Modeling and Guiding the Creation of Ethical Human-AI Teams

Christopher Flathmann
Clemson University
Clemson, SC, USA
cflathm@clemson.edu

Beau G. Schelble
Clemson University
Clemson, SC, USA
bschelb@clemson.edu

Rui Zhang
Clemson University
Clemson, SC, USA
rzhang2@clemson.edu

Nathan J. McNeese
Clemson University
Clemson, SC, USA
mcneese@clemson.edu

ABSTRACT

With artificial intelligence continuing to advance, so too do the ethical concerns that can potentially negatively impact humans and the greater society. When these systems begin to interact with humans, these concerns become much more complex and much more important. The field of human-AI teaming provides a relevant example of how AI ethics can have significant and continued effects on humans. This paper reviews research in ethical artificial intelligence, as well as ethical teamwork through the lens of the rapidly advancing field of human-AI teaming, resulting in a model demonstrating the requirements and outcomes of building ethical human-AI teams. The model is created to guide the prioritization of ethics in human-AI teaming by outlining the ethical teaming process, outcomes of ethical teams, and external requirements necessary to ensure ethical human-AI teams. A final discussion is presented on how the developed model will influence the implementation of AI teammates, as well as the development of policy and regulation surrounding the domain in the coming years.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Human-AI Teamwork, Human-AI Ethics, Artificial Intelligence, AI Ethics

ACM Reference Format:
Christopher Flathmann, Beau G. Schelble, Rui Zhang, and Nathan J. McNeese. 2021. Modeling and Guiding the Creation of Ethical Human-AI Teams. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462573

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462573

1 INTRODUCTION

With new innovations in the field of Artificial Intelligence (AI) being achieved every day, the number of roles AI can facilitate, or has the potential to, is rapidly advancing [27]. Unfortunately, this advancement has not always been met with similar levels of passion and productivity in regulating and monitoring the ethical use of these technologies [16]. Specifically, this paper adheres to the definition of ethics being the "various ways of understanding and examining the moral life" [3]. The past ignorance towards ethics has recently led to a rapid increase in the production of research centered around the ethical creation and operation of AI systems [21, 103]. The concerns that motivate this research become especially poignant when the potential for close human-AI collaboration becomes a consideration [6]. As AI technology becomes more intertwined with society, the possibility for this interaction gets all the more likely, thus leading to the conclusion that the importance of understanding the role ethics plays in human-AI interaction will similarly increase [49].

While ethics in human-AI interactions is being studied, a gap has formed around a more specific field of ethics in human-AI teamwork. Unlike general human-AI interaction, human-AI teamwork creates a more complex relationship between humans and AI as: (1) the compositions of these teams are often different from the 1-1 relationship often seen in traditional human-AI interaction research, (2) the lifespan of teams is generally longer, increasing the importance of meeting human factors requirements, and (3) AI takes on the larger more human-like role of a teammate, as opposed to a simple tool [73]. Thus, it would be irresponsible to assume that the research in general ethical human-AI interaction would fully translate to the field of human-AI teamwork. Similarly, while human-human teaming research has demonstrated the importance of ethics [8], the addition of AI teammates adds additional complexities and requirements to the teamwork process [29]. Thus, a similar assumption as above should not be made, as ethical human-human teaming research cannot be blindly transplanted into human-AI teaming. Due to the inability to make these assumptions, a research gap has emerged that requires the outlining and guidance of ethical human-AI teamwork.

Therefore, this paper pursues the research goal of outlining the ethical requirements of both AI systems and teamwork while also drawing attention to the unique aspects of human-AI teaming that separate them from their human-only counterparts. The interactions of these requirements are outlined, demonstrating more

specific principles and requirements that are required to achieve ethical human-AI teamwork. These interactions are communicated through the creation of a model that contextualizes the requirements and outputs of ethical human-AI interaction in a human-AI team (HAT), thus resulting in a model that acts as a guide to building ethical human-AI teamwork. This model is informed by a wide variety of past literature that has been synthesized to produce the novel model, which follows the guidelines for producing impactful conceptual models [66, 105]. This paper also takes the opportunity to demonstrate how the existence and need of ethical human-AI teaming may lead to changes in the current path of research on developing and implementing AI technologies, as well as the policies that will govern AI's integration into society. Thus, by achieving the research goal outlined in this paper, research into ethical AI systems can become more comprehensive and robust since the impacts and requirements of human-AI teaming have been properly contextualized.

2 BACKGROUND

Modeling the interaction of ethics in human-AI teamwork is a complex task that requires an interdisciplinary research perspective. Specifically, a strong understanding of the importance of ethics to both the field of human-human teamwork, as well as AI, is required. Additionally, a strong grasp of human-AI teaming will allow the above to be applied to the state of the art in teamwork science. With the review of these three disciplines as foundations, an interaction model can be created that meets the requirements of ethical AI, ethical teamwork, and human-AI teamwork, thus resulting in the creation of a model depicting ethical human-AI teamwork.

2.1 Ethics in Human-Human Teamwork

With ethics and ethical behaviors being vital in the daily lives of humans, it would make sense for ethics to become an important factor in teamwork. However, the field of teamwork necessitates a multidimensional understanding of ethics as teams must act responsibly to both those external to the team [57] and those internal to the team [44]. In terms of external interaction, similar to individuals, teams must contemplate how their actions and decisions will affect others in society [97]. Additionally, teamwork, being the collaboration of diverse individuals, often results in the interaction of differing ethical perspectives within a team, where new team members can impact their teams and their environments with their ethical ideologies [104]. Unfortunately, the complexities that arise with teamwork, which carry over to teamwork ethics, can often make it difficult for teams to operate in accordance with every individual team members' ideal ethical standards [84, 95]. These challenges facing ethical teamwork have become far more complex and critical as global teams, made up of multiple cultures, have been brought together by a variety of advanced technologies [86]. Therefore, the impact of ethics and morality on modern teamwork, enhanced by the existence of intra-team interaction, needs to be properly understood to ensure teams are not harmed by unethical behaviors or by misunderstandings around ethical standards.

The most apparent examples of these confrontations exist in the medical field, where teams are constantly tasked with scenarios that require ethical considerations [4, 28, 38]. Medical teams are a clear and practical example of how ethics and teamwork are intertwined and affect those within the team and those the team interacts with [8]. Due to the high stakes nature of medical teams, it is imperative that teams are able to act both quickly and ethically, meaning time cannot be wasted on ethical debates [55, 78]. The significance of ethics in the medical domain has also led to the rise of ethical consultation, where oversight and advice can be provided to ensure high ethical standards are not only proposed but continuously met by medical professionals [1]. Therefore, the efficient operation of medical teams requires the prior establishment of an ethical code or standard that ensures ethical conduct is a foundational priority, rather than an afterthought, which has led ethical training to an important part of a complete education [23]. Without the consideration of ethical standards, ethical oversight, and ethical education, it would be extremely difficult, if not impossible, for medical teams to confidently execute vital and ethical medical care.

In addition to the importance of teaming in the medical industry, the explosion of the technology sector in recent decades has been similarly met with an increased interest and importance of ethics within technology teams and organizations [67, 75, 99]. Due to this increased growth, the necessity of ethical training has been further identified for the tech sector, especially in leadership roles [69, 98]. Other contexts which have demonstrated a strong relationship between ethics and teamwork include military operations [56] and disaster response [60]. Similar to the medical domain, technology and military domains also see the key to ethical education and establishing one's ethical compass as occurring early in one's career, preferably as a student, thus allowing an ethical foundation to be created before entering workforce teams [43, 90]. The variety of domains in which ethics and teamwork are shown to depend on each other demonstrate that ethics and teamwork are deeply intertwined concepts, where teamwork should not effectively exist without ethics.

Fortunately for teams, when the interaction of ethical ideologies within a team is handled correctly, it can be an important factor related to high performance [51]. Ethical factors and considerations of individual team members can become a point of harmony through which teamwork is further improved [2, 82]. Additionally, ethical environments can become highly impactful towards other teaming factors, such as creativity [115] and trust [88], which is an essential factor in long-term teaming [30, 54]. Due to the importance of ethics to teamwork outlined above, recent innovations have also seen teamwork become a tool to create education in the field of ethics. Essentially, by learning about ethics and ethical principles in a team setting, individuals become better equipped to contribute to and understand the differing ethical ideologies present in the future teams they may be a part of [68]. Not only do individuals learn ethics through this education, but they also learn the process of building a shared ethical understanding with a team, which is a vital skill to acquire in the modern workforce [80]. Additionally, this method provides the opportunity for applied demonstration by accomplished, ethical teams who have the ability to lead by example [42].

The above literature provides a detailed account of how ethics both broadly and deeply affect teamwork, and any domains that utilize teams. If teams are unable to build these ethical foundations, then they may not only perform worse, but their actions could have

extremely harmful consequences to others in society. Incorporating and practicing ethics in teamwork is not only a benefit, but a necessity to ensuring the longevity and morality of a variety of industries, including medical, military, and technology to name a few.

2.2 Ethics in Artificial Intelligence

Rising alongside AI technology are concerns over the unethical use of AI and the unethical actions made by these technologies. Previous research has shown that the use of malevolent AI may cause security issues and have a negative influence on humans [12, 85]. Hence, how to address these ethical issues has been a research focus in AI development and human-computer interaction [34, 112].

Researchers have been working on creating ethical guidance for AI development over the past several years [53]. One possible approach for building AI ethics is to include functionality in AI systems that can make ethical decisions either by programming ethical principles into the AI system or by using machine learning to learn human ethical and unethical behaviors [34]. The former may require AI development teams to have ethicists working on ethics integration [71], while the latter method indicates the building of *ethics bots*, which learn the ethical preference of users and then apply it to the user's machine, e.g., autonomous vehicle [34]. However, this type of implementation may actually lead to unethical AI behavior if users behave unethically. Additionally, though a body of research has investigated ethical principles and guidelines in AI development, the principles proposed have a high diversity. A previous study analyzed 84 documents that include ethical principles and concluded that a convergence emerged around five ethical principles, transparency, justice and fairness, non-maleficence, responsibility, and privacy [53]. A recent study also utilized traditional, human-derived ethical concepts, i.e., virtue ethics, deontological ethics and consequentialist ethics, to develop ethical frameworks. The traditional ethical notions indicate that AI's behavior is justified as ethical when the behavior conforms to virtue, obligation, or consequences in a moral way [25].

Another approach applied in the development of AI ethical frameworks is to shift the responsibilities onto humans [9, 101]. In addition to the ethical design and development of AI themselves, the ethical use of AI is also an essential component in AI ethics, which gives humans more responsibility in the development and implementation process. The Department of Defense has proposed an ethical principle framework targeting the ethical use of AI, including humans being responsible, and the development and deployment of AI being equitable, traceable, reliable, and governable [9]. Importantly, humans have to manage what AI systems are to be developed and what ethical policy should be followed in the use of AI systems [37]. Thus, policies that define human responsibilities in the use and development of ethical AI are necessary [100].

Similar to the unethical behaviors of humans which harm others, AI systems without ethical principles implemented could cause harm to human security and human trust. One such example is in the domain of autonomous vehicles, which have a risk of harming humans on the road. According to an ethical risk assessment in robotics, ethical hazards include loss of trust [107]. Previous work has indicated that ethical governance, a collection of processes and values, is vital in building trust between humans and AI [108]. That is to say, without ethical guidelines and principles, it may be difficult for humans to develop trust in AI.

Important to success in ethical AI, researchers have explored various strategies to apply ethical principles to real implementation. For example, a previous study constructed a typology to help the implementation of AI principles into practices, which described how to apply ethical principles (beneficence, non-maleficence, autonomy, justice and explicability) using various methodologies and tools [77]. Another study evaluated a business ethics strategy into ethical AI development [102]. While plenty of work has explored the implementation of ethical frameworks in AI, that does not mean that the path towards implementing practical AI is entirely clear. One reason is that it is difficult to translate ethical principles into successful practice due to infancy of methods used to implement ethical principles practically [76]. Moreover, high-level ethical principles may cause misunderstanding in real implementation, leading to the opposite outcome of building an ethical AI [100]. According to these studies, detailed ethical principles are more likely to be translated into practice [76, 106]. To solve this dilemma, previous research has proposed that understanding different values and goals in real practice and specific contexts plays an essential role in bridging the gap between ethical theories and implementation [76].

2.3 Human-AI Teamwork

Teams have long been utilized to accomplish work of varying complexity from a multitude of different contexts [91]. Naturally, humankind has always sought to enhance the performance of teams and thus joining them with current technology is a constant endeavor. For example, the new capabilities presented by technologies such as the internet and the personal computer in the 1990's allowed for the creation of a new type of team, the virtual team [62]. These virtual teams were defined by team members that were not immediately present with their peers, having varying amounts of dispersion across temporal and or spatial bounds [33]. As would be expected, a new type of team also came with a multitude of challenges for practitioners and researchers alike to confront. Specifically, nonverbal cues could be lost when virtual team members interact with one another while also increasing the level of abstraction the team members must confront [35]. Nearly two decades of research and applied use of virtual teams solved many of these issues and now virtual teams and remote work were the immediate answer once the COVID-19 pandemic began [13]. Technologies have made another major advancement with the democratization and widespread use of AI, and many have already identified the potential benefits AI presents to teams, not just as a helpful assistant or tool, but as a full-fledged member of the team.

AI joining together with humans as teammates introduces a new type of team known as a HAT, which has at least one human and one AI teammate [73]. HATs share the same traditional requirements of human-human teams (interdependence, shared goals, see [32]), but include a few additional requirements regarding the AI teammate in order to be defined as a true HAT [81]. While AI is still a growing technology, humans have high expectations for these AI teammates

and are interested in their future implementation [114], and these expectations will become especially important when designing additional technologies to support HATS [36]. The AI teammate of a HAT must: (1) possess at least partial levels of decision making autonomy (Level 5 and above, see [81, 83]), (2) occupy a distinct role within the team, and (3) be interdependent in its activity and outcomes. HATs are also distinctly different from traditional human-human teams as transparency [74], reliability [18], and autonomy level [110, 111] become important predictors of team performance [81]. Yet while HATs are distinctly different from traditional human-human teams in many ways, they still retain an empirical need for human elements such as trust.

Trust is a particularly essential component to successful HATs, and trust's unique relationship to ethics makes it especially relevant to the current paper. Existing studies have empirically shown the importance of the relationship between trust in AI teammates and team performance [72], and additional work has harped on its importance in ethically fraught subjective contexts such as healthcare. In this context of teaming, autonomous agents should be able to satisfy a role within the team and gain trust over time as ethics is directly related to trust [47, 107, 108], which will in turn affect overall team performance. Due to AI's general use in more objective contexts, relevant research must be conducted to ensure AI's acceptance and utilization in teams that focus on subjective tasks. Providing methods for increasing the explainability and fairness of these systems, which have been identified as necessary in human-AI teaming [19, 87], may better equip HATs to handle subjective tasks.

Creating ethical AI can lead to enhanced trust in HATs and subsequently enhanced team performance and fairer team decision making. Human members of HATs have been shown to produce better shared mental models when AI teammates display characteristics similar to their personality [48], and culture has also been shown to play a significant role in the amount of trust humans place in their AI teammates [20]. These findings show that ethical AI can be developed in such a way as to not only maximize the fairness of decisions but also maximize the level of trust between the AI and its human teammates. This assertion is in addition to studies directly linking ethical behavior to trust in human-AI interactions [108]. For example, recent research has focused on outlining the steps needed to develop value-based ethical AI systems [79]. Such advanced value-based ethical AI systems could be produced that do their best to align their values with their human teammates to maximize trust and performance within the team, while remaining within the bounds of ethicality most applicable to the situation. While developing ethical AI is far easier said than done [5], the endeavor is incredibly important to the future of HATs and numerous other fields as well, and with it, the challenges posed by this new technology can be met and surpassed in the coming decades.

3 MODELING ETHICAL HUMAN-AI TEAMWORK

Creating a comprehensive model of ethical human-AI teamwork requires the consolidation and interaction of the previous literature review [66, 105]. The resultant model this paper creates through the method above is presented in Figure 1. Figure 1 contains three distinct parts that can be isolated and discussed separately. First, the societal requirements necessary for ethical human-AI teamwork are represented through the grey hexagons; these requirements are seen as external and societal factors that HATs may not have direct control over. Second, the large, upper rectangle contains the representation of the constructive process of utilizing ethics within a team, ultimately resulting in a team-wide ethical ideology. This section is the core representation of ethical human-AI teamwork, as it demonstrates the complex interaction of ethics in teamwork when mediated by the introduction of AI. Finally, the outputs of ethical teamwork are shown, with the team's internal outputs being represented by the left blue rectangle and the team's external outputs to society being shown in the right red rectangle. These outputs serve as both a motivating factor for ensuring ethical teamwork as well as a form of awareness in understanding the important outputs to which ethical HATs can be evaluated. Thus, when these elements are joined and related to each other, the comprehensive model of ethical human-AI teamwork shown in Figure 1 can be created. This model serves as a guide to building ethical HATs, which is a necessity for achieving effective human-AI teaming. The following elaborates on how ethical/unethical human-AI teaming principles can significantly benefit/harm the potential of HATs in society.

3.1 Requirements for Ethical Human-AI Teamwork

External requirements are vital in shaping ethical human-AI teamwork by setting behavior baselines for humans and AI in HATs. Previous literature has pointed out that humans, as a core component of ethical models in human-AI teamwork, undertake the responsibilities to form, regulate and guide the development and deployment of ethical AI [9, 101]. This section focuses on what external requirements are needed to facilitate the construction of ethical HATs.

3.1.1 Ethical Education and Ethical AI Standards. Education on ethics in HATs and ethics regulation and standards lay the cornerstone of ethical human-AI teamwork. Ethical training is taken as an essential mechanism in the education of various domains [43, 89, 90]. Though the education of AI ethics is still in the early stage of development, the field is considered as an important aspect of the AI education that ought to be integrated into college education in a systematic way [10, 40]. Only when individuals have a basic understanding of AI ethics, and ethical behaviors and decisions, are they able to make ethical decisions in the development or deployment of AI in the real-world [39]. In addition to including ethical knowledge in curriculum, workshops are an effective way to improve the delivery of ethical knowledge to people who are going to be involved in ethical considerations as part of their career [68]. Equipped with ethical knowledge, individuals and AI also need ethical standards to direct and guide the team to behave in an ethical way. Ethical standards are defined as principles that address ethical issues and concerns as well as point out the potential harm caused by unethical behaviors [107]. These principles act as a reference for individuals to follow, which assists the development of shared ethical ideologies in human-AI teamwork.

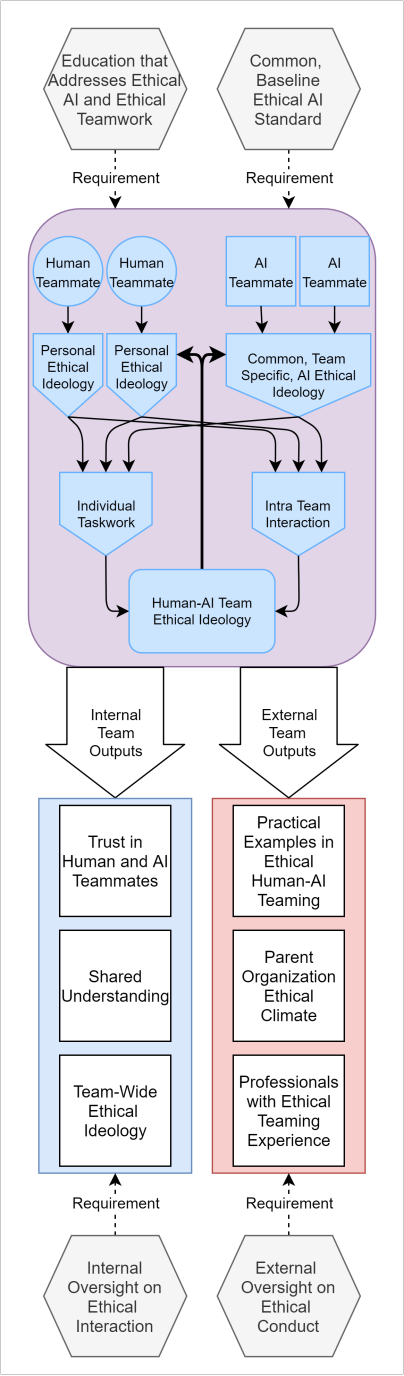

Figure 1: A model depicting the teaming process in regards to ethical human-AI teamwork. Human-AI teaming outputs are shown (red and blue rectangles). Requirements for ethical Human-AI teaming are shown as grey hexagons. The interactions of teamwork ethics are also shown in the large, upper rectangle.

3.1.2 Internal and External Oversight on Ethical Issues. In addition to ethical knowledge and ethical standards, internal and external oversight on ethical issues in collaborative environments are additional necessities in building ethical human-AI teamwork. By applying internal and external oversight, HATs are monitored in their pursuit of ethical human-AI teamwork and shared ethical ideologies. Internal oversight refers to an internal member or committee ensuring that human teammates and AI teammates conduct ethical actions or make ethical decisions in teams. Though the internal refers to one or more team members for most cases in HATs, it could also describe one or more individuals who belong to the same organization or community, but are not involved in the HAT. One such example is how hospitals ensure ethicality. The internal oversight may come from the hospital but not the team consisted of healthcare personnel and AI. Internal oversight makes sure each individual behaves ethically in HATs, which contributes to ethics on a team-level, including enabling human teammates to build trust with AI teammates, building a shared understanding in HATs, and more importantly, constructing team-wide ethical ideologies.

In contrast to internal oversight, external oversight is ensured by people who are excluded from the teams. For instance, the development and deployment of AI teammates in industry should be supervised by an ethicist. Moreover, the ethicist who oversees the deployment of ethics in human-AI teamwork has to be at a high level of hierarchy in the community to ensure the *real* implementation of ethical principles. Importantly, the cost of a violation of AI ethics needs to be considered in the implementation of ethics in human-AI teamwork. In other words, ethicists need to identify specific penalties at the beginning of the AI development or deployment. A penalty rule is beneficial to decide the penalty for a specific behavior, e.g., setting different instances of not following ethical standards with a corresponding penalty. External oversight provides a vita security in ensuring the implementation of ethical principles and guidance in real-world practice. Furthermore, it builds up a robust and solid ethical environment in large organizations using a systematic ethical framework designed for them specifically. Professional ethicists ensure that HATs are guided in the proper direction using their accumulated ethical teaming experience, as well as their expert knowledge.

In summary, four requirements regarding HATs are proposed to approach high-quality ethical human-AI teamwork: (1) individuals in HATs need to learn how to address ethical issues in human-AI teamwork, (2) common ethical AI standards are required to make sure AI is implemented and behaves ethically, (3) internal oversight should be implemented to ensure each individual's ethical behavior, and (4) external oversight is necessary in regard to the implementation of ethics in HATs, which is usually covered by ethicists.

3.2 Building an Ethical Human-AI Team Through Teammate Interactions

Building ethical HATs requires a strong understanding of how ethical ideologies within a team interact and positively build on each other to create a team's ethical ideology. Figure 1 (specifically the large, upper rectangle) illustrates these interactions where individual ideologies are eventually consolidated into an overall ethical

team ideology. The interactions are based on the known interactions of ethical human-human teamwork when taking into consideration the factors and requirements presented by ethical AI and human-AI teamwork.

3.2.1 Different Individual Ethical Ideologies. Before understanding the consolidation process, it is important to understand the specific, individual ideologies that are present, which are shown at the top of the large, upper rectangle. First, each human is identified as having an individual ethical ideology, which is not guaranteed or even likely to be consistent from person to person [11, 61]. While these individual differences may seem unreliable, they are essential to preserving team diversity and individuality, a vital factor in effective teaming [50, 52]. On the other hand, AI teammates should be designed to share a specific ethical ideology, which is an important aspect in ensuring an ethical AI standard [107] and preventing possible disagreements between AI systems [70]. This shared ideology could be a consistent set or list of duties or virtues that all AI teammates remain consistent with. Additionally, the shared ideology ensures a level of ethical consistency across AI teammates, which is a necessity to guaranteeing the implementation of trusted AI systems [94]. These individual ideologies are the building blocks for ethical human-AI teamwork; however, the mere existence of these internal ideologies is not enough to build ethical HATs.

3.2.2 Ethical Ideologies Influence Teaming Actions. The ethical ideologies outlined above physically manifest through teamwork actions, which is represented by the conjoining of ethical ideologies from both human and AI teammates into two different categories shown in Figure 1. Specifically, the individual, role-based tasks one performs (left of the Figure 1) and one's interactions with their teammates (right of Figure 1) are they two key teamwork actions that allow teammates to express their ethical ideologies in a practical way. It is the result of these actions that yields the creation of a team ethical ideology, meaning practice, not just principle, is key to achieving ethical human-AI teamwork. While the emphasis of practice in ethical human-machine interaction is documented [58, 76], the model in Figure 1 demonstrates how practice becomes a broader and more varied concept within human-AI teamwork. Unlike general human-AI interaction, which sees an emphasis on the task related actions and decisions made by individual systems (the action represented on the left of Figure 1 [7, 22], human-AI teamwork adds the more complex action of intra-team interaction (the action pictured on the right of Figure 1. The existence of these additional interactions between team members means that one's ethical ideology is given additional opportunities and mediums to manifest. This is especially important to human-AI teaming ethics, as the context specific nature of AI ethics is often seen as a challenge or limitation [22, 64]. Therefore, specific considerations need to be built into AI teammates' ethical ideologies if they are going to be constructive, regardless of the teaming action being taken.

3.2.3 Creating a Shared HAT Ethical Ideology. The ultimate result of human-AI teaming actions driven by ethical ideologies will be a common ethical ideology, which is representative of a team's ideology. It is important to note that this team-level ideology is not a replacement for any individual's ideology, but rather a wholly new ideology that is owned by the team. Importantly, as depicted in the bottom of the large, upper rectangle of Figure 1, the team ethical ideology is able to continuously affect individual ideologies. This effect is the result of team members continuously learning from their fellow teammate ideologies through natural teaming actions and processes. This means that human teammates will be able to continuously learn about both human and AI teammate ethical ideologies through the teamwork process.

Similarly, AI teammates should be able to learn from this team-level ethical ideology, thus allowing AI team members to build on their foundational ethical ideology and tailor it to the specific human teammates they work with. Although this results in AI ethical ideologies that are different from team to team, it is a vital step as human-AI teaming, like AI ethics, is contextually dependent [92], meaning no two teams should be limited to having the exact same ethical ideology. This additional, team specific, interaction also means that the common AI ethical ideology shared by AI teammates needs to be able to add additional ethical constraints to remain consistent with the team-wide ideology. On that note, it is important that the process of building this team-wide ethical ideology should always be constructive, meaning, the interactions from the team-wide ideology should never grossly violate one's individual ethical foundation, regardless of one being a human or AI teammate. Thus, the ethical foundation of both human and AI teammates is ensured through a teaming process that consistently prioritizes the constructive development of human-AI teaming ethics.

3.3 The Outputs of Ethical Human-AI Teams

As detailed in the previous two sections covering the ethical human-AI teamwork model seen in Figure 1, the unique and purposeful interactions between teammates combined with the proper education and oversight can produce extremely high-quality HATs. The current section overviews how the teammate interactions and supporting requirements affect and produce the internal and external team outputs, then explores how such outputs are inevitable but are completely dependent upon the ethics of said inputs and oversight.

3.3.1 Internal Team Outputs. Internal team outputs (pictured in the left, blue rectangle of Figure 1) are the product of the intra-team interactions presented in the prior section and make up the immediate individual and team level benefits seen in the model. The intra-team interactions produce a team-wide ethical ideology and allow for improvements in team characteristics such as trust between human and AI team members, and enhanced shared understanding amongst teammates. Building robust trust between team members in HATs can come from human team members sharing their personal ethical ideology with each other and their AI teammates sharing their common team specific AI ethical ideology. This sharing and continuous task based interaction allows both the AI and human teammates to adapt to one another and create the human-AI team ethical ideology, which would closer align the human and AI team members personal ethical viewpoints improving transparency, significantly enhancing trust between the two [74]. The model also enables teams to improve their level of shared understanding, in the vein of team cognition, which is a construct in teams consisting of their team level knowledge and structure [14, 26]. The model enhances the teams shared understanding by

aligning a team's ethical perspectives over time to create the team-wide ethical ideology, which, much like the enhancements to trust, represents individual team members that are more closely aligned on a personal level [48]. Finally, the driving force behind the other outputs is an output itself, the team-wide ethical ideology. This team ideology empowers the team to make collectively ethical decisions that align to and further reinforce the existing ethical perspective shared by the team members.

3.3.2 External Team Outputs. External team outputs (pictured in the right, red rectangle of Figure 1) are another product of intra-team interactions but consist of the secondary consequences of the model, affecting the parent institution and other teams in the same domain. This change starts with individuals being provided with valuable experience working in highly ethical HATs that are a consequence of the model. Such teams conduct themselves in an ethical and appropriate manner throughout the process of performing the task and when deliberating and making team decisions together. The decisions and actions made by the teams are ethical, and the process the team took to reach their outputs provides the individuals with positive and valuable experience regarding ethical teaming. These positive experiences provide a host of additional benefits to both the organization as well as the individual regarding future teaming as the positive experiences lead to enhanced trust in future HATs [45, 46] and improved performance if the task is spatial in nature [17]. The model also improves the ethical climate of the parent organization and their respective domain by requiring the development of a series of educational programs, baseline expectations, and responsible ethical oversight policies. These policies are also part of what helps produce individuals with ethical teaming experience that also contribute to significantly improving the ethical climate of an organization the longer such policies and teams exist. Finally, teams created through the insight of the model provide real-world examples of ethical human-AI teaming. Parent organizations can then utilize the wide variety of successful HATs as part of continuing education to improve both the performance and culture of their workforce [59], and for managers to review and critique in order to identify any potential improvements in both performance and ethical standards.

3.3.3 Output Relationships with Requirements. The outputs reviewed here are inevitable given the nature of the model seen in Figure 1; however, whether or not these outputs are positive are entirely dependent upon the inputs or, requirements, outlined in the previous section. These inputs can just the same be faulty and unethical but will produce a shared ideology amongst the team members just the same as ethical inputs would. The nature of this principle reinforces the importance of the inputs, but also highlights the duality of the inputs location. For example, even if the education and common baseline ethical standard for AI are unethical, the unethical shared ideology will still be caught and have the potential to be rectified as long as proper internal and external oversight is provided. Therefore, it is truly the ethical aspects of the individual *teams* that determine if these are high-quality outputs, and the requirements exist to ensure those teams begin on the right path (top requirements) and stay on the right path (bottom requirements). Additionally, these requirements may be tweaked, as alluded to at the end of the previous paragraph (management review of teams), to maximize the internal and external outputs. Such changes to the requirements may seek to target maximization of internal or external factors the parent organization or broader domain deem most important, such as further enhancing shared understanding, trust, and or providing more ethical teaming experience.

In summary, the outputs of the model presented in the current paper are a natural result of maximizing team interactions to create a high quality human-AI team ethical ideology. This ideology in turn allows for improved trust, shared understanding, and a team-wide ethical ideology, in addition to providing practical examples of ethical human-AI teaming, improving the ethical climate of the teams parent organization, and giving the individual team members vital ethical teaming experience. As these benefits are a natural consequence of the functioning of the model, it is essential that the requirements be implemented properly and carefully as they can still produce an unethical human-AI team ideology if the education, ethical AI standards, and oversight are improper. Fortunately, the model has the ability to account for mistakes in education and AI model training with proper oversight, enabling changes and tweaks to be made to the requirements in order to maximize the most important outputs to the parent organization.

4 DISCUSSION

While the model above presents a guide for building ethical HATs through the synthesizing of a model, the model in Figure 1 needs to be related to its role in society to ensure it properly influences other disciplines surrounding AI. Specifically, this paper finds it important to understand how (1) the model should influence the design and development of HATs, as well as (2) influence the policy and regulation surrounding the rapidly advancing technology. For (1), ensuring future developers and research consider ethics when building HATs is a must as HATs will become a common use of AI technologies. For (2), recent incidents in the AI sector have demonstrated the need for explicit regulation and policy to ensure the ethical use of AI technologies, and these regulations should extend to that of human-AI teamwork.

4.1 Building Ethical AI Teammates for Human-AI Teams

The model in Figure 1 outlines the ethical teaming process as well as the outcomes of ethical teamwork; understanding how to operationalize these processes is additionally necessary to build ethical HATs. First, the most critical aspect is building the ability for AI teammates to perform the teamwork actions depicted (i.e. individual action and teamwork communications) with a consideration for ethics. While it is still controversial as to whether or not AI systems should be permitted to tackle tasks with such ethical ambiguity with proprietary ethical decision-making capabilities [31], fully preventing them from incorporating ethical decision making skills would heavily reduce the number of potential roles they could fill, and consequentially, the performance they could have in a HAT. The balancing of these technical benefits of AI systems with ethical concerns is not a new challenge in the AI space [41], and this paper discusses possible development goals and methods for ensuring the safe integration of these capabilities into AI teammates.

Therefore, building ethical HATs that have the potential to contribute to shared ethical ideologies, as depicted in Figure 1, will require the addition of ethical decision making skills in addition to the AI teammates' task related skillset. While this may open the possibility of intentionally unethical behavior from AI teammates [103], the possible benefits are undeniable with the potential to heavily benefit shared understanding, team cohesion, and trust through a shared ethical ideology. While the actual method for incorporating ethical knowledge into these systems is debated, this paper believes the perspective taken by Wright 2019 concerning duty and Kantian ethics would be best as it would allow a more natural creation of the shared ethical ideology used by all AI teammates in a single HAT. However, despite this paper's support for building ethical decision making abilities into AI teammates, this does not mean that these systems should be blindly integrated, regardless of the method taken for building AI ethical ideologies. It will also be necessary to develop automatic systems for monitoring the ethics and fairness of the AI teammates and ensuring they are inline with human teammates and teaming standards [64, 96]. Once these systems are developed and integrated through the collaboration of researchers, developers, and potential HATs, AI teammates that have the ability to contribute to teaming ethics can begin to be integrated into the workforce.

Once practitioners who follow this paper's model finish building AI teammates that posses the ability to demonstrate and contribute to ethical ideologies, immediate attention should turn to developing the ability for AI teammates to modify their common ethical ideology based on their team interactions. The development of this functionality needs to happen after the previous development goal as the changes AI systems make to their ideologies should never contradict the basic ethical foundation they were integrated with. In regard to Kantian ethics, the modifications to AI ethical ideology could be seen as additional, team specific duties that they must abide by. While this ability could be harmful if done maliciously, it is necessary in ensuring AI systems can adapt to the ethical climates they are placed in, which may not always be consistent due to the specific context of the HAT [22]. Once this functionality is built, preexisting HATs can begin to fully develop dynamic, team ethical ideologies, thus leading to a greater sense of shared understanding and cohesion between team members.

The development of these two functionalities is not the only aspects required for ethical AI teammates, as they also need to be trained in their assigned task; however, these functionalities are the vital aspects of AI teammates in ethical HATs. While these functionalities could be ignored and left out, doing so may not only result in unethical AI teammates but also a reduction in shared understanding within HATs. The development of these AI features should be considered necessary in the design and implementation of AI teammates if AI systems are expected to become fully functional teammates and not just tools humans use.

4.2 External Regulation for Ethical Human-AI Teams

Despite the model in Figure 1 providing a roadmap towards building ethical HATs, that does not mean that practitioners and developers of AI systems of required to build in these ethical considerations. This very reason is why this paper explicitly labels the external requirements for ensuring ethical human-AI teaming. Unfortunately, the explicit representation of these requirements is not enough, and it is necessary for external regulation to ensure HATs, in addition to all AI systems, operate ethically. While the centralization of AI regulation is still up for debate [24, 93, 113], the model presented in this paper necessitates external assurance and policy be applied to the context of human-AI teamwork, similar to other contexts within the AI domain [58, 65].

First, regulation and accreditation within AI educational domains needs to include a requirement for ethical human-AI teamwork. While ethics is seen as an essential part to a contemporary AI education [40], the complexities of teamwork outlined above necessitate its explicit consideration. In addition to serving as a guide for building ethical HATs, the model could also provide an essential educational resource in teaching students and professionals to understand the human-AI teaming relationship that they may be responsible for implementing and monitoring. Thus, future developers and designers of human-AI teams will be familiar with and ready to develop the concepts and functions outlined in the previous section. Without this education, professionals may be severely hindered from implementing HATs in a timely and ethical fashion.

Moreover, future policies and regulations need to consider the oversight requirements explicitly outlined by the model presented. While a large amount of ethical AI regulation is focused on the external interactions of AI in society, which is important [15, 63], the extra dimension of intra-team communication in HATs merits the requirements of additional oversights targeted towards the ethical use of AI teammates. Unfortunately, this regulation may not be as apparent as these systems may not be public facing, but rather internal and team facing. Ensuring the ethical use of these systems, whether they are internal tools or external entities, is a necessity to ensuring AI technologies, as a whole, are used ethically. Therefore, the concerns of internal AI usage within an organization should become an additional point of concern within AI regulation and policy, with an explicit consideration for the context of human-AI teamwork.

While the model presented in this paper is a good initial first step and guide towards building ethical human-AI teams, action is still necessary in assuring (1) ethical teammates are developed and (2) HATs operate ethically, both internally and externally. While the consideration of human-AI teaming ethics adds an extra layer of complexity as well as an extra regulatory burden, one of AI's greatest potential roles in society is dependent on ensuring it can become an ethical teammate. Being an ethical teammate is more than simply acting ethically, but being able to contribute to an overall ethical climate and prevent the decay or compromise of one's ethical ideology. Moving forward, developers, researchers, teams, and regulators are going to need to collaboratively work together to not only ensure AI agents are ethical, but the systems and teams they are a part of continuously maintain and promote a positive ethical ideology.

ACKNOWLEDGMENTS

This research was partially supported by AFOSR Award FA9550-20-1-0342 (Program Manager: Laura Steckman).

5 CONCLUSION

Building ethical human-AI teams is an increasingly complex task as AI technologies, and therefore the teams they will be a part of, continue to advance. While substantial innovations have been made in the field of AI ethics over the past few years, the increased complexity seen in HATs justifies a separate, more focused initiative towards building and ensuring ethical HATs. This paper provides a foundation for this initiative by consolidating decades of ethical teamwork research with the rapidly advancing fields of AI ethics and human-AI teaming to create a comprehensive model depicting ethical human-AI teamwork. The model, shown in Figure 1, serves as a guide for building ethical human-AI teaming. Without the consideration of the factors presented by this model, one of the most promising implementations of AI technology may not only underperform, but ultimately harm the society they were designed to help.

REFERENCES

[1] George Agich. 2004. Joining the Team: Ethics Consultation at the Cleveland Clinic. *HEC forum : an interdisciplinary journal on hospitals' ethical and legal issues* 15 (Jan. 2004), 310–22. https://doi.org/10.1023/B:HECF.0000011973.18422.16

[2] Anne Arber and Ann Gallagher. 2009. Generosity and the Moral Imagination in the Practice of Teamwork. *Nursing Ethics* 16, 6 (Nov. 2009), 775–785. https://doi.org/10.1177/0969733009343134 Publisher: SAGE Publications Ltd.

[3] Tom L. Beauchamp, Professor of Philosophy and Senior Research Scholar Tom L. Beauchamp, James F. Childress, and University Professor and Hollingsworth Professor of Ethics James F. Childress. 2001. *Principles of Biomedical Ethics*. Oxford University Press. Google-Books-ID: _14H7MOw1o4C.

[4] Martin Benjamin. 1992. *Ethics in Nursing*. Oxford University Press.

[5] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[6] Piercosma Bisconti Lucidi and Daniele Nardi. 2018. Companion Robots: the Hallucinatory Danger of Human-Robot Interactions. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 17–22. https://doi.org/10.1145/3278721.3278741

[7] Edvard P. Bjørgen, Simen Madsen, Therese S. Bjørknes, Fredrik V. Heimsæter, Robin Håvik, Morten Linderud, Per-Niklas Longberg, Louise A. Dennis, and Marija Slavkovik. 2018. Cake, Death, and Trolleys: Dilemmas as benchmarks of ethical decision-making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 23–29. https://doi.org/10.1145/3278721.3278767

[8] Alan Bleakley. 2006. A Common Body of Care: The Ethics and Politics of Teamwork in the Operating Theater are Inseparable. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 31, 3 (Jan. 2006), 305–322. https://doi.org/10.1080/03605310600732826

[9] Defense Innovation Board. 2019. AI principles: Recommendations on the ethical use of Artificial Intelligence by the Department of Defense. *Supporting document, Defense Innovation Board* (2019).

[10] Jason Borenstein and Ayanna Howard. 2020. Emerging challenges in AI and the need for AI ethics education. *AI and Ethics* (2020), 1–5.

[11] David S. Bright, Bradley A. Winn, and Jason Kanov. 2014. Reconsidering Virtue: Differences of Perspective in Virtue Ethics and the Positive Social Sciences. *Journal of Business Ethics* 119, 4 (Feb. 2014), 445–460. https://doi.org/10.1007/s10551-013-1832-x

[12] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228* (2018).

[13] Erik Brynjolfsson, John J Horton, Adam Ozimek, Daniel Rock, Garima Sharma, and Hong-Yi TuYe. 2020. *COVID-19 and remote work: an early look at US data*. Technical Report. National Bureau of Economic Research.

[14] JA Cannon-Bowers, E Salas, and SA Converse. 1990. Cognitive psychology and team training: Shared mental models incomplex systems. In *5th Annual Conference ofthe Society for Industrial and Organizational Psychology, Miami, florida*.

[15] Stephen Cave, Kate Coughlan, and Kanta Dihal. 2019. "Scary Robots": Examining Public Responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 331–337. https://doi.org/10.1145/3306618.3314232

[16] Stephen Cave and Seán S. ÓhÉigeartaigh. 2018. An AI Race for Strategic Advantage: Rhetoric and Risks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 36–40. https://doi.org/10.1145/3278721.3278780

[17] Jessie YC Chen and Michael J. Barnes. 2010. Supervisory control of robots using RoboLeader. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 1483–1487. Issue: 19.

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

[19] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.

[20] Shih-Yi Chien, Michael Lewis, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 2016. Influence of cultural factors in dynamic trust in automation. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 002884–002889.

[21] Amit K Chopra and Munindar P Singh. 2018. Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 48–53.

[22] Amit K. Chopra and Munindar P. SIngh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 48–53. https://doi.org/10.1145/3278721.3278740

[23] D. A. Christakis and C. Feudtner. 1993. Ethics in a short white coat: the ethical dilemmas that medical students confront. *Academic Medicine* 68, 4 (April 1993), 249–54. https://journals.lww.com/academicmedicine/Abstract/1993/04000/Ethics_in_a_short_white_coat_the_ethical_dilemmas.3.aspx

[24] Peter Cihon, Matthijs M. Maas, and Luke Kemp. 2020. Should Artificial Intelligence Governance be Centralised? Design Lessons from History. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 228–234. https://doi.org/10.1145/3375627.3375857

[25] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. 2016. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems.. In *AAMAS*. 1106–1114.

[26] Sharolyn Converse, JA Cannon-Bowers, and E Salas. 1993. Shared mental models in expert team decision making. *Individual and group decision making: Current issues* 221 (1993), 221–46.

[27] Subhro Das, Sebastian Steffen, Wyatt Clarke, Prabhat Reddy, Erik Brynjolfsson, and Martin Fleming. 2020. Learning Occupational Task-Shares Dynamics for the Future of Work. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 36–42. https://doi.org/10.1145/3375627.3375826

[28] Clare Delany and Melati Conwell. 2012. Ethics and teamwork for pediatric medical imaging procedures: insights from educational play therapy. *Pediatric Radiology* 42, 2 (Feb. 2012), 139–146. https://doi.org/10.1007/s00247-011-2271-4

[29] Mustafa Demir, Nathan J. McNeese, Nancy J. Cooke, Jerry T. Ball, Christopher Myers, and Mary Frieman. 2015. Synthetic Teammate Communication and Coordination With Humans. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59, 1 (Sept. 2015), 951–955. https://doi.org/10.1177/1541931215591275 Publisher: SAGE Publications Inc.

[30] Darleen M. DeRosa, Donald A. Hantula, Ned Kock, and John D'Arcy. 2004. Trust and leadership in virtual teamwork: A media naturalness perspective. *Human Resource Management* 43, 2-3 (2004), 219–232. https://doi.org/10.1002/hrm.20016 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hrm.20016.

[31] Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, Galit Haim, Malte S. Kließ, Maite Lopez-Sanchez, Roberto Micalizio, Juan Pavón, Marija Slavkovik, Matthijs Smakman, Marlies van Steenbergen, Stefano Tedeschi, Leon van der Toree, Serena Villata, and Tristan de Wildt. 2018. Ethics by Design: Necessity or Curse?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 60–66. https://doi.org/10.1145/3278721.3278745

[32] Jean L Dyer. 1984. Team research and team training: A state-of-the-art review. *Human factors review* 26 (1984), 285–323.

[33] Clarence A Ellis, Simon J Gibbs, and Gail Rein. 1991. Groupware: some issues and experiences. *Commun. ACM* 34, 1 (1991), 39–58.

[34] Amitai Etzioni and Oren Etzioni. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 21, 4 (2017), 403–418.

[35] Stephen M Fiore, Eduardo Salas, Haydee M Cuevas, and Clint A Bowers. 2003. Distributed coordination space: toward a theory of distributed team process and performance. *Theoretical Issues in Ergonomics Science* 4, 3-4 (2003), 340–364.

[36] Christopher Flathmann, Beau Schelble, Brock Tubre, Nathan McNeese, and Paige Rodeghero. 2020. Invoking Principles of Groupware to Develop and

Evaluate Present and Future Human-Agent Teams. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 15–24.
[37] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines* 28, 4 (2018), 689–707.
[38] Sara T Fry, Megan-Jane Johnstone, and Marla Fletcher. 2003. Ethics in nursing practice: a guide to ethical decision making. *The Canadian Nurse* 99, 4 (2003), 20. Publisher: Canadian Nurses Association.
[39] Heidi Furey and Fred Martin. 2019. AI education matters: a modular approach to AI ethics education. *AI Matters* 4, 4 (2019), 13–15.
[40] Natalie Garrett, Nathan Beard, and Casey Fiesler. 2020. More Than" If Time Allows" The Role of Ethics in AI Education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 272–278.
[41] Timothy Geary and David Danks. 2019. Balancing the Benefits of Autonomous Vehicles. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 181–186. https://doi.org/10.1145/3306618.3314237
[42] Glenn C. Graber and Christopher D. Pionke. 2006. A team-taught interdisciplinary approach to engineering ethics. *Science and Engineering Ethics* 12, 2 (June 2006), 313–320. https://doi.org/10.1007/s11948-006-0029-4
[43] Barbara J Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. 2019. Embedded EthiCS: integrating ethics across CS education. *Commun. ACM* 62, 8 (2019), 54–61. Publisher: ACM New York, NY, USA.
[44] L Beth Gunn and others. 2005. Working well with others: The evolution of teamwork and ethics. *Public Choice* 123, 1-2 (2005), 115–131. Publisher: Springer.
[45] Feyza Merve Hafizoglu and Sandip Sen. 2018. The Effects of Past Experience on Trust in Repeated Human-Agent Teamwork. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 514–522.
[46] Feyza Merve Hafizoglu and Sandip Sen. 2018. Reputation based trust in human-agent teamwork without explicit coordination. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. 238–245.
[47] Mark A Hall. 2005. The importance of trust for ethics, law, and public policy. *Cambridge Q. Healthcare Ethics* 14 (2005), 156.
[48] Nader Hanna and Deborah Richards. 2015. The Impact of Virtual Agent Personality on a Shared Mental Model with Humans during Collaboration. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 1777–1778.
[49] José Hernández-Orallo and Karina Vold. 2019. AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 507–513. https://doi.org/10.1145/3306618.3314238
[50] Sujin K. Horwitz and Irwin B. Horwitz. 2007. The Effects of Team Diversity on Team Outcomes: A Meta-Analytic Review of Team Demography. *Journal of Management* 33, 6 (Dec. 2007), 987–1015. https://doi.org/10.1177/0149206307308587 Publisher: SAGE Publications Inc.
[51] Tammy G. Hunt and Daniel F. Jennings. 1997. Ethics and Performance: A Simulation Analysis of Team Decision Making. *Journal of Business Ethics* 16, 2 (Feb. 1997), 195–203. https://doi.org/10.1023/A:1017987224590
[52] Susan E. Jackson and Aparna Joshi. 2011. Work team diversity. In *APA handbook of industrial and organizational psychology, Vol 1: Building and developing the organization*. American Psychological Association, Washington, DC, US, 651–686. https://doi.org/10.1037/12169-020
[53] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.
[54] Gareth R Jones and Jennifer M George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review* 23, 3 (1998), 531–546. Publisher: Academy of Management Briarcliff Manor, NY 10510.
[55] Aileen R Killen. 2002. Stories from the Operating Room: moral dilemmas for nurses. *Nursing Ethics* 9, 4 (July 2002), 405–415. https://doi.org/10.1191/0969733002ne524oa Publisher: SAGE Publications Ltd.
[56] Dongkyu Kim and Christian Vandenberghe. 2020. Ethical Leadership and Team Ethical Voice and Citizenship Behavior in the Military: The Roles of Team Moral Efficacy and Ethical Climate. *Group & Organization Management* 45, 4 (Aug. 2020), 514–555. https://doi.org/10.1177/1059601120920050 Publisher: SAGE Publications Inc.
[57] A. Kossaify, W. Hleihel, and J. C. Lahoud. 2017. Team-based efforts to improve quality of care, the fundamental role of ethics, and the responsibility of health managers: monitoring and management strategies to enhance teamwork. *Public Health* 153 (Dec. 2017), 91–98. https://doi.org/10.1016/j.puhe.2017.08.007
[58] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. Defining AI in Policy versus Practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/3375627.3375835
[59] Marjan Laal, Ashkan Laal, and Arsalan Aliramaei. 2014. Continuing education; lifelong learning. *Procedia-social and behavioral sciences* 116 (2014), 4052–4056.
[60] Gregory Luke Larkin. 2010. Unwitting Partners in Death: The Ethics of Teamwork in Disaster Management. *AMA Journal of Ethics* 12, 6 (June 2010), 495–501. https://doi.org/10.1001/virtualmentor.2010.12.6.oped1-1006. Publisher: American Medical Association.
[61] Wilfred W. F. Lau and Allan H. K. Yuen. 2014. Internet ethics of adolescents: Understanding demographic differences. *Computers & Education* 72 (March 2014), 378–385. https://doi.org/10.1016/j.compedu.2013.12.006
[62] Jessica Lipnack and Jeffrey Stamps. 1999. Virtual teams: The new way to work. *Strategy & Leadership* (1999).
[63] Alex John London and David Danks. 2018. Regulating Autonomous Vehicles: A Policy Proposal. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 216–221. https://doi.org/10.1145/3278721.3278763
[64] Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K. Brent Venable. 2018. Preferences and Ethical Principles in Decision Making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 222. https://doi.org/10.1145/3278721.3278723
[65] Matthijs M. Maas. 2018. Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 223–228. https://doi.org/10.1145/3278721.3278766
[66] Deborah J MacInnis. 2011. A framework for conceptual contributions in marketing. *Journal of Marketing* 75, 4 (2011), 136–154.
[67] A. Mahalle, J. Yong, and X. Tao. 2019. Ethics of IT Security Team for Cloud Architecture Infrastructure in Banking and Financial Services Industry. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 506–511. https://doi.org/10.1109/CSCWD.2019.8791928
[68] Sarah A. Manspeaker, Elena V. Donoso Brown, Sarah E. Wallace, Leesa DiBartola, and Allison Morgan. 2017. Examining the perceived impact of an ethics workshop on interprofessional values and teamwork. *Journal of Interprofessional Care* 31, 5 (Sept. 2017), 628–637. https://doi.org/10.1080/13561820.2017.1336992 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13561820.2017.1336992.
[69] C. Dianne Martin. 2007. Leadership, teamwork, and ethics in the development of IT professionals. *ACM SIGCSE Bulletin* 39, 2 (June 2007), 8–9. https://doi.org/10.1145/1272848.1272850
[70] Jeanna Neefe Matthews, Graham Northup, Isabella Grasso, Stephen Lorenz, Marzieh Babaeianjelodar, Hunter Bashaw, Sumona Mondal, Abigail Matthews, Mariama Njie, and Jessica Goldthwaite. 2020. When Trusted Black Boxes Don't Agree: Incentivizing Iterative Improvement and Accountability in Critical Software Systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 102–108. https://doi.org/10.1145/3375627.3375807
[71] Stuart McLennan, Amelia Fiske, Leo Anthony Celi, Ruth Müller, Jan Harder, Konstantin Ritt, Sami Haddadin, and Alena Buyx. 2020. An embedded ethics approach for AI development. *Nature Machine Intelligence* 2, 9 (2020), 488–490.
[72] Nathan McNeese, Mustafa Demir, Erin Chiou, Nancy Cooke, and Giovanni Yanikian. 2019. Understanding the role of trust in human-autonomy teaming. In *Proceedings of the 52nd Hawaii international conference on system sciences*.
[73] Nathan J McNeese, Mustafa Demir, Nancy J Cooke, and Christopher Myers. 2018. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors* 60, 2 (2018), 262–273.
[74] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors* 58, 3 (2016), 401–415.
[75] Jacob Metcalf, Emanuel Moss, and danah boyd. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476. http://muse.jhu.edu/article/732185 Publisher: Johns Hopkins University Press.
[76] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* (2019), 1–7.
[77] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics* 26, 4 (2020), 2141–2168.
[78] Carlos R. Degrandi Oliveira. 2019. Ethics in the operating room; the anesthesiologist's responsibility. *Anaesthesia, Pain & Intensive Care* (2019), 333–336. https://doi.org/10.35975/apic.v23i4.1163
[79] Osonde A Osoba, Benjamin Boudreaux, and Douglas Yeung. 2020. Steps Towards Value-Aligned Systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 332–336.
[80] Levent Ozgonul and Mustafa Kemal Alimoglu. 2019. Comparison of lecture and team-based learning in medical ethics education. *Nursing Ethics* 26, 3 (May 2019), 903–913. https://doi.org/10.1177/0969733017731916 Publisher: SAGE Publications Ltd.

[81] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2020. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors* (2020), 0018720820960865.

[82] Michael E. Palanski, Surinder S. Kahai, and Francis J. Yammarino. 2011. Team Virtues and Performance: An Examination of Transparency, Behavioral Integrity, and Trust. *Journal of Business Ethics* 99, 2 (March 2011), 201–216. https://doi.org/10.1007/s10551-010-0650-7

[83] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.

[84] Alexander M. Petersen, Ioannis Pavlidis, and Ioanna Semendeferi. 2014. A Quantitative Perspective on Ethics in Large Team Science. *Science and Engineering Ethics* 20, 4 (Dec. 2014), 923–945. https://doi.org/10.1007/s11948-014-9562-8

[85] Federico Pistono and Roman V Yampolskiy. 2016. Unethical research: how to create a malevolent artificial intelligence.

[86] Alfred Presbitero and Mendiola Teng-Calleja. 2019. Ethical leadership, team leader's cultural intelligence and ethical behavior of team members: Implications for managing human resources in global teams. *Personnel Review* 48, 5 (Jan. 2019), 1381–1392. https://doi.org/10.1108/PR-01-2018-0016 Publisher: Emerald Publishing Limited.

[87] David V Pynadath, Ning Wang, Ericka Rovira, and Michael J Barnes. 2018. Clustering behavior to recognize subjective beliefs in human-agent teams. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1495–1503.

[88] Hamid Rahimi and Farshad Baharlooei. 2018. The Effect of Ethical Climate on Trust in Teamwork with the Meditating Role of Ethical Behavior. *Organizational Behaviour Studies Quarterly* 7, 2 (Aug. 2018), 129–158. http://obs.sinaweb.net/article_32518.html

[89] Laura W Roberts, Teddy D Warner, Katherine A Green Hammond, Cynthia MA Geppert, and Thomas Heinrich. 2005. Becoming a good doctor: perceived need for ethics training focused on practical and professional development topics. *Academic Psychiatry* 29, 3 (2005), 301–309.

[90] Paul Robinson. 2007. Ethics training and development in the military. *Parameters* 37, 1 (2007), 23. Publisher: US Army War College.

[91] Eduardo Salas, Nancy J Cooke, and Michael A Rosen. 2008. On teams, teamwork, and team performance: Discoveries and developments. *Human factors* 50, 3 (2008), 540–547.

[92] Beau G Schelble, Christopher Flathmann, and Nathan McNeese. 2020. Towards Meaningfully Integrating Human-Autonomy Teaming in Applied Settings. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 149–156.

[93] Daniel Schiff, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 153–158. https://doi.org/10.1145/3375627.3375804

[94] Arathi Sethumadhavan. 2019. Trust in Artificial Intelligence. *Ergonomics in Design* 27, 2 (April 2019), 34–34. https://doi.org/10.1177/1064804618818592 Publisher: SAGE Publications Inc.

[95] Graham Sewell. 2005. Doing what comes naturally? Why we need a practical ethics of teamwork. *The International Journal of Human Resource Management* 16, 2 (Feb. 2005), 202–218. https://doi.org/10.1080/0958519042000311408 Publisher: Routledge _eprint: https://doi.org/10.1080/0958519042000311408.

[96] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 166–172. https://doi.org/10.1145/3375627.3375812

[97] Vahid Soltanzadeh, Afshar Kabiri, Afshar Kabiri, and Hassan Galavandi. 2014. A Study of the Relationship between Social Responsibility and Teamwork among the Staff of Urmia University. *Journal of Applied Sociology* 25, 1 (March 2014), 111–120. https://jas.ui.ac.ir/article_18336.html Publisher: University of Isfahan.

[98] Alan P. Sprague and Raquel Diaz-Sprague. 2019. Tying Ethics to Teamwork Training in a Minimodule. In *Proceedings of the 8th Computer Science Education Research Conference (CSERC '19)*. Association for Computing Machinery, New York, NY, USA, 120–122. https://doi.org/10.1145/3375258.3375275

[99] B. C. Stahl and D. Wright. 2018. Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security Privacy* 16, 3 (May 2018), 26–33. https://doi.org/10.1109/MSP.2018.2701164 Conference Name: IEEE Security Privacy.

[100] Andreas Theodorou and Virginia Dignum. 2020. Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence* 2, 1 (2020), 10–12.

[101] Jim Torresen. 2018. A review of future and ethical perspectives of robotics and AI. *Frontiers in Robotics and AI* 4 (2018), 75.

[102] Ville Vakkuri and Kai-Kristian Kemell. 2019. Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEDD Strategy. In *International Conference on Software Business*. Springer, 260–275.

[103] Dieter Vanderelst and Alan Winfield. 2018. The Dark Side of Ethical Robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 317–322. https://doi.org/10.1145/3278721.3278726

[104] Rebecca A. VanMeter, Douglas B. Grisaffe, Lawrence B. Chonko, and James A. Roberts. 2013. Generation Y's Ethical Ideology and Its Potential Workplace Implications. *Journal of Business Ethics* 117, 1 (Sept. 2013), 93–109. https://doi.org/10.1007/s10551-012-1505-1

[105] David A Whetten. 1989. What constitutes a theoretical contribution? *Academy of management review* 14, 4 (1989), 490–495.

[106] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 195–200.

[107] Alan Winfield. 2019. Ethical standards in robotics and AI. *Nature Electronics* 2, 2 (2019), 46–48.

[108] Alan FT Winfield and Marina Jirotka. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180085.

[109] Ava Thomas Wright. 2019. Rightful Machines and Dilemmas. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 3–4. https://doi.org/10.1145/3306618.3314261

[110] Julia L Wright, Jessie YC Chen, and Michael J Barnes. 2018. Human–automation interaction for multiple robot control: the effect of varying automation assistance and individual differences on operator performance. *Ergonomics* 61, 8 (2018), 1033–1045.

[111] Julia L Wright, Jessie Y Chen, Stephanie A Quinn, and Michael J Barnes. 2013. *The effects of level of autonomy on human-agent teaming for multi-robot control and local security maintenance.* Technical Report. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD.

[112] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953* (2018).

[113] Baobao Zhang and Allan Dafoe. 2020. U.S. Public Opinion on the Governance of Artificial Intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 187–193. https://doi.org/10.1145/3375627.3375827

[114] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 246:1–246:25. https://doi.org/10.1145/3432945

[115] Jinguo Zhao, Wei Sun, Shujie Zhang, and Xiaohong Zhu. 2020. How CEO Ethical Leadership Influences Top Management Team Creativity: Evidence From China. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.00748 Publisher: Frontiers.