [56]. These considerations and challenges persist as AI is developed for more involved and complex roles, such as teaming.

2.2 Trust within human-autonomy teams (HATs)

While research on trust has its roots in human interpersonal trust literature, we expound upon this understanding to look at trust as a function of perceived or actual interpersonal dynamics on human-autonomy teams. While human–human manifestations of trust are different, the teaming dimensions make this literature relevant to our discussions and our participants' later evaluations of their trust in these agents. Therefore, we adopt Lee and See's [53, 54] definition of trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability," which was originally applied to human-automation trust but has also been used when discussing trust in autonomous systems (e.g., [15, 20]). This distinction between automation and autonomy is critical in understanding trust calibration, which can differ based on the type of agent [61]. For example, Madhavan and Wiegman posit that trust in humans is initially low and increases over time with demonstrated displays of competency and efficacy [58]. Conversely, initial trust in automated systems tends to be high (i.e., automation bias) and is damaged by behavior which violates expectations. Inappropriately calibrated trust (i.e., over-or under-trust) can result in disuse, misuse, or abuse of automation [77]. A great deal of research has investigated how trust is calibrated in the realms of human-automation interaction [3, 68], which has expanded recently to include autonomous systems [18, 81, 104].

Early research has demonstrated the importance of trust calibration in HATs as it relates to appropriate usage and other outcomes, such as performance. For example, McNeese and colleagues had teams work with a synthetic pilot teammate in a simulated RPAS task environment, finding that low-performing teams trusted the autonomous teammate more compared to medium- or high-performing teams [66]. Another study by Chen and colleagues investigated the impact of transparency on HAT effectiveness [12]. Participants interacted with an autonomous squad member, which provided information about the environment, the rationale for its decisions, and projections of future states. They found that increased transparency was associated with better performance and greater trust in the agent. However, trust failed to increase when the autonomous teammate expressed uncertainty. Appropriately calibrated trust is a reflection of an agent's true capabilities, avoiding under- or over-trust. As with any technology, autonomous systems cannot be expected to be perfectly reliable all the time, meaning investigation must also be done into how trust can be repaired.

2.2.1 Trust repair

Early studies revealed that trust repair strategies like apologies and denials can effectively repair trust in human trustees [47, 48]. However, the effectiveness of a trust repair strategy depends on various factors, including the nature of the trust violation. For example, apologies have been found to be more effective when a trustee attributes a fault to their incompetency, whereas denials are more effective when a trustee attributes a fault to their integrity [48]. Subsequent research identified additional factors that influence trust repair efficacy, including internal/external blame attribution [47], individual differences [29], and perceived repentance [23]. More recently, trust repair research has extended to human–machine interaction with mixed results. For instance, apologies from an automated system are more effective when the system commits a competency rather than integrity violation [50, 79], replicating findings from interpersonal trust repair literature. However, denials have been shown to be equally effective between competency and integrity violations [79], a departure from previous findings. In contrast, a human–robot interaction study replicated the apology-competence and denial-integrity interpersonal trust repair findings [88].

The current state of research indicates that there may be a discrepancy between trust repair with automated systems and those with greater autonomy. Robots and autonomous systems may one day participate in teaming contexts [35], changing the fundamental paradigm of human–machine interaction from technology as a tool to technology as a teammate [74]. It is possible that this paradigm shift will also influence the effectiveness of trust repair. Additionally, the paradigm shift may demand additional trust repair strategies beyond apologies and denials (e.g., blame, explain, and promise; [103]). Research on new methods for trust repair will become critical as advanced systems (e.g., autonomy) participate in teaming contexts that demand considerations not present in human-automation interaction.

2.2.2 Acceptance

Successfully building, maintaining, and repairing trust can help to foster automation acceptance. The connection between trust and automation acceptance has been found across different automation types, such as decision [25] and diagnostic aids [55]. The factors which influence automation acceptance are related to those which affect trust.

The Technology Acceptance Model (TAM) posits that the two main factors which influence attitudes toward (and subsequent usage of) technology are perceived usefulness and ease of use [17]. Automation failure may negatively impact both of these perceptions while also damaging trust. However, demonstrated behavior changes or insight into an automated system's true capabilities may help repair trust and foster acceptance. Conversely, acceptance of AI may be more challenging if a system's underlying processes are complex and it cannot easily demonstrate improved performance or provide insight into its decision-making processes. While the promise of AI is vast, concerns about acceptance must be addressed.

Early research on AI suggests that widespread adoption of AI can lead to various societal benefits, including increased productivity [1], improved health outcomes (e.g., AI for cancer screening; [65, 100]), and facilitated education [69, 113]. However, there are concerns associated with AI acceptance and implementation. For example, AI making its own decisions (e.g., autonomy) creates legal and moral issues where it is unclear who should take the responsibility for an AI's actions [10]. Similarly, even if an AI is programmed to reach a desired goal, it may achieve that goal in an undesirable manner (e.g., harming humans; [93]). Finally, AI often relies on black box methods (e.g., artificial neural networks, deep learning), making it difficult, if not impossible, for humans to understand how an AI makes a decision. As AI capabilities continue to increase, systems may complete actions that defy human intelligence and rationale, which can serve as another barrier to the acceptance of AI [44].

The perspectives of human collaborators must be considered to responsibly develop autonomous teammates. While few truly autonomous teammates currently exist, early examples have been implemented in military settings (e.g., Boeing's Loyal Wingman within the military domain). Actions deemed unethical can result in unintentional conflict escalations, damage to physical infrastructure, loss of life, and global instability. For an autonomous teammate to foster acceptance, it must have some comprehension of human values to interpret human desires correctly [44]. This requires the autonomous teammate to not only understand a human's goal but also their preferences for how an action should be performed. An autonomous teammate with ethical guidelines can increase reliance and acceptance [44]. Including future human collaborators in the process of developing an autonomous teammate's ethical guidelines will likely lead to an overall more successful autonomous teammate, as has been shown with the development of AI within human resources [97]. Therefore, the successful development of an ethical autonomous teammate requires the inclusion of the perspectives of humans, which will 1 day be teaming with these advanced systems.

2.3 Intersection of ethicality and trust in human-autonomy teams (HATs)

Grounded in prior literature on both ethical and trust implications in HATs and with the use of HATs expected to rise in future [27], it is necessary to understand how an autonomous teammate's perceived ethicality may influence trust in that teammate. Ethical judgments play a major role in the development and sustainment of trust between persons [39, 108, 109]. Similarly, interpersonal interactions are influenced by one's ethical judgments of others' actions [42, 45, 95]. Therefore, the ethical judgments of an autonomous teammate may influence how human teammates interact with and trust the autonomous teammate. Initial research on trust and ethics within HATs suggests that unethical actions often, but not always, reduce trust in an autonomous teammate [86, 99]. However, it is unclear how human teammates are making ethical judgments of their autonomous teammates. As discussed above, ethical judgments are influenced by internal and external factors [94], with the external factors varying by domain. For example, military organizations promote adherence to ethical standards to establish a strong ethical climate and limit a team's acceptable courses of action [36, 46, 57, 64, 92]. Given a strong ethical climate, it is possible that the ethical judgments of an autonomous teammate will mostly be influenced by organizational factors.

Given the increased capabilities of autonomous systems, one such consideration is that autonomy may eventually be included in teams that perform ethical decision-making. No system is infallible, so an autonomous teammate may make decisions that its human teammates consider to be unethical. The current research on trust in human–machine interaction does not allow us to predict how unethical actions may influence trust. Indeed, current models posit that system performance is a key determinant of trust, but performance is often described in relation to how well the system helps a trustor achieve a goal [40, 54]. With ethical decision-making, it is possible that there are multiple methods with varying ethicalities that will satisfy the trustor's goal to equal levels. Therefore, it is possible for an autonomous teammate to perform an unethical action that meets or exceeds the expected performance. Furthermore, if unethical actions damage trust, research has yet to address how trust damaged by unethical actions can be repaired. It is possible that ethics-damaged trust may require trust repair strategies unique to ethical actions. These research questions must be addressed before autonomy can be introduced into these teaming contexts.

3 Methods

3.1 Recruitment and participants

To recruit expert users of advanced AI systems, we collaborated with the United States Air Force Academy to distribute a recruitment email to leaders of various Air Force groups. As a result, fourteen F-35A pilots were recruited for our study. After recruiting the fourteen pilots, we recognized that they all identified as white males. We communicated our desire to recruit a more diverse sample, but our contacts at the Air Force Academy were unable to help us satisfy this request. Ranks of the participants varied from First Lieutenant to Lieutenant Colonel, years of service ranged from less than 1–20 years of experience within the Air Force, and years flying ranged from about 1 year to over 18 years of flight experience. Two of the fourteen participants are not active military but fly in reserve units. Stated positions or assignments varied by the participant, with three identifying as F-35A flight instructors, seven identifying as pilots, two identifying as flight leads, and one identifying as a wingman. The pilots all indicated using AI technologies onboard F-35A aircraft, which is extensive and one of the most advanced examples of autonomous systems that humans interact with and rely upon.

3.2 Interviews

Prior to the interview, participants were mailed an informed consent document and a short introductory message describing the purpose of the interview. This short description contains definitions for key concepts in this research: autonomy, trust, ethics, and human-autonomy teams (see Table 1). The goal of this description is to standardize our respondents' conceptualizations of the key concepts in this study.

Interviews were conducted and recorded over video conferencing software and lasted between 45 and 60 min. Interviews began by asking participants about their rank and experiences while serving in the Air Force. To ensure the participants understood the terms used in the interview, researchers read the above definitions for the participant before proceeding to the interview questions. Next, participants were asked to provide an example of an autonomous teammate, either from their own experiences or a future hypothetical system. This example was used as the subject for future questions. For example, if a pilot provided the example of an autonomous wingman, we would ask the question: "Can you think of an unethical behavior that an autonomous wingman could commit?" We believe that this allowed participants to easily provide examples of unethical behaviors, which we could then probe to determine the underlying moral principles being violated. Following this initial portion, we asked participants to provide exemplary unethical behaviors the hypothetical autonomous teammate could perform, with probing questions regarding why that action would be considered unethical, how the participant would feel if an AT performed that action, how that action may influence trust, and how trust may be repaired (if the participant indicates that trust would be harmed). To determine if there are any contextual elements to this unethical behavior, we asked if there could be any situations where the action could be considered ethical or morally neutral. If the participant provided answers in the affirmative, we provided the same probes regarding the expected effect on trust and the possibility of trust repair. This line of questioning allowed us to discern actions that would invoke moral outrage from those that the pilot merely disagrees with, providing us with potential differences in how trust is influenced and how trust can be repaired for different ethical violations. When suitable, we would include additional questions regarding (1) their expected role within a HAT, (2) the onus for an AT's unethical actions, (3) how an AT's ethicality may influence team effectiveness, (4) the importance of having an AT whose ethical guidelines and decisions align with human teammates', and (5) the potential tradeoff between maintaining moral character in lieu of influencing team effectiveness.

3.3 Data analysis

We performed a thematic analysis [9, 91] to explore how a hypothetical autonomous teammate's ethical actions influence trust in that autonomous teammate and how ethics-damaged trust can be repaired. Following these guidelines

Table 1 Autonomy, trust, and ethics definitions for use in interview sessions

Term	Definition
Autonomy	An agent can be considered autonomous if it exerts at least a partial degree of independence in decision-making and strategy
Trust	A willingness to have someone else complete a task in a situation where you can't be certain that the task will be correctly completed
Ethics	Your own [pilot's] personal judgments of what is morally right or wrong
Human-autonomy teams	Human-autonomy teams involve one or more humans and one or more autonomous agents, wherein each human and autonomous agent is recognized as a unique team member occupying a distinct role on the team and in which the members strive to achieve a common goal as a collective

detailed by previous literature, the interview data were analyzed using the following phases.

3.3.1 Phase 1: familiarizing the research team with the data

Two authors read every interview to obtain a high-level understanding of our interviewees' thoughts on ethics, trust, and trust repair with autonomous teammates. These authors started this process by going through multiple readings of the interview transcripts to identify pieces of information related to the research questions of the current study and each interviewee's thoughts on the relevant topics described previously (i.e., ethics, trust, AI teammates).

3.3.2 Phase 2: generating initial codes

During Phase 2, the same two authors engaged in an iterative coding process, reviewing each of the pieces of information related to the research questions identified in Phase 1. The identified pieces of information were reviewed for relevancy, and if it was determined to still be related to the research questions, it was incorporated into an existing code or given a new code. Once completed, the two authors met to combine codes, eliminating redundant codes and re-organizing existing codes through discussion with one another. This process identified the primary themes that pertained to our research questions.

3.3.3 Phase 3: searching for themes

With the individual codes identified, the same two authors from the previous phases began identifying patterns among the codes that would merge into themes. The authors generated themes through discussion with one another and clustered codes together to create themes and subthemes with the data.

3.3.4 Phase 4: reviewing potential themes

Once the themes and subthemes were identified, the two authors came together to provide a detailed review of each theme and subtheme. This process was carried out by critically debating and discussing each of the themes and subthemes in terms of their relevance to the research questions and their potential to be grouped into other themes.

3.3.5 Phase 5: defining and naming themes

In the final phase of the thematic process, all the authors came together to review and further refine the themes and subthemes. Any differences in interpretation or agreement between authors were resolved through open discussion, and this phase also saw all the authors decide on theme names and representing quotations.

4 Results

In this section, we first explore how pilots perceive their autonomous teammates' roles in ethical or unethical actions (RQ1). Then we present the pilots' cognitive barriers to accepting an autonomous teammate (RQ2). Finally, we describe how ethics-damaged trust in autonomous teammates can be repaired (RQ3).

4.1 Humans' perceptions of AI teammates' roles in ethical or unethical actions

Throughout the interviews, the pilots frequently relied upon human relationship metaphors in order to conceptualize their potential roles in HATs, especially in relation to their AI teammates' ethical or unethical actions. Notably, multiple participants referred to these AI teammates as children, situating their own roles as parents to AI in a HAT context. From this perceived role as a parent, the pilots expressed that (1) ATs are not accountable for unethical actions, (2) ATs will require human guidance to become ethical decision-makers, and (3) ATs have the potential to become independent ethical decision-makers.

4.1.1 AI agents are not accountable for their unethical actions

As these pilots' interactions with autonomous agents are often restricted to low-autonomy, non-humanoid, and other simplistic versions of AI, their current mental picture for these teammates can readily be compared to a toddler that they are now adopting. Indeed, these pilots look at themselves as being newly saddled with the responsibility of parenthood, guiding the nascent beings through the initial stages of their life. Because the pilots cannot trust the agent to act on its own, they also do not hold the agents accountable for unethical actions. As a result, trust and trust repair issues are redirected toward those in the organization who are building and guiding the agent's development. Indeed, the participants frequently referenced their trust in the military organization as a key contributor to their expectations of an ethical AT, as shared by a pilot:

> I have an open mind. [If] that's what's told to me, then I'll hold my [autonomous teammate] to that center. If [an ancillary] says like, "Hey, this thing can barely fly next to you, but we just want to make sure it's safe." Like, "Okay." (P10)

This pilot shares that their expectations of an AT will be largely based on "what's told to [them]." Given the context of the interview, their ethical expectations of their autonomous teammate will likely also be influenced by the information their ancillaries provide. The interviewee's provided example of an autonomous teammate that could "barely fly" suggests that pilots are willing to work alongside an AT with limited capabilities if their expectations are set by a trusted source (e.g., the engineers who designed the AT). Thus, this statement implies a high level of trust in their ancillaries within the military organization. However, it seems there may be a threshold for initial trust with an AT. When asked if he could trust an AT, one pilot shared that:

> I would trust the same way that I would a two-year-old if it makes a mistake, I don't blame the robot or I don't blame the AI, I blame us as those who empowered it. Once it becomes more advanced, maybe we can have different conversations, because then it's more of a human in nature, hence robot rights (P1)

This pilot compared his initial perception of the agent to a two-year-old that makes mistakes because it is still growing and learning about the environment around it. Like a parent to a young child, the pilot does not "blame" the AI when it makes a mistake but instead blames those who set the rules for the child and guide its behavior through life. Indeed, when rearing a child, caregivers are ultimately responsible for watching the child and ensuring that they learn important lessons about societal norms and rules, including that which is moral and acceptable behavior. Thus, bridging that comparison, the agent is also learning what the ethical standards are for the military and how they must operate around these standards. Violating these rules does not necessarily reflect on the agent but on those "who empowered it" and gave it both the capabilities and commands to operate in the context in which it is operating. As such, ethical and trustworthiness standards for a new AI teammate are not the same as those for a fully fledged member of the team, and these expectations will change as the agent "becomes more advanced" and grasps a greater understanding of the moral rules that may guide its behavior. Furthermore, in its present state, much like a young child, the agent is not fully actualized, meaning that the pilots, for the most part, perceive these agents as capable of growing to more fully understand the implications of their actions and gain greater capacity to be trusted with important decisions.

Similarly, another pilot focused on the growth potential of these agents from their formative states and unpacked their assessment of responsibility given their current functional status. This pilot shared that:

> I view it as more like you're trying to teach a toddler, take a toddler. I've got two kids at home. And sometimes they hit their sister, or they do things, like they make bad actions. But I can't necessarily blame them because I know that they're not a mature adult. And I know that they have limitations. And I can get upset and flustered with the actions, but that doesn't mean I love them any less, or want to help them any less (P3)

Unlike P1, P3 more closely situates themselves as the direct parent and source of interaction with the agent, expressing true concern not only over when they "make bad actions" but also acknowledging that these bad actions would not make them "want to help them any less." This pilot's remarks demonstrate that, once again, the belief that agents, in their formative stages, are not necessarily responsible for their actions and that the level of trust in the agents is mediated by their experiences when guided by supportive "parents." The pilots desire to trust in the agents but need to see the agents grow into maturity before they can trust them fully, particularly given the potentially grave consequences if the agent does not behave within the ethical parameters set by the organization. The openness of this pilot to help the agent adapt and understand the rules necessary to act as a full teammate, though, demonstrates that there is a belief in the potential for both technological advancement and collaboration with those working with these agents to get them to ready state.

In line with their newfound responsibility for these agents, the pilots also believe that the agents themselves cannot be trusted to act on their own, given the severity of the consequences that could result from their actions. Indeed, the pilots fear that, much like new parents, they would need to keep a constant watch over the agents and monitor their every action to ensure that they did not do unintended harm. When considering the comparison between agent and child, one pilot shared that people:

> …Have low expectations for this thing. But I wouldn't be nearly as patient as I would with a child. Because the consequences are too great. I would have a pretty short leash, I think (P9)

While the pilot agrees with the comparison between agents and children as a matter of "low expectations," he finds that the comparison overlooks the "consequences" that may result from their actions and thus advocates for a "pretty short leash" on their activities, seeking restrictions on the agent's autonomy. Much like the other pilots, P9 believes that the agent should not be fully trusted at its current stage, particularly when it comes to situations involving ethics and potentially grave consequences such as civilian death. Later on in the conversation, this pilot shifts the discussion from children to inexperienced pilots. This pilot shares that, like an inexperienced wingman, an agent would be:

> …Totally draining of my time and energy and my brain bytes because I'm just worried about [the teammate and they must be] always watching over it like, 'All right, is it doing what I expect it to do? Is it doing what I told it to do'? (P9)

P9 is worried about having to "watch over" this agent rather than be able to trust it as a fully actualized teammate; much like a baby, they cannot leave the agent to act on its own accord, and that constant monitoring is mentally and physically exhausting for pilots and new parents alike. Indeed, its evident through these comparisons to agents as toddlers that the fear for pilots is that they must take on responsibility for the actions of a being that cannot comprehend the ethicality of their actions and, if they fail to shape them properly, they may never be trustworthy enough to be put out into situations that test those boundaries. Thus, as they continue to view these agents in their infantile form, they cannot possess the level of trust necessary to test the limits and proverbially let the agent off the leash in combat.

4.1.2 AI can grow to handle decisions of an ethical nature with human guidance

Comparisons between agents and children continued throughout the interviews and shifted to more discussions around how the agents would transform through time and experience with the pilots. Indeed, these pilots explored how these agents might grow and how their relationship would shift, given those changing expectations. One pilot contextualized this within the framework of the organization and how that molds the priorities and behaviors of the agent, noting that it is mainly about:

> …How it has been programmed. Obviously it doesn't have the benefit of years of upbringing and childhood, et cetera. It's formed its own opinions. So, I just think if you would play with it and let it, since you have to understand how it's been programmed, to sort of judge it from that standpoint (P6)

This pilot believes that what is important about establishing the agent as a being with the potential to navigate ethical situations stems from how it's programmed, which in his viewpoint, is comparable, though not replicable, to "the benefit of years of upbringing and childhood." Because the agent lacks a strong moral and social upbringing that shapes ethical perceptions, its opinions are formed through those humans that have contributed to its formation thus far. In this instance, the pilot acts more as a stand-in parent, picking up where the programmers left off to "play with it" and understand how to judge its capabilities for ethical actions and trustworthiness from there. Thus, while the pilot does not take responsibility for how the agent has come to be, he does exhibit the willingness to understand it and contribute to its development.

Another pilot also explored what it would be like to contribute to the development of an agent but focused more on how the pilots would respond as they pushed or even crossed ethical or otherwise established boundaries. Unlike the previous pilot, this pilot made a direct comparison to parenthood, sharing that this is:

> …Sort of like raising a kid, maybe you would take away its privileges for a while until it slowly demonstrates iteratively kind of that it can handle decisions like that, because obviously if you tell... Let's say we're in country X and they're extremely upset because you killed 100 kids, if you tell them the system says it's sorry, it's not going to do it again, they're not going to be like, "Okay, we're fine with it." I think in that scenario you'd have to keep developing the software, I guess, and let it continue to slowly prove itself that it can handle decisions like that (P2)

Like when a child grows older, the agent begins to slowly gain independence within a scaffolded environment where the agent must show that it "can handle decisions" of an ethical nature, such as missions where mistakes involving innocent civilians cannot be overlooked. In this case, if an agent violated the trust of a pilot or the entirety of the organization, he believes that it should get "its privileges" taken away until it can learn its lesson. The pilot believes in restricting the autonomy of the agent until it can "prove itself" to take on immense responsibility–this restriction involves "developing the software" to better fit the ethical understanding and framework for impactful decisions. This development is akin to the development that a child undergoes as they grow and the lessons that parents instill as their child navigates mistakes. Similar to initial trust, the continued development of trust will be influenced by input from a pilot's ancillaries, as shared by another pilot during a discussion of trust-damaging behaviors:

> So the relevancy of the question is, I have to rely on a combination of my leadership and the engineers and maintainers who are telling me that [an issue with my AT] has been fixed, right?...If I fly and this particular system isn't working, I land, I tell maintenance it's not working, and then they do something and then they give it back to me and say it's fixed (P01)

Rather than immediately turning to the AT to address its trust-damaging behavior, this pilot states that their trust will be restored by the engineers and maintainers, who are expected to have the ability to change the AT's behavior. The response also suggests that the pilot would seek an option to end the current mission as soon as possible, seeking the next opportunity to land the aircraft and hand off the 'fixing'

process to others. Under the paradigm of human-autonomy teaming as parenting, this could be considered analogous to a parent seeking a behavioral expert to correct their child's actions. Considered alongside a prior pilot's willingness to "play with" an AT as the agent's capabilities develop, this response showcases a willingness to seek help to improve an AT's capabilities from outside sources, too.

4.1.3 AI has potential to act autonomously and independently as a fully actualized teammate on critical, real-world missions with ethical implications

Finally, the pilots explored instances in which the agent can act more independently and whether it is feasible to place all the responsibility on the agent in sensitive situations. In a sense, the pilots imagined what it would be like for the agents to, metaphorically, leave the nest and act autonomously as a fully actualized teammate on critical, real-world missions. That is, they unpacked whether these agents could be fully autonomous teammates who can make ethical decisions while fostering trust. One pilot conditionally accepted the agent taking authority over mission capabilities, stating that:

> I would say pretty much 99% of the time, I would be very happy with that autonomous agent always making those autonomous decisions for defensive purposes, without any need. Maybe I could see like again, the counter measures thing, just do it. Like just do it. There might not be a lot of time. Just go for it. Completely autonomous decision, it feels the threat, it does it (P4)

P4 expressed a much more open mind to the agent "making those autonomous decisions" with the caveat that it be for "defensive purposes" in which the agent would respond with "countermeasures" to a "threat." A few participants shared this belief that these agents could possess full autonomy over combat-related decisions when it was nested within defensive measures rather than offensive attacks that carry much greater risk and ethical considerations. Still, this pilot's willingness to let the agent act autonomously over these decisions demonstrates an openness to this design and to a future in which these agents can be independent of human oversight. As such, this pilot demonstrates an acceptance of future agent independence, much like when a parent watches their child leave the home as an adult, even if that acceptance is conditional.

As with other stages of the agent's existence, the pilots frequently cited the lack of current autonomous functionality and ability to process ethical decision-making scenarios as reasons for not trusting an agent. However, some pilots expressed the belief that, eventually, AI may reach a general intelligence level that would allow them to take on tasks on their own. Furthermore, they expressed a belief that through their interactions and relationship-building opportunities, this confidence and trust in the agent could be achieved. P15 demonstrated this by sharing:

> It's not human, but it's getting darn close, and I think the relationship analogy is the best one I can give. If you're asking if I'm going to trust it, if you're asking if I'm going to give it responsibilities, no different than if that's another person or my spouse or my child or a coworker. It's a relationship, and relationships are built on trust and very small decisions and interactions every day (P15)

As the agent approaches human-level intelligence and relational capacities, the pilot views the agent as closer to a fully fledged human capable of proving its trustworthiness to act on its own. The pilot focuses on the "relationship analogy" akin to a spouse, child, or coworker and asserts that if the relationship is "built on trust" from "small decisions and interactions every day," then there is an opportunity for that agent to be given eventual decision-making powers in more critical scenarios. Thus, 1 day, the pilot envisions that the agent can leave the metaphorical nest, given a proven track record of responsibility, and start acting with agency as a human being would. Important in this comparison is the willingness of the pilot to treat the agent as something whom he can have a genuine relationship with rather than an object or tool. His openness to this relationship-building and eventual emergence into an actualized form indicates that, while far off, some pilots can envision a future in which their parenthood duties are eventually shifted away from the early stages of heavy monitoring and later stages of development toward trusting in their relationship and background to view the agent as an autonomous being.

Despite some pilots describing instances in which they believe the agent will be competent enough to move out from under their close guidance and care, others struggle with fully trusting in the agent to be on their own. For some pilots, it was an issue of their personality and the general personalities of those who hold their positions, stating that they need:

> …Some level of control over that. Being a, and probably most fighter pilots are, type A personality kind of guys and girls, that we rely on that ability to be able to correct the situation. And in this scenario, an autonomous thing I don't think would necessarily have that. So there could be a mistake going on. I could, as the flight leader or whatever, recognize that mistake is happening and have the inability to do anything about it. (P6)

In this particular case, the pilot feels that full autonomy should never occur and that the pilot should "be able to

correct the situation" if they believe a "mistake is happening." As the participant states, many pilots have similar "type A" personalities, making it difficult for them to let go of their control over the agent, much like a parent struggling to see their child grow up and become independent. However, unlike a parent, the pilots do not necessarily see the agent ever becoming fully autonomous in that the technology could act on its own accord without the consent of the "flight leader." Much like the structure and culture of the military organization invokes a chain of command, the pilots would want an AT to follow a similar hierarchy and request permission before making ethical decisions. Much like P6, P7 expressed hesitance about full, unassisted autonomy from an agent in their missions, sharing:

> For us, there's just too many decisions that need to be made, which is why you hear a fighter pilot say, "You're never going to pull us out of the cockpit." It's because you need a brain to be making the decisions. But, maybe one day they prove us wrong, but that's just kind of where we currently stand (P7)

P7 believes that humans should never be pulled "out of the cockpit" due to the "many decisions that need to be made" because a "brain" needs to decide in combat. Both P6 and P7 demonstrate a belief among participants that human capacities to make these decisions can never be replicated in an entity that is not human. While P7 recognizes the potential for artificial intelligence to "prove us wrong," he has little faith in the current conceptions of AI teammates to take on these roles and work autonomously in combat scenarios.

While several participants could not imagine a future where the agent acts autonomously in these combat scenarios, others unpacked the idea in their interviews and discussed their discomfort with these agents becoming independent beings. These participants positioned themselves between the groups that were open to autonomy and those completely against it, recognizing the inevitability of these agents acting on their own accord while also worrying about the potential ramifications related to autonomy. Relatedly, P3 explained that:

> I think the human will still, in the end, own that action. Even if the drone is making its own decision, and through automation, executing that action, I think that's why it will still fall back on whoever, whichever human is closest to it or working with it the closest (P3)

The pilots often shared that they believe that humans will always be, in the end, responsible for the agents and must "own" the actions they take, including the ethical ramifications of those actions. Even when a fully autonomous being is "executing" the action, the pilots, programmers, or anyone else "working with it" is ultimately responsible for the outcome of their action as they shaped how it came to be. As in a parent–child relationship, the parent does their best to mold the child as they grow to make the right choices, guiding them through their own set of moral beliefs and appropriate actions; in the end, the child is then entrusted to go into the world and make those correct choices. However, that now-adult is still a reflection of their parents and upbringing in a way. Unlike the parenting metaphor, though, these autonomous agents would be more directly controlled and shaped through their interactions with humans, justifying more of this perceived culpability over the agent's actions from the pilots. As such, it is evident that AI agents should be treated with different care than a child regarding ethical decision-making and trust, even if there are evident parallels that are taken between the two.

4.2 Barriers to accepting autonomous teammates when ethical decisions are involved

In 4.1, we have described how pilots perceive their AI teammates' roles in possible ethical or unethical actions and their expectations for AI's potential to act autonomously and independently as fully actualized teammates on critical, real-world missions with ethical implications. Additionally, the pilots described the large influence their ancillaries' statements will have on the pilot's expectations of and initial trust in an AT. However, during our interviews, the pilots expressed that they have reservations about teaming with an autonomous agent that is truly autonomous (i.e., does not require human input), especially in ethically ambiguous scenarios in which trust is essential and, as it stands, unearned. In this section, we identify two core barriers for human teammates to accept autonomous teammates in a HAT, especially when ethical decisions are involved: (1) who will be held responsible for an autonomous teammate's unethical actions, and (2) the potential inability to understand how and why an autonomous teammate makes decisions.

4.2.1 Belief that autonomous agents cannot be held accountable

As mentioned in 4.1.1, pilots in this study believe that an autonomous teammate will not be held accountable for its actions, as one participant shared:

> I believe that when I employ a tool, I am responsible for the outcomes of that tool. So if I am choosing to bring these Loyal Wingman to a fight and knowing that, "Hey, this is how they operate and they will make that decision." When that fails, copy, the machine was the one that made the decision, but I was the one that put the machine in that battle space to do that thing. So that's on me (P13)

While clearly acknowledging the ability of an autonomous teammate to make their own decisions, P13 still views autonomy as another "tool." As such, they believe that, like any other tool, the person using a tool is culpable for every outcome, not the tool itself. From his perspective, humans *should* be held accountable for an autonomous teammate's actions because this teammate is not able to truly take responsibility for its actions as a tool. Rather than pointing to any particular human in the process, P13 appears to place the blame on the person that "put the machine in that battle space." Therefore, P13 would place the blame on whoever decided to include an autonomous teammate, Similarly, two other pilots believe that an autonomous system cannot be held responsible for unethical actions:

> If everyone in mission planning was aware of its shortcomings and it still did it, then the blame would probably be on the operators and the mission planners themselves (P08)
>
> And our bosses are telling us that we have to fly with this thing so we're going to fly with it... It's going to hit us or somebody else, or it's going to bomb the wrong target and I'm going to get blamed for it…I don't want to go near this thing until it's fully vetted and as good or better than a human (P09)

Like P13, P08 places the blame onto any humans, whether inside or outside the HAT, that decide to include an unethical autonomous teammate in an operation. P09's response suggests that the pilots believe they would not be technically responsible for the issue but would be held responsible for an autonomous teammate's actions. P09 argues this view by reintroducing an earlier point about pilots' limited autonomy when deciding which systems belong on the F-35A and when they are used. Even if P09 believes that mission planners are responsible for including the autonomous teammate on a mission, the human pilots will receive the blame. For that reason, P09 does not want to "go near this thing" until the autonomous teammate has proven it is at least as competent as a human. Despite differences in their beliefs on who initiates the blame, they all believe that the humans involved are ultimately at fault for choosing to employ this autonomous agent. Another pilot with a similar view proposed another approach to making them more willing to team up with a machine:

> I wouldn't want it, at least in the beginning, making [hostile target identification] decisions on its own. I would want to make sure that I had the final say in that. Because at the end of the day, I'm probably going to be the one still held responsible if something goes wrong (P03)

Like the other pilots, P03 does not believe that autonomous teammates will be accountable for their actions. P03 states that "at least in the beginning," they would want an autonomous teammate to seek consent from human teammates before it performs an action, thereby limiting its autonomy. Notably, P03 made this statement when discussing the task of identifying a target as hostile or non-hostile, a task that can result in a friendly fire incident if done incorrectly. Because of the potentially catastrophic consequences of misidentifying a target, P03 would not want an autonomous teammate to perform that action unassisted until it has matured as a system, which echoes P09's concern about working with an unvetted system. In sum, because the pilots believe that humans will always be held accountable for an autonomous teammate's actions, they want an initial limit to its autonomy until it has proven itself as a reliable system.

4.2.2 Challenges to understanding an autonomous teammate's decision-making process

As part of the years of training Air Force pilots endure, the military trains its personnel on the ethical principles that guide military action:

> We take a class, basically academics, called LOAC, Law of Armed Conflict. That would be, I think, one of the starting places for ethics, if you will, as far as how to conduct war…and then match that with also what we call the ROE, rules of engagement (P11)

The LOAC provides some guiding principles for ethical decision-making but is only a "starting place." As P11 describes it, LOAC provides general guidelines for warfare. Instead, pilots will rely on the ROE. Whereas LOAC provides general principles, ROE provides a list of acceptable and unacceptable actions. As described by P04, "in the rules of engagement, they [mission planners] would have specified who is an enemy, who is not and then what collateral damage is acceptable." Therefore, acceptable ethical actions are dictated by mission planners before a mission occurs, not the pilots during an operation. Despite mission planners' attempts to explicitly state when certain actions should be taken, ROE cannot encompass all the variables present during an operation:

> Every flight is different, every mission has its own things that the team is going to have to solve without procedural guidance to direct it, and that's going to be based off our team's mental model (P13)

P13's statement that every team will have to make decisions "without procedural guidance" was echoed by all our interviewees. As P04 stated, ROEs attempt to provide an acceptable amount of collateral damage, but there can be situations where a pilot cannot accurately predict the amount of collateral damage an action will incur. To address these situations, P13 suggests that teams rely on their team's

mental models. Presumably, P13 also includes the possibility for a team ethical mental model to exist such that team members will make decisions that follow the team's ethical framework, not that of any individual.

While the pilots like the idea of a HAT being able to construct a team ethical mental model, they are concerned that an autonomous teammate's rationale for its decisions may be impossible to understand:

> Once it starts making decisions that are not understood by me, although it could be a better decision, it may induce confusion into the rest of the formation and that could potentially be detrimental…I don't necessarily want to inhibit the artificial intelligence creative problem-solving. If it can solve the problems better, then I don't want to induce confusion within the formation (P14)

P14 acknowledges that autonomous teammates may be able to make better decisions than human teammates. However, P14 is concerned that the decision could be incomprehensible for human teammates, thereby leading to "confusion with the formation." This confusion could impact the development of team cohesion, potentially hampering or preventing the development of a team's ethical mental model. This puts P14 in a compromising position; he wants autonomy to present the best possible solutions but doesn't want its processing to become so advanced that team processes are hindered. A couple of other pilots shared this concern about an autonomous teammate whose intelligence is superior to humans':

> It should never be, it can kind of be like a black box thing. We never need to see the code…pretty much every tactic we'll start with the mission system and talk a little bit about how it actually works, so we can kind of gauge that context in our mind…it would need to be very transparent (P04)
>
> I know that we're going to be quickly exceeded in the ability to process information by those types of systems, and I have confidence in their ability to execute it. I just need to have it well-defined for me. I don't need to know the internal workings necessarily on it, but I need to know that when input A happens, output C happens…otherwise, I can't mitigate risk in a war time scenario (P13)

In line with P14, P04, and P13 address the possibility for future autonomy to become super-intelligent, yet, they see the benefit from including super-intelligent teammates within HATs. P14 did not state whether they want to understand the underlying algorithms guiding an autonomous teammate's behavior (although it can be implied), but P04 and P13 explicitly state that they do not want to see the code. For P04, the autonomous teammate just needs some level of transparency so they can "gage" their autonomous teammate's capabilities and consider it as they plan the tactics for a mission. P13's perspective is similar, stating that the understanding of an autonomous teammate's "internal workings" is unnecessary; he only wants some understanding of the system's inputs and outputs so that he can reliably predict an autonomous teammate's behaviors in a "wartime scenario." Despite the varying perspectives on how much transparency and explainability an autonomous teammate must provide, the pilots agree that an autonomous teammate, even one that makes better decisions than humans, must be able to share some level of processing.

4.3 Repairing ethics-damaged trust

While understanding the pilots' perspective on how trust may be established and maintained is valuable, it is inevitable that trust will be damaged at some point. All pilots reported some degree of expected trust degradation immediately following an unethical action. When discussing ethics-damaged trust, three primary strategies emerged which would support short- and long-term repairs.

4.3.1 Strategy 1: limit autonomous teammate's task involvement

Trust tends to decrease following expectation violations which can be jarring and concerning for those involved. Immediately following an unethical action, pilots believed that an autonomous teammate's autonomy should be limited:

> [Pilots should be able to] dial back its [the autonomous teammate's] level of automation to consent for certain actions. Like it will not execute XYZ unless I tell it it can. (P03)

P03 feels it's important for him to be in control of the teammate's abilities, recommending an immediate decrease its ability to make decisions. The pilot believes he should be able to "dial back" decision-making freedom following an ethical violation, and the autonomous teammate should consult a human before proceeding with future decisions. However, he does not believe that complete removal from the team is appropriate. Dismantling a team has its own implications, and maintaining that structure was important to the pilots. Another pilot stated even more explicitly that even after multiple ethical violations, some level of trust would be maintained:

> … If you start to have multiple issues of like, they say it's fixed and it's not fixed and it keeps having an error, you get to the point where you don't trust it. Would I ever write it off completely? No. (P01)

Even though P01 stated he might "get to the point where [he doesn't] trust it", he still would not "write it off completely," indicating that he still sees value in the teammate. Another pilot agreed that following an ethical violation, an autonomous teammate is still valuable. Similarly to P03, P02 believed that restrictions should be put in place:

> You can give it [the autonomous teammate] slow processes, give it less, start with smaller decisions. Decisions that have a smaller effect and slowly build that up again, I think you can re-establish trust. (P02)

Following an unethical action, P02 believes that the best course of action would be to reduce the criticality of the autonomous teammate's decisions until trust is restored. Similarly to a child who misbehaves, pilots believe that an unethical autonomous teammate should be dealt with by reducing their freedoms. Following that immediate penalty, the process of trust repair can begin.

4.3.2 Strategy 2: investigate why autonomous teammates acted unethically

Immediately restricting autonomy provides a short-term solution for relieving uncertainties around an autonomous teammate's behavior. Moving forward, pilots also want to investigate the situation to determine *why* the autonomous teammate behaved unethically, considering initial impressions may be misguided. For example, perceived unethical behavior may have been the best option given the information an agent had access to. The pilots' responses suggest that they consider three main variables when assessing the ethicality of a decision: the probability of success, the autonomous teammate's intentions, and the autonomous teammate's capabilities.

One way to judge the ethicality of a decision in hindsight is to evaluate the probability of success at the time of the decision. That is, an action may not have the desired effect and result in damages or casualties. However, the pilots acknowledge that few, if any, actions have a 100% chance of success and judging ethicality solely by outcomes is not always appropriate:

> We have to make a decision to try to comply with the rules of engagement, but there's a lot of uncertainty there. Like we've got out targeting pod on something and we're like pretty sure it's the thing we're supposed to hit, but like also, it might not be…it's more of, yeah, risk management (P04)

According to P04, decisions often involve risk and uncertainty where they must make a judgment call on whether an action is abiding by the ROEs or not. The probability of success assessments may be informed by gaining insight into the autonomous teammate's intentions. For example, P09 stated, "As soon as it does something bad intentionally, that thing is not to be trusted by any stretch". Understanding the intentionality and motivation behind a decision can help in painting a fuller picture of a situation defined by uncertain gray zones. P02 also indicated that a window into the autonomous teammate's rationale would help repair trust following an ethical violation:

> If he [the autonomous teammate] screwed up one thing and I was like, "Don't do that." And he was like, "Okay, I won't do that." And then we go and we complete the rest of the mission just fine, it would be a debrief point where it's like, "Hey autonomous teammate, why did you start doing that?"

This pilot indicated that he would require a "debrief" following a mission where the autonomous teammate would indicate the rationale and intentions behind a particular action. In the quoted example, the teammate integrated the pilot's feedback and altered their behavior, but the pilot still asked for additional information surrounding the initial violation. This indicates that a shift in behavior does not eliminate the need for additional insight into the reasoning behind an autonomous teammate's unethical behavior.

The final factor the pilots investigate during this part of the trust repair process is the autonomous teammate's capabilities. Understanding a system's capabilities and limitations is crucial for setting expectations. It is possible that an autonomous teammate would be placed in a situation where it is out of its depths. P08 explained how he would handle this type of scenario:

> If everyone in mission planning was aware of its [the autonomous teammate's] shortcomings and it still [performed an unethical action], then the blame would probably be on the operators and the mission planners themselves (P08)

If an autonomous teammate was placed in a role that it was not capable of assuming and subsequently behaved unethically, the pilot would not lose trust in the teammate. Rather, trust in the other team members who allowed the situation to occur might be damaged. Pilots might build trust in an autonomous teammate's capabilities by understanding how it processes information:

> I would want to look at how it takes in information, all of its sensors and all of the inputs and then how it processes those. Whatever the decision-making process is to lead from input to output and then finally then the execution, so once it's made that decision, how it executes it and see where the breakdown was. First of all, did it perceive the situation accurately? Maybe it didn't know there was a school there (P02)

P02 wanted to know at what specific point the system failed to gain a better understanding of the situation. He wants to know "all of the inputs and then how it processes those" (P02) to understand the system's capabilities. He suggests, "maybe it didn't know there was a school there," implying that the autonomous teammate didn't have access to sufficient or accurate information to make an optimal decision. For him, trust repair is dependent on painting a detailed picture of the system's information intake and decision-making processes.

4.3.3 Strategy 3: confirm that autonomous teammates' performance improves

After restricting an autonomous teammate's decision-making abilities and investigating the reason behind an unethical action, a teammate can be given the opportunity to demonstrate behavior change. Pilots want to ensure that the autonomous teammate is improving and is less likely to repeat past unethical behaviors. The pilots clarified that they would like some method for personally confirming the autonomous teammate's performance has improved. One pilot suggests that he "want[s] to see the data behind it" (P09) in order to fee assured that previous issues have been resolved. Others report that they want to be part of the training process to improve their autonomous teammate:

> I would want to see that [the autonomous teammate's performance] in execution. Like a training or a real life…However it would be other than just a computer simulation. I would want to see that in action personally to see how that all goes down (P10)

P10 is suggesting a highly involved approach to trust repair. He wants to be in the cockpit of an F-35A as the autonomous teammate improves its ethical decision-making. Although flight simulations are not enough for P10, P01 stated, "hopefully, you can throw some simulations at [the autonomous teammate] before you actually put it into a combat scenario" (P01). This position is held by P03, who believed that training the autonomous teammate in a "non-real environment" would be necessary following an unethical decision:

> I think you just keep working and training with it in a non-real environment. You don't bring it into the real world, real environment, and give it those real repercussion chances until you are really, really damn sure that it's not going to act unethically. Then if it does, "Well, okay. We are going back and we are continually working on this." (P03)

Training in a real-world scenario was out of the question for P03. Since real-world scenarios can be highly consequential, he would need to be "really damn sure that it [the autonomous teammate] is not going to act unethically". P03 was also committed to the training process, stating that they would be "continually working" on getting the autonomous teammate to a place where he was 100% confident that it would make ethical decisions in the field. P09 affirmed that trust can be regained through "retesting, reevaluating, or a lengthy period of time with multiple data points". Regardless of the preferred method, the pilots' responses indicate that ethics-damaged trust will require evidence that the autonomous teammate's performance has improved.

5 Discussion

To explore the potential interplay between trust and ethics within HATs, we conducted semi-structured interviews with F-35A pilots, a population of individuals who regularly interact with highly automated technologies and will likely be some of the first humans to participate in realized HATs. Given the importance of team ethical decision-making in their profession, the pilots' perspectives on future ethical ATs served as a valuable resource to address our research questions. Our study produced multiple key highlights. First, the pilots view ATs as agents that will initially be incapable of performing comparably to a human teammate in ethical decision-making. For ATs to improve their performance, the pilots expect human teammates to act as parents that guide, monitor, and seemingly punish the AT until it develops into an agent capable of acting autonomously (RQ1). Second, two barriers exist to accepting an AI as a teammate for ethical decision-making: concerns about the culpability for an AT's unethical actions and the potential inability of human teammates to understand how an AT makes ethical decisions (RQ2). Last, repairing trust damaged by unethical actions is possible, but it will require the AT to thoroughly explain the rationale behind its actions and display improved ethical decision-making in future teaming events (RQ3). In the following sections, we discuss the implications of our findings on the current literature on ethical HATs.

5.1 Reapproaching AI's role in HAT's ethical decision-making

Through the participants' discussions, the conversations often returned to how the ethical decision-making context forced them to reconsider agents' current and future roles on a HAT. The roles focused on the present evaluations of autonomous agents, in which the participants cannot trust them to act independently in critical scenarios, as well as their imagined future roles that acknowledge agents' potential to be nurtured by humans toward ethical decision-making and agency in critical, real-world missions. The assignment of human-like characterizations of the AI, comparing

their roles to children, expands upon the discussion on the evolution of AI from a tool to AI as a teammate [89], demonstrating that acceptance of these teammates in critical scenarios requires greater care when designing and implementing these autonomous agents. We discuss how these roles compare to the present literature in terms of their evaluations of present inadequacies and future potential for growth.

In its current state, the participants felt neither comfortable nor confident in the present manifestations of AI, particularly as they deemed the AI unable to take accountability for its unethical actions. This finding affirms previous research that human teammates struggle to trust AI teammates [84], particularly regarding their level of machine intelligence [32]. Indeed, the participants' comparisons between agents and toddlers indicate that they are infantilizing these agents in their current conceptions and view them as lacking intelligence, independence, and understanding of how to apply their abilities and knowledge for a given scenario appropriately. However, unlike previous research that finds that humans expect human-like behaviors and almost unlimited potential in an AI teammate [114], the current study adds complexity to the perceptions of AI teammates when applied to critical, ethical scenarios by demonstrating that these contextual considerations stymie their desire to act on that unlimited potential. While the pilots stated that they could eventually consider an agent a teammate capable of making their own ethical decisions, in its current state, they feel there are limitations on their desire for the agent to have complete autonomy. For instance, these pilots viewed these agents much like children; they wanted to limit their actions and abilities in combat and other ethically ambiguous scenarios until they could prove themselves trustworthy. Even then, some pilots did not believe that agents could be independent actors in teaming scenarios, thus questioning the possibility of ATs becoming fully actualized teammates.

While there are evident hurdles to adopting ATs given present perceptions, the pilots did present openness to the potential future roles of ATs in handling ethical decision-making with human guidance and even acting autonomously; as fully actualized teammates on critical missions. Indeed, comparisons to human children indicate that the participants are open to and understand the potential for agents to grow in autonomy, extending beyond the status of a tool through the guidance of human teammates. These findings bolster the need for Rix's propositions for enhancing human–AI collaboration, particularly those focused on establishing the team as a social entity [81]. The participants in this study desired and made clear connections between a pillar of human social existence, the family, and their relationships with the agent on the team for nurturing ethically aware decision-making in ATs. While these dynamics are nuanced based on the particular task, the need for relationship building is even more critical to building and maintaining trust in HATs when these teams operate in potentially hazardous environments.

Furthermore, collaborative behavior, which relies on teammate perceptions, including the ability to "exhibit proactive, iterative, responsive as well as competent behavior" [81], is paramount for these pilots to trust the agents to act on their own accord. Indeed, extensions of the technology acceptance model (TAM) of AI, in particular, highlight the role of trust in the acceptance of AI [14], and our findings support other research on how obtaining the trust needed for effective teaming and AT acceptance becomes much more difficult in the face of ethical dilemmas [85, 86]. While some work suggests that understanding an AI's potential for agency may help promote teamwork among human members and greater cooperation overall [58], these pilots still presented cognitive barriers to teaming with these agents. Many also felt less trusting and secure in working with fully autonomous agents, which aligns with the research on desires for adaptive autonomy levels in ATs, given the task demands and contextual bounds [38, 39]. As such, interventions that support healthy teamwork, particularly building a greater understanding of and empathy for the AI, are wholly necessary to facilitate trust in the HAT and agent, especially for scenarios in which trust is both imperative and challenging to build and maintain, such as in combat [105, 110]. Future design considerations for these imagined roles, then, must be carefully crafted and center these needs to allow these ATs to be fully implemented in practice.

5.2 Ethics complicates the trust dynamic as autonomous systems become teammates

Across all present and imagined roles of AI was the pervasive concept of trust, demonstrating how imperative it is for researchers to understand the trust dynamics that shape these teammate perceptions to facilitate adoption and effective teaming. Participants detailed how the question of ethics alters their trust relationship with an AT, specifically complicating it with organizational contexts, the explainability of the AT's actions or processes, and trust repair after damaging actions.

While a central focus of the interviews regarded how humans saw ATs through the lens of a parent-to-child relationship, a third entity that also played a role in developing and sustaining trust in human-autonomy teams was brought up. Specifically, the pilots discussed the organizational guidelines and expectations on the team's operation and execution of their shared objective in the pilot interviews. Within the context of the current interviews, the pilots and their teams must operate within a strict set of guidelines like the Rules of Engagement (RoE) and the Uniform Code of Military Justice (UCMJ), which includes obeying and respecting the chain of command. There was

an expectation that technology, including ATs, would also be designed to operate within the organization's guidelines and expectations. This expectation represents a potential benefit for implementation in rigid environments like the military. Pilots reported that this trust in the organization helped instill and maintain the trust that the AT would be capable of ethically fulfilling its role on the team. However, it might also limit the potential for trust development as it creates another set of expectations for the AT to fulfill, which ultimately falls on developers to ascertain and meet, adding to a now-growing list of expectations for ATs [114]. Additionally, this finding demonstrates the transitive nature of trust within the organization: pilots trust their superiors, and superiors trust ATs enough to implement them; therefore, pilots trust ATs. While ATs are distinctly different from human teammates, research has shown that trust is positively related to coworkers' perceived benevolence and integrity [98]. Research must explore further the similarities and differences between human–human and human-autonomy teams to understand their complexities.

Another factor ethics adds to the complex nature of trust within human-autonomy teams is how pilots wanted insight into their AT's decision-making process following unethical behavior. The AI literature echoes this perspective, which hails the importance of explainable AI (XAI) [5, 33] to improve trust between humans and AI. However, the literature also points out its implementation challenges [34], as AI often bases its decisions on complex datasets, making model interpretation and communication difficult. Furthermore, some literature has shown evidence of 'algorithm aversion,' which refers to distrust in systems utilizing AI and preference for human decision-makers [11, 21, 22]. Dietvorst et al. found that participants who saw an algorithm perform were less confident and less likely to choose it over a human forecaster, even when the algorithm outperformed the human [21]. Follow-up research found that participants were more likely to rely on algorithmic forecasts if they could change a fraction of the AI's prediction [22]. In the context of HAT, this may be analogous to collaboration. Instead of an AT simply describing a decision and its rationale, the human teammate might influence the decision-making process. AI explainability may also improve more than trust for human-autonomy teams as it is highly related to situational awareness at many levels for these teams [28]. While these approaches may help to build and instill trust in an AT, findings also showed that a focus might also need to include trust repair.

In further support of ATs meeting organizational (and individual) expectations and providing explanations for their actions, the pilots repeatedly emphasized the need for demonstrated behavioral changes following an ethical violation by the AT. After an AT commits an ethical violation of either the expectations set forth by the individual human teammate or the organization, an explanation from the AI is not likely to be enough to restore original levels of trust. Interestingly, existing mainstream trust repair strategies (apology and denial) are largely ineffective for ethical violations in human–AI teams [86, 99], reinforcing the complex nature of this relationship. Work from Industrial Organizational Psychology has shown that trustworthy behavior over a period of time or a series of interactions can aid in trust repair [90]. However, this method of trust repair requires repeated interaction following a trust violation. The findings of this study showed that depending on the violation type and severity, pilots were more or less open to interaction following an unethical action. Future systems may be implemented in settings where human teammates can easily override or shut down their autonomous counterparts. Therefore, lower-stakes environments might be better proving grounds for ATs than those with more significant ethical consequences. Alternatively, several pilots pointed out that their trust in the AT may be reformed if the individuals responsible for maintaining the AT fixed the problem that caused the infraction, further complicating the trust relationship between humans and ATs. By introducing another party (i.e., technician or developer), the maintenance of trust in the AT becomes further dependent on outside sources, and the level of repair may also be subject to the human's trust in the technician. Addressing the complexities that ethics introduces to the trust relationship between humans and ATs requires a clear understanding of the expectations at several levels, alongside a robust understanding of AI algorithmic functioning for explainability and trust repair.

5.3 Limitations and future directions

One primary limitation of our study is our reliance on a sample of F-35A pilots, which is a highly specialized group. For their line of work, ethical decision-making often involves decisions that can result in loss of life. Therefore, an ethical autonomous teammate's moral principles will significantly differ in contexts like business ethical decision-making. However, other elements, like the importance of agent transparency, will likely remain highly relevant regardless of context. A transparent agent helps humans to align their mental models with their performance [63] and engenders trust in decision-making scenarios, including medical decisions [73]. As such, an autonomous teammate's use of an explanatory trust repair strategy will also be important for contexts where a human teammate does not understand why its autonomous teammate performed an action. If an AI capable of explaining its decision may influence humans to perceive it to have mental capacities beyond a simple machine, then it can promote trust [111]. Further, the role of collateral damage and the judgment of the ethicality of such a decision is a major avenue for future research on ethics in

human–AI interaction. For example, investigating how the same decision resulting in collateral damage is judged based on whether the one making such a decision was an artificial entity or a human.

Related to the specificity of the F-35A population, another limitation is the lack of diversity within our sample. This may be a reflection of the overwhelmingly white and male members within the Air Force, especially as military rank increases [2]. Even the military recognizes the lack of diversity within their ranks, which has led to a push to improve recruitment and retainment efforts [31]. Other fields that will contain HATs are also showing a trend of increasing diversity, such as annual increases in women and persons of color attending medical school [6]. In order to best design agents to suit HATs, future research should consider and investigate how demographic factors can influence design considerations.

6 Conclusion

When we consider the fact that autonomous teammates will likely see first use within environments with ethical considerations (e.g., the military), we must consider how to best design an ethical teammate that can promote trust within teams, a significant antecedent to team effectiveness. In our investigation of ethical team decision-making within F-35A flight teams, we uncovered multiple themes around the interaction between trust and ethics. Within a forced teaming context, pilots feel as if they will be given the role of a parent as they team alongside developing machine teammates, and as such, they feel responsible for their actions and may not lose trust if it performs an unethical action. However, once the system develops, unethical actions will likely lower trust that can only be repaired with strategies that explain the teammate's processing and indicate growth. Together, these findings amplify the importance of factors like teaming contexts, human–machine interaction paradigms, transparency, accountability, and trust repair within future HATs whose actions will bear moral weight. Hopefully, these findings can lead to future research that can better explore these topics and eventually design guidelines for trustworthy, ethical autonomous teammates.

Acknowledgements The authors would like to acknowledge the contributions of Ella Kokinda to the development of this research and manuscript. This research was supported by AFOSR Award FA9550-20-1-0342 (Program Manager: Laura Steckman).

Funding Air Force Office of Scientific Research, FA9550-20-1-0342, Nathan McNeese.

Data availability The dataset generated by the current research is not openly available to preserve the privacy of study participants.

References

1. Aghion, P., Jones, B.F., Jones, C.I.: Artificial intelligence and economic growth. In: The economics of artificial intelligence: an agenda, pp. 237–282. University of Chicago Press, Chicago (2018)
2. Research Air Force Personnel Center Analysis, Data Division: Air Force Demographics - Active Duty. Air Force Personnel Center. Retrieved from https://www.afpc.af.mil/The-Air-Forces-Personnel-Center/Demographics/. Accessed 11 May 2023 (2023)
3. Akash, K., McMahon, G., Reid, T., Jain, N.: Human trust-based feedback control: dynamically varying automation transparency to optimize human-machine interactions. IEEE Control Syst. Mag. **40**(6), 98–116 (2020)
4. Akman, I., Mishra, A.: Ethical behavior issues in software use: an analysis of public and private sectors. Computers Human Behav. **25**(6), 1251–1257 (2009)
5. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**(2020), 82–115 (2020)
6. Association of American Medical Colleges: 2022 Fall Applicant, Matriculant, and Enrollment Data Tables. AAMC. Retrieved from https://www.aamc.org/media/64176/download?attachment. Accessed 11 May 2023 (2022)
7. Barnett, T., Valentine, S.: Issue contingencies and marketers' recognition of ethical issues, ethical judgments and behavioral intentions. J. Bus. Res. **57**(4), 338–346 (2004)
8. Beauchamp, T.L., Childress, J.F.: Principles of biomedical ethics, 5th edn. Oxford University Press, Oxford (2001)
9. Braun, V., Clarke, V.: Thematic analysis. American Psychological Association, Washington (2012)
10. Bryson, J.J.: Robots should be slaves. In: Close engagements with artificial companions: key social, psychological, ethical and design issues, vol. 8, pp. 63–74. John Benjamins Publishing Company, Amsterdam (2010)
11. Castelo, N., Bos, M.W., Lehmann, D.R.: Task-dependent algorithm aversion. J. Mark. Res. **56**(5), 809–825 (2019)
12. Chen, J.Y.C., Barnes, M.J., Selkowitz, A.R., Stowers, K:. Effects of agent transparency on human-autonomy teaming effectiveness. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 001838–001843. DOI:https://doi.org/10.1109/smc.2016.7844505 (2016)
13. Chiu, R.K.: Ethical judgment and whistleblowing intention: Examining the moderating role of locus of control. J. Bus. Ethics **43**(2003), 65–74 (2003)
14. Choung, H., David, P., Ross, A.: Trust in AI and its role in the acceptance of AI technologies. Int. J. Human Computer Interact. **2022**, 1–13 (2022)
15. Cohen, M.C., Demir, M., Chiou, E.K., Cooke N.J.: The Dynamics of Trust and Verbal Anthropomorphism in Human-Autonomy Teaming. In 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), IEEE, 1–6 (2021)
16. Cointe, N., Bonnet, G., Boissier, O.: Ethical judgment of agents' behaviors in multi-agent systems. In Proceedings of the 2016 international conference on autonomous agents & multiagent systems, 1106–1114 (2016)
17. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **1989**, 319–340 (1989)
18. De Visser, E.J., Peeters, M.M.M., Jung, M.F., Kohn, S., Shaw, T.H., Pak, R., Neerincx, M.A.: Towards a theory of longitudinal trust calibration in human–robot teams. Int. J. Soc. Robot. **12**(2), 459–478 (2020)

19. Dean, K.L., Beggs, J.M., Keane, T.P.: Mid-level managers, organizational context, and (un) ethical encounters. J. Bus. Ethics **97**(1), 51–69 (2010)
20. Demir, M., McNeese, N.J., Gorman, J.C., Cooke, N.J., Myers, C.W., Grimm, D.A.: Exploration of teammate trust and interaction dynamics in human-autonomy teaming. IEEE Trans. Human Mach. Syst. **51**(6), 696–705 (2021)
21. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. **144**(1), 114 (2015)
22. Dietvorst, B.J., Simmons, J.P., Massey, C.: Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Manag. Sci. **64**(3), 1155–1170 (2018)
23. Dirks, K.T., Kim, P.H., Ferrin, D.L., Cooper, C.D.: Understanding the effects of substantive responses on trust following a transgression. Organ. Behav. Human Decis. Process. **114**(2), 87–103 (2011)
24. Drath, R., Horch, A.: Industrie 4.0: hit or hype?[industry forum]. IEEE Ind. Electron. Mag. **8**(2), 56–58 (2014)
25. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Int J Human Computer Stud **58**(6), 697–718 (2003)
26. Eigenstetter, M.: Ensuring trust in and acceptance of digitalization and automation: Contributions of human factors and ethics. In Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Communication, Organization and Work: 11th International Conference, DHM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22, Springer, 254–266 (2020)
27. Endsley, M.R.: From here to autonomy: lessons learned from human–automation research. Hum. Factors **59**(1), 5–27 (2017). https://doi.org/10.1177/0018720816681350
28. Endsley, M.R.: Supporting Human-AI Teams: transparency, explainability, and situation awareness. Comput. Hum. Behav. **140**(2023), 107574 (2023)
29. Fehr, R., Gelfand, M.J.: When apologies work: How matching apology components to victims' self-construals facilitates forgiveness. Organ. Behav. Human Decis. Process. **113**(1), 37–50 (2010)
30. Fullerton, S., Kerch, K.B., Robert Dodge, H.: Consumer ethics: an assessment of individual behavior in the market place. J. Bus. Ethics **15**(1996), 805–814 (1996)
31. Garamone, J.: Diversity, Equity, Inclusion Are Necessities in U.S. Military. U.S. Department of Defense. Retrieved from https://www.defense.gov/News/News-Stories/Article/Article/2929658/diversity-equity-inclusion-are-necessities-in-us-military/https%3A%2F%2Fwww.defense.gov%2FNews%2FNews-Stories%2FArticle%2FArticle%2F2929658%2Fdiversity-equity-inclusion-are-necessities-in-us-military%2F. Accessed 11 May 2023 (2022)
32. Glikson, E., Woolley, A.W.: Human trust in artificial intelligence: review of empirical research. Acad. Manag. Ann. **14**(2), 627–660 (2020)
33. Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (XAI) program. AI Mag. **40**(2), 44–58 (2019)
34. Hagendorff, T., Wezel, K.: 15 challenges for AI: or what AI (currently) can't do. AI Soc. **35**(2020), 355–365 (2020)
35. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., De Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. Hum. Factors **53**(5), 517–527 (2011)
36. Hannah, S.T., Jennings, P.L., Bluhm, D., Peng, A.C., Schaubroeck, J.M.: Duty orientation: theoretical development and preliminary construct testing. Organ. Behav. Human Decis. Process. **123**(2), 220–238 (2014)
37. Haslanger, S.: Three moral theories, pp. 1–5 (2017)
38. 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 2022, pp. 1–7, Carnegie Mellon University Software Engineering Institute (SEI) (2021)
39. Hauptman, A.I., Schelble, B.G., McNeese, N.J., Madathil, K.C.: Adapt and overcome: perceptions of adaptive autonomous agents for human-AI teaming. Comput. Hum. Behav. **138**(2023), 107451 (2023)
40. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors **57**(3), 407–434 (2015). https://doi.org/10.1177/0018720814547570
41. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
42. Jones, T.M., Bowie, N.E.: Moral hazards on the road to the "virtual" corporation. Bus. Ethics Q. **8**(2), 273–292 (1998)
43. Kaber, D.B.: Issues in human–automation interaction modeling: presumptive aspects of frameworks of types and levels of automation. J. Cogn. Eng. Decis. Mak. **12**(1), 7–24 (2018)
44. Kaplan, A., Haenlein, M.: Rulers of the world, unite! the challenges and opportunities of artificial intelligence. Bus. Horiz. **63**(1), 37–50 (2020)
45. Kasper-Fuehrera, E.C., Ashkanasy, N.M.: Communicating trustworthiness and building trust in interorganizational virtual organizations. J. Manag. **27**(3), 235–254 (2001)
46. Kim, D., Vandenberghe, C.: Ethical leadership and team ethical voice and citizenship behavior in the military: the roles of team moral efficacy and ethical climate. Group Organ. Manag. **45**(4), 514–555 (2020)
47. Kim, P.H., Dirks, K.T., Cooper, C.D., Ferrin, D.L.: When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. Organ. Behav. Human Decis. Process. **99**(1), 49–65 (2006)
48. Kim, P.H., Ferrin, D.L., Cooper, C.D., Dirks, K.T.: Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. J. Appl. Psychol. **89**(1), 104 (2004)
49. Ko, Y.-H., Leem, C.-S.: The influence of AI technology acceptance and ethical awareness towards intention to use. J. Digit. Converg. **19**(3), 217–225 (2021)
50. Kohn, S.C., Quinn, D., Pak, R., De Visser, E.J., Shaw T.H.: Trust repair strategies with self-driving vehicles: An exploratory study. In Proceedings of the human factors and ergonomics society annual meeting, Sage Publications Sage CA: Los Angeles, CA, 1108–1112 (2018)
51. Kuntz, J.R.C., Kuntz, J.R., Elenkov, D., Nabirukhina, A.: Characterizing ethical cases: a cross-cultural investigation of individual differences, organisational climate, and leadership on ethical decision-making. J. Bus. Ethics **113**(2), 317–331 (2013). https://doi.org/10.1007/s10551-012-1306-6
52. Langer, M., König, C.J., Back, C., Hemsing, V.: Trust in Artificial Intelligence: comparing trust processes between human and automated trustees in light of unfair bias. J. Bus. Psychol. **2022**, 1–16 (2022)
53. Lawrence, M., Roberts, C., King, L.: Managing automation: employment, inequality and ethics in the digital age. Discussion paper presented at The IPPR Commission on Economic Justice, pp. 1–56, IPPR, London (2017)
54. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**(1), 50–80 (2004)

55. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**(10), 1243–1270 (1992)
56. Leveringhaus, A.: Ethics and Autonomous Weapons. Springer (2016)
57. James Lemoine, G., Hartnell, C.A., Leroy, H.: Taking stock of moral approaches to leadership: an integrative review of ethical, authentic, and servant leadership. Acad. Manag. Ann. **13**(1), 148–187 (2019)
58. Li, J., Huang, J., Liu, J., Zheng, T.: Human-AI cooperation: modes and their effects on attitudes. Telematics Inform. **73**(2022), 101862 (2022)
59. Lopez, T.B., Babin, B.J., Chung, C.: Perceptions of ethical work climate and person–organization fit among retail employees in Japan and the US: a cross-cultural scale validation. J. Bus. Res. **62**(6), 594–600 (2009)
60. Mabkhot, M.M., Al-Ahmari, A.M., Salah, B., Alkhalefah, H.: Requirements of the smart factory system: a survey and perspective. Machines **6**(2), 23 (2018)
61. Madhavan, P., Wiegmann, D.A.: Similarities and differences between human–human and human–automation trust: an integrative review. Theor. Issues Ergon. Sci. **8**(4), 277–301 (2007)
62. Malik, P., Pathania, M., Rathaur, V.K.: Overview of artificial intelligence in medicine. J. Fam. Med. Prim. Care **8**(7), 2328 (2019)
63. Matthews, G., Lin, J., Panganiban, A.R., Long, M.D.: Individual differences in trust in autonomous robots: implications for transparency. IEEE Trans. Human Mach. Syst. **50**(3), 234–244 (2020). https://doi.org/10.1109/THMS.2019.2947592
64. Mayer, D.M., Nurmohamed, S., Treviño, L.K., Shapiro, D.L., Schminke, M.: Encouraging employees to report unethical conduct internally: it takes a village. Organ. Behav. Human Decis. Process. **121**(1), 89–103 (2013)
65. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**(7788), 89–94 (2020)
66. 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)
67. 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)
68. Merritt, S.M., Huber, K., LaChapell-Unnerstall, J., Lee, D.: Continuous calibration of trust in automated systems. Missouri Univ-St Louis, St Louis (2014)
69. Mirzaeian, V.R., Kohzadi, H., Azizmohammadi, F.: Learning Persian grammar with the aid of an intelligent feedback generator. Eng. Appl. Artif. Intell. **49**(2016), 167–175 (2016)
70. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
71. Muir, B.M.: Trust between humans and machines, and the design of decision aids. Int. J. Man Mach. Stud. **27**(5–6), 527–539 (1987)
72. Nardo, M., Forino, D., Murino, T.: The evolution of man–machine interaction: the role of human in Industry 4.0 paradigm. Prod. Manuf. Res. **8**(1), 20–34 (2020)
73. Nesset, B., Robb, D.A., Lopes, J., Hastie, H.: Transparency in hri: Trust and decision making in the face of robot errors. In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, 313–317 (2021)
74. O'Neill, T., McNeese, N., Barron, A., Schelble, B.: Human-autonomy teaming: a review and analysis of the empirical literature. Hum. Factors **64**(5), 904–938 (2022). https://doi.org/10.1177/0018720820960865
75. Osborn, K:. The F-35 will soon be equipped with artificial intelligence to control drone wingmen. Business Insider. From https://www.businessinsider.com/f-35-artificial-intelligence-drone-wingmen-2017-1. Accessed 23 Mar 2023 (2017)
76. Othman, K.: Public acceptance and perception of autonomous vehicles: a comprehensive review. AI Ethics **1**(3), 355–387 (2021)
77. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. Hum. Factors **39**(2), 230–253 (1997)
78. Peterson, E., Mitchell, T.R., Thompson, L., Burr, R.: Collective efficacy and aspects of shared mental models as predictors of performance over time in work groups. Group Process. Intergroup Relat. **3**(3), 296–316 (2000)
79. Quinn, D.B., Pak, R., de Visser, E.J.: Testing the efficacy of human-human trust repair strategies with machines. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA, 1794–1798 (2017)
80. Ramaswamy, S., Joshi, H.: Automation and ethics. In: Springer handbook of automation, pp. 809–833. Springer, New York (2009)
81. Rix, J.: From Tools to Teammates: Conceptualizing Humans' Perception of Machines as Teammates with a Systematic Literature Review. In Proceedings of the 55th Hawaii International Conference on System Sciences. (2022)
82. Schaefer, K.E., Straub, E.R., Chen, J.Y.C., Putney, J., Evans III, A.W.: Communicating intent to develop shared situation awareness and engender trust in human-agent teams. Cogn. Syst. Res. **46**(2017), 26–39 (2017)
83. Schelble, B.G., Flathmann, C., McNeese, N.: Towards Meaningfully integrating human-autonomy teaming in applied settings. In Proceedings of the 8th International Conference on Human-Agent Interaction (HAI '20), Association for Computing Machinery, New York, NY, USA, 149–156. DOI:https://doi.org/10.1145/3406499.3415077 (2020)
84. Schelble, B.G., Flathmann, C., McNeese, N.J., Freeman, G., Mallick, R.: Let's think together! assessing shared mental models, performance, and trust in human-agent teams. Proc. ACM Hum. Comput. Interact. **6**(GROUP), 13:1-13:29 (2022). https://doi.org/10.1145/3492832
85. Schelble, B.G., Lancaster, C., Duan, W., Mallick, R., Mcneese, N,J., Lopez, J.: The effect of AI teammate ethicality on trust outcomes and individual performance in human-AI teams. In Proceedings of the Annual Hawaii International Conference on System Sciences, 322–331. (2023)
86. 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 **2022**, 00187208221116952 (2022)
87. Schwepker, C.H., Jr.: Ethical climate's relationship to job satisfaction, organizational commitment, and turnover intention in the salesforce. J. Bus. Res. **54**(1), 39–52 (2001)
88. Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 57–65
89. Seeber, I., Bittner, E., Briggs, R.O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A.B., Oeste-Reiß, S., Randrup, N., Schwabe, G., Söllner, M.: Machines as teammates: a research agenda on AI in team collaboration. Inf. Manage. **57**(2), 103174 (2020). https://doi.org/10.1016/j.im.2019.103174
90. Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T.: Promises and lies: Restoring violated trust. Organ. Behav. Hum. Decis. Process. **101**(1), 1–19 (2006)

91. Smith, J.A., Osborn, M.: Chapter 4: Interpretive phenomenological analysis. In: Qualitative psychology: a practical guide to methods, pp. 53–80. Sage Publications, London (2003)
92. Sosik, J.J., Chun, J.U., Ete, Z., Arenas, F.J., Scherer, J.A.: Self-control puts character into action: examining how leader character strengths and ethical leadership relate to leader outcomes. J. Bus. Ethics **160**(2019), 765–781 (2019)
93. Sotala, K., Yampolskiy, R.V.: Responses to catastrophic AGI risk: a survey. Phys. Scr. **90**(1), 018001 (2014)
94. Sparks, J.R., Pan, Y.: Ethical judgments in business ethics research: definition, and research agenda. J. Bus. Ethics **91**(2010), 405–418 (2010)
95. Sutton, G.W., Washburn, D.M., Comtois, L.L., Moeckel, A.R.: Professional ethics violations gender, forgiveness, and the attitudes of social work students. J. Coll. Charact. **7**(1), 1–7 (2006)
96. Sweeney, B., Arnold, D., Pierce, B.: The impact of perceived ethical culture of the firm and demographic variables on auditors' ethical evaluation and intention to act decisions. J. Bus. Ethics **93**(2010), 531–551 (2010)
97. Tambe, P., Cappelli, P., Yakubovich, V.: Artificial intelligence in human resources management: challenges and a path forward. Calif. Manage. Rev. **61**(4), 15–42 (2019)
98. Tan, H.H., Lim, A.K.H.: Trust in coworkers and trust in organizations. J. Psychol. **143**(1), 45–66 (2009)
99. Textor, C., Zhang, R., Lopez, J., Schelble, B.G., McNeese, N.J., Freeman, G., Pak, R., Tossell, C., de Visser, E.J.: Exploring the relationship between ethics and trust in human-artificial intelligence teaming: a mixed methods approach. J. Cogn. Eng. Decis. Mak. **2022**, 15553434221113964 (2022)
100. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al.: Human–computer collaboration for skin cancer recognition. Nat. Med. **26**(8), 1229–1234 (2020)
101. Tzafestas, S.G.: Roboethics: fundamental concepts and future prospects. Information **9**(6), 148 (2018)
102. Valentine, S., Fleischman, G.: Ethical reasoning in an equitable relief innocent spouse context. J. Bus. Ethics **45**(2003), 325–339 (2003)
103. de Visser, E.J., Pak, R., Shaw, T.H.: From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. Ergonomics **61**(10), 1409–1427 (2018)
104. de Visser, E.J., Pak, R., Neerincx, M.A.: Trust Development and Repair in Human-Robot Teams. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17), Association for Computing Machinery, New York, NY, USA, 103–104. DOI:https://doi.org/10.1145/3029798.3038409 (2017)
105. Walliser, J.C., Mead, P.R., Shaw, T.H.: The perception of teamwork with an autonomous agent enhances affect and performance outcomes. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 61(1):231–235. https://doi.org/10.1177/1541931213601541 (2017)
106. Weger, K., Matsuyama, L., Zimmermann, R., Mesmer, B., Van Bossuyt, D., Semmens, R., Eaton, C.: Insight into user acceptance and adoption of autonomous systems in mission critical environments. Int. J. Human Computer Interact. **2022**, 1–15 (2022)
107. Wilson, H.J., Daugherty, P.: Collaborative intelligence: humans and AI are joining forces. Harv. Bus. Rev. **96**(4), 114–123 (2018)
108. Winfield, A.: Ethical standards in robotics and AI. Nat. Electron. **2**(2), 46–48 (2019)
109. Winfield, A., Jirotka, M.: Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos. Trans. R. Soc. A **376**(2133), 20180085 (2018)
110. Wynne, K.T., Lyons, J.B.: An integrative model of autonomous agent teammate-likeness. Theor. Issues Ergon. Sci. **19**(3), 353–374 (2018). https://doi.org/10.1080/1463922X.2016.1260181
111. Young, A.D., Monroe, A.E.: Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. J. Exp. Soc. Psychol. **85**(2019), 103870 (2019)
112. Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V.R., Yang, Q.: Building ethics into artificial intelligence. arXiv preprint arXiv:1812.02953 (2018)
113. Zhang, K., Aslan, A.B.: AI technologies for education: Recent research & future directions. Computers Educ. **2**, 100025 (2021)
114. Zhang, R., McNeese, N.J., Freeman, G., and Geoff Musick: An Ideal Human" Expectations of AI Teammates in Human-AI Teaming. Proceedings of the ACM on Human-Computer Interaction 4, CSCW3 (2021), 1–25