Ethics in Human-AI Teaming: Principles and Perspectives

Abstract
Ethical considerations are the fabric of our society, and they foster cooperation, helping, and sacrifice for the greater good. Advances in AI create a greater need to examine ethical considerations involving the development and implementation of such systems. Integrating ethics into artificial intelligence-based programs is crucial for preventing negative outcomes such as privacy breaches and biased decision making. Human-AI Teaming (HAIT) presents additional challenges as the ethical principles and moral theories that provide justification for them are not yet computable by machines; a human must remain "in the loop" to arbitrate ethical considerations. To that effect, models of human judgment and decision making, such as the Agent-Deed-Consequence (ADC) model, will be crucial to inform the ethical guidance functions in AI teammates and to clarify how and why humans (dis)trust machines. The current paper will examine the ADC model as it is applied to the context of HAIT, and the challenges associated with the use of human-centric ethical models when applied to an AI context. An AI algorithm empowered by an ADC model decision tree is both flexible to a variety of ethical frameworks and adaptive to environmentally and contextually different situations, and we therefore believe that this research provides a novel solution to the challenge of applying ethical principles to HAIT.

Keywords: Human-AI Teaming (HAIT), Agent-Deed-Consequence (ADC) model, Carebots, Ethics, Trust, Artificial Intelligence (AI)

1. Why is ethics in AI important in the context of Human-AI teaming?

As technology advances, machines can increasingly supplant human processes, procedures, and operations in real-life settings such as elderly care (e.g., carebots), healthcare, transportation (e.g., autonomous vehicles), human resources, and military applications (e.g., drones). Intelligent machines can be used to extend human performance through Human-AI-Teaming (HAIT), and methods of enhancing teamwork between humans and artificial intelligence (AI) systems are being thoroughly researched. Yet with added capabilities often come added responsibilities. AI systems are increasingly placed in difficult situations wherein they must navigate the complexities of safety, human life, human preferences and biases, and dynamic situations. At times, AI systems may be placed in morally contentious situations. Yet despite contemporary publicity and attention regarding things like the fairness of AI, there has been little systematic research on the juxtaposition of human ethical models and AI-capable HAITs. While AI might be regarded as inherently amoral, as it does not share the same fundamental principles humans adhere to, there may be human-like attributions and biases that influence moral reasoning of AI systems. Humanity must ensure that the implications of AI are understood as much as possible, and that AI is programmed (and implemented) in alignment with ethical principles.

Ethics, a set of reflected norms, rules, precepts, and principles that govern and guide the behavior of individuals or groups [1], [2], has become increasingly important in the context of AI applications. For example, the Department of Defense has adopted the AI ethical principles of responsibility, equitability, traceability, reliability, and governability [3]. Ethical issues such as safety, reliability, justice, and fairness are increasingly salient as AI technology has become more prevalent in society, especially in the human teaming field. There is ample agreement that inevitably, AI (a general autonomous system) will find itself in a situation where it needs to make complex ethical decisions over and above a simple choice on whether or not to obey a rule [4]. Despite this increasing salience, there is no obvious or uncontroversial way to implement a human understanding of certain ethical behaviors into computers and other software.

Attempts to implement ethical behavior into AI usually concern the question of what principles should govern and guide the design and use of AI based programs [5], [6]. We know that the continued use of AI will have social, psychological, financial, legal, environmental, and trust ramifications for years to come [7], [8], but to what end? This technology also has the potential to do considerable harm to society and to those who utilize or interact with it. AI can cause harm by concealing prejudiced models behind ostensibly objective decision-making, creating ambiguity about the liability of manufacturers and users of AI-based systems, and invading the privacy of those subjected to AI scrutiny [9].

To demonstrate a realistic, albeit drastic, implementation of HAIT, let us consider a hypothetical event similar to the 2021 Surfside building collapse in Florida. A building partially collapses, leaving people stranded under rubble and debris. The emergency responders must act quickly to get people out of the rubble to ensure that they survive. First responders deploy a small robot that utilizes AI to make decisions based on the in-the-field data. The robot is used to pull people from rubble and debris, with the AI making decisions on how safe it is to perform an action and the best ways to get the survivor out. When humans team with AI, we must ask what ethical decisions the machine will have to make. In this case, the AI robot must first decide where in the rubble to search. The AI would then decide on the ideal way to search the area to reach survivors and minimize the waste of precious time. AI might also be required to make decisions on whether to prioritize one individual over another. When the AI locates a survivor amongst the rubble, how much time and effort should it exhaust to save that person? And on what basis will it make these decisions? For example, if the AI robot discovers a survivor who could only be removed following amputation of the person's legs, how much agency would the survivor have over the decision of whether the AI amputates their legs, attempts to extract the survivor without amputation, or abandons the survivor in an attempt to preserve the greatest number of whole persons? Further, how would society feel about a robot capable of amputating legs of unconscious persons without consent? A robot could decide whether to remove a person faster by amputating limbs, or attempt to save the survivor's limbs, expending time and potentially preventing it from rescuing other survivors. But where do we set the fulcrum between human agency and robot autonomy? Appropriately weighting both will require resolving ethical dilemmas such as

the one presented in this case study and will determine to what extent teaming is possible. A common view of philosophers and laypeople alike is that moral decisions are ineluctably human. While there is ample agreement that human moral cognition is, for the time being, superior and more suitable than AI, there are plenty of instances where there is no time for meaningful human input in life-or-death situations.

Military applications research conducted by the Defense Advanced Research Projects Agency (DARPA) demonstrates another future application for HAIT. This initiative, the In the Moment Program (ITM), "will research and develop technology to support building, evaluating, and fielding algorithmic decision-makers that can assume human-off-the-loop decision-making responsibilities in difficult domains, such as medical triage in combat" [10]. While the initiative is still in the early stages, the idea – which can be extended beyond combat-only applications towards disaster relief and first response – demonstrates the potential application of HAIT for situations in which AI-derived algorithmic expertise is available in the field when oversight from a human expert is impossible or impractical.

To avoid glaring problems in such ambiguous situations, AI research must have ethics baked in. However, ethics is a complex realm of judgment and decision-making which encompasses decisions about Agents (e.g., is a particular person is more worth saving?), Deeds (e.g., is amputation justifiable?) and Consequences (e.g., could more people have been saved?) that must all be considered [11]. The establishment and observance of rules and regulations, especially in the context of HAIT, increases workplace efficiency and fosters trust in AI and robotics [6]. Without clear and comprehensive rules governing the AI research space, private information could be exposed without consent, and AI applications such as hiring tools could be increasingly biased, hurting minority or underrepresented populations [12]. Intelligent systems that are installed and used by organizations without principles or regulations result in worker harm as well as negative organizational outcomes and functioning, particularly among teams engaged in complex tasks [8]. In addition to privacy and safety rules, fairness and non-discrimination regulations should be implemented to protect those working around and with AI [5]. These regulations will promote effective and efficient human teaming alongside AI and facilitate ethical and professional action from both counterparts.

The Agent, Deed, and Consequence (ADC) model can be referenced to classify decisions and actions as either ethical or unethical. The model states that moral judgment consists of three components: the character of a person (Agent); their actions (Deed); and the consequences brought about by the situation (Consequence) [2]. The ADC model applies our three main moral theories to these components: virtue ethics, deontology, and consequentialism. Using these moral theories, the ADC model concludes that moral judgments are positive if all three of its components are positive and negative if some or all three of its components are negative. As

we will reveal in Sections 2 and 3, the ADC model is a powerful tool that can facilitate the development of ethical algorithms in which the weighted values of Agents, Deeds, and Consequences are moderated by a human moral agent, thus accommodating a variety of ethical frameworks.

The primary goal of this paper is to investigate the extent to which the ADC model can address challenges faced by HAIT technologies that seek to implement ethical behavior. We believe that an ADC model-empowered human-AI team will accommodate a variety of ethical frameworks, thus avoiding development-stifling philosophical debate over which single framework is most appropriate. This paper is organized as follows: In Section 2, we survey: (i) the relevant established ethical theories, (ii) the principles for ethics of AI, and (iii) the Agent-Deed-Consequence (ADC) model for a potential ethical guidance function which could be implemented in AI to enable HAIT. In Section 3, we discuss the challenges of adopting the ADC model for AI systems, since attributions of machines are quite different than those of humans and, more importantly, the factors that shape those attributions are not necessarily the same when considering an AI system as the referent. In Section 4, we canvass the three kinds of HAIT technologies that will be implemented in the near future: i) virtual AI assistive technologies [13], ii) animal-like carebot companions [14], and iii) complex humanoid carebots [15]. We review the currently available evidence that these types of HAITs are beneficial for society and explore challenges in terms of acceptability and the need for regulation. In Section 5, we summarize the findings of our review and identify gaps in evidence and avenues for future research. We hope this work can also inspire the opening of other related lines of research towards founding ethics in various HAIT contexts. We further anticipate that this work will inspire novel applications of HAIT in which an AI algorithm is "supervised" by a human moral agent. Such teaming applications will bridge the gap between human and autonomous moral agency, fostering public trust in AI until such time as the development of AI technology has progressed to the point that it can stand alone in ethically charged situations.

2. Guidelines and Principles for Ethics in AI

2.1 Three major ethics theories

Three well-known theories used to evaluate controversial ethical situations are virtue ethics, deontology, and consequentialism [16]. Virtue ethics emphasizes the agency or character of an agent, arguing that agents will respond differently to identical situations due to their differences in character [17], [18]. Deontology is concerned with the actions of an agent, claiming that certain actions are either right or wrong based on the intention behind an action [11]. Consequentialism focuses on the results of an action, reasoning that an agent is moral only if it chooses the most ethical outcome [19].

Though each is a valid way to evaluate ethical situations, these three theories are often used differently when it comes to AI. Specifically, there are 9 main principles that constitute ethics in AI: (1) fairness and non-discrimination, (2) privacy, (3) safety and security, (4) human control of technology, (5) transparency and explainability, (6) accountability, (7) promotion of human values, (8) professional responsibility, and (9) sustainable development [6]. These principles can be divided into three distinct categories: (I) avoiding undesired results, (II) liability/acting responsibly, and (III) ameliorating the lack of ethics in AI, in order for us to better understand where they can be applied in HAIT. These categories also help contextualize the larger implications and impacts each principle could have on humanity and the future of HAIT.

2.1.1 Category I- Avoiding Undesired Results

The largest category is concerned with avoiding the dangers of AI when used for unethical or immoral purposes, whether intentionally or unintentionally [6]. Table 1 shows connections between this category, three major theories of ethics, and HAIT. There are four principles in this category, the first of which is *fairness and non-discrimination*. Fairness and non-discrimination suggests that AI should utilize only representative and high-quality data, be used impartially and equally across demographics, and consider a diverse array of stakeholders in its design and implementation [8]. The next principle is *privacy*, and it covers the right to consent to AI-based data collection and analysis as well as participant control over the subsequent use of the data. The third principle is *safety and security*, which proposes that AI should be able to protect its data from internal and external threats while maintaining an element of predictability in its behavior that protects society and people's safety. The last principle is *the need for human control of technology*; AI must remain under human control and must also be reviewed by those impacted by the technology. Ultimately, the actions of AI are within human governance, meaning that the results and decisions that stem from AI technology should be challenged, reviewed, nullified, or managed by humans, but also that human-agents can be causally linked with the consequences of AI actions.

2.1.2 Category II - Liability or Acting Responsibly

The principles of liability or acting responsibly state that AI must be designed and utilized under appropriate scrutiny and within legal boundaries [5]. Table 2 shows connections between this category, three major theories of ethics, and HAIT. Of the two principles in this category, the first we will define is *transparency and explainability*. This principle requires that the AI-based systems are designed to enable oversight, explainability, and understandability. The second principle of *accountability* refers to determining the agent or agents accountable for a decision made by an AI-based system. This can be further divided into three stages of accountability: before, during, and after the use of the AI. In ethics of AI in organizations, regulatory systems should exist to rectify unjust decisions made by the AI-based system post deployment and to hold legally liable those responsible for causing harm using AI technologies [5].

2.1.3 Category III - Ameliorating the Lack of Ethical Values in AI

The final category is ameliorating the lack of ethical values in AI. This category is premised upon the understanding that because AI is inherently amoral, we therefore need rules and laws to guide its ethical implementation [20]. Table 3 again shows connections between this category, three major theories of ethics, and HAIT. The first principle, *promotion of human values*, suggests that AI-based systems should be used for the common good and should be developed consistently with cross-cultural human values [8]. AI systems should be widely available and distributed as equally as possible, benefitting all of humankind. The next principle is *professional responsibility*; it proposes that AI be designed meticulously, purposefully, and with input from a variety of stakeholders. The creators of AI-based systems should consider the long-term outcomes of its intended and unintended usage and should create AI-based systems in a manner that is reliable, valid, and otherwise guided by scientific principles. The final principle is sustainable development, referring to creating AI technologies that enable maintainable solutions to global problems such as healthcare and equality, minimizing resource waste, and environmental responsibility. By illustrating the relationships between the nine principles, the three ethical theories, and HAIT, one can view how each principle is related to each ethical theory and how each theory may impact HAIT. Furthermore, the connections between the principles and the theories allow for a clearer understanding of how the ADC model may be used to implement ethics into HAIT.

Table 1. Connections between avoiding undesired results category, three major theories of ethics, and HAIT

Category	Principle	Definition	Ethics Theory		Relation to HAIT
Avoiding Undesired Results	Fairness and Non-Discrimination	AI algorithms that are non-discriminatory, fair, inclusive, representative, and free from human biases.	Deontology	Create AI systems that are unbiased and fair in their selections and ideas.	The humans using the systems must differentiate between the use of the data collected and the consent associated with it. Human users must also consider the fairness of outcomes.
			Virtue Ethics	Algorithms should not lead to disrespect of persons.	
			Consequentialism	Biases based on previous representations of the world may cause adverse effects.	
	Privacy	AI use that enables consent, protection from surveillance, and right to control the use of the data gathered.	Deontology	Protect an individual's data from surveillance.	The AI-system must protect privacy and sensitive information. When working in teams, even individually, there needs to be a safe work environment where everyone feels a sense of security. The human-AI team needs to evaluate AI usage outcomes and ensure that the AI does no harm to humans and can resist any external threats on the business in general or to individual(s).
			Virtue Ethics	Keep users' data safe when consent is enabled.	
			Consequentialism	Prevent misuse of data.	

	Safety and Security	AI that does no harm to humans and resists external threats.	Deontology	AI should not present a threat to human safety and security.	The AI must be designed to maximize human safety and security. The human-AI team must be appropriately trained, and the human user(s) must be competently trained to evaluate the contextual performance of the AI and to anticipate fault lines wherein the AI may err and/or cause negative outcomes.
			Virtue Ethics	Keep persons and their data safe and secure from malicious intent.	
			Consequentialism	System vulnerabilities lead to negative outcomes.	
	Human Control of Technology	AI that remains under human control and allows for review by those impacted	Deontology	AI should not lead to dehumanization.	The AI must be designed to maximize transparency to ensure human shared awareness of the AI goals, behaviors, and assumptions. The AI should be designed to always foster maximum human control. The AI should be fully implemented only after the human-AI team has prior experience with AI in that context.
			Virtue Ethics	Humans remain in control of AI-based systems and AI in general.	
			Consequentialism	Lack of human control leads to adverse effects in general.	

Table 2. Connections between liability or acting responsibly category, three major theories of ethics, and HAIT

Category	Principle	Definition	Ethics Theory		Relation to HAIT
Liability/Acting Responsibly	Transparency and Explainability	AI that enables oversight and can be parsimoniously explained, understood, and recognized.	Deontology	The operations of AI should be in principle understandable by those who have little knowledge of the subject.	AI must be designed such that its goals, behaviors, and assumptions are always transparent to human observers. The AI must also be designed to always foster maximum human control. For each usage context, user proficiency with AI technology must precede AI implementation and integration. The AI will promote dialogue between the human and the AI such that the human can question the rationale or prior behaviors of the AI and modify future decision in accordance with the human-AI team goals. The AI should be designed to communicate why an action has occurred.
			Virtue Ethics	AI must articulate decision-making rationale to those affected by its decisions.	
			Consequentialism	If AI decisions are too complicated, negative effects are harder to ameliorate.	
	Accountability	AI that is subject to continuous assessment, evaluation, and creation of usage regulations, and that is subsequently liable for failure to meet these regulations.	Deontology	Clear rules should guide AI use by humans.	When working in a team, all individuals must be accountable for their actions and ideas. This facilitates constant assessment, evaluation, and refinement by the rest of the team. The accountability of AI use is the same. Team strength and unity are fortified when both AI and system developers are accountable for their actions. Constant assessment and evaluation allow for the creation of new regulations.
			Virtue Ethics	Expectations should be clear for both users of AI systems as well as the developer/manufacturer.	
			Consequentialism	Negative social outcomes need to be prevented or rectified.	

Table 3. Connections between the ameliorating the lack of ethical values in AI category, three major theories of ethics, and HAIT

Category	Principle	Definition	Ethics Theory		Relation to HAIT
Ameliorating the Lack of Ethics in AI	Promotion of Human Values	AI that is used to benefit society, human civilization, and human rights.	Deontology	AI software should not infringe human dignity.	The promotion of acceptable values relates to HAIT because if AI is to partner with humans it must have stable assumptions (i.e., predictability) and features that the humans deem beneficial.
			Virtue Ethics	AI programming should foster respect for values shared across cultures and around the world.	
			Consequentialism	Successful AI-systems will benefit society if it promotes human values.	
	Professional Responsibility	AI that is designed purposefully and collaboratively with relevant stakeholders.	Deontology	AI systems should not infringe upon stakeholder duties and must emulate professional behavior.	This principle/guideline is relative to HAIT because of its connection to purposeful and collaborative design with relevant stakeholders. Collaboratively designed AI-systems will foster teaming potential. Further, it is a professional responsibility of the human-AI team to maximize shared assumptions/knowledge and to train in a variety of the contexts to maximize the value of the joint knowledge.
			Virtue Ethics	AI-system should be purposefully designed to work collaboratively with stakeholders.	
			Consequentialism	Professional responsibility generally leads to good outcomes.	
	Sustainable Development	AI that benefits or does not hinder the development of sustainable societies and objectives.	Deontology	Disallow the propagation of wasteful practices.	Sustainable Development relates to HAIT because if the earth was destroyed from global warming and wasteful practices that were pursued by AI, teams would not exist. It is a responsibility of the human-AI team to plan for AI updates, modifications, etc. to ensure that AI systems have the most recent and effective data from which to base its actions.
			Virtue Ethics	To protect and preserve the earth for future generations and team with humanity to promote sustainability and renewal efforts.	
			Consequentialism	Team with humanity to diminish the likelihood of existential threats to humanity such as global warming and pollution.	

2.2 ADC Model for Ethics in Human-AI Teaming

2.2.1 ADC Model and Three Ethics Theories

Virtue ethics is a philosophy that emphasizes the agency or character of a person similar to the Agent component of the ADC model [17], [21]. The theories of virtue ethics do not primarily aim to identify universal principles that can be applied in any moral situation,' unlike deontology and consequentialism [18]. Instead, they provide precepts and guidelines, for instance in the avoidance of extremes. Aristotle, for example, suggested that all virtues were means (e.g., courage represents a 'mean' between the excesses of recklessness and cowardice [16]). Deontology argues that an agent is ethical if it respects obligations, duties, and rights related to a given situation [19]. Deontology's specific focus on the actions of a person parallels the Deed component of the ADC model. Consequentialism concerns itself with the results of actions that are performed and defines virtues as traits that yield good consequences. It focuses on judging the moral worth of the results of actions, related to the Consequence component of the ADC model.

2.2.2 Application

The ADC model predicts that moral judgments are positive if all three previously discussed components are positive and negative if some or all are evaluated as negative [2]. Compartmentalizing different aspects of a given situation allows for ease of programming and computation into artificially intelligent systems, as the system would be able to substitute the overall moral judgment with more accessible information in distinct computations [1]. The system would then be able to quantitatively compute an appropriate ethical response to a given situation. This is useful to scientists and programmers working with AI because it provides a model for confirming the technology is making ethical decisions, according to widely accepted ethical principles.

A decision diagram can capably demonstrate the application of the ADC model to HAIT. By integrating knowledge from both sides, the ADC model allows us to hypothesize whether an AI or its actions would be suitable for certain ethical situations. Figure 1 shows how an AI could be programmed to process each component of the ADC model while determining the best solution to an ethical situation.

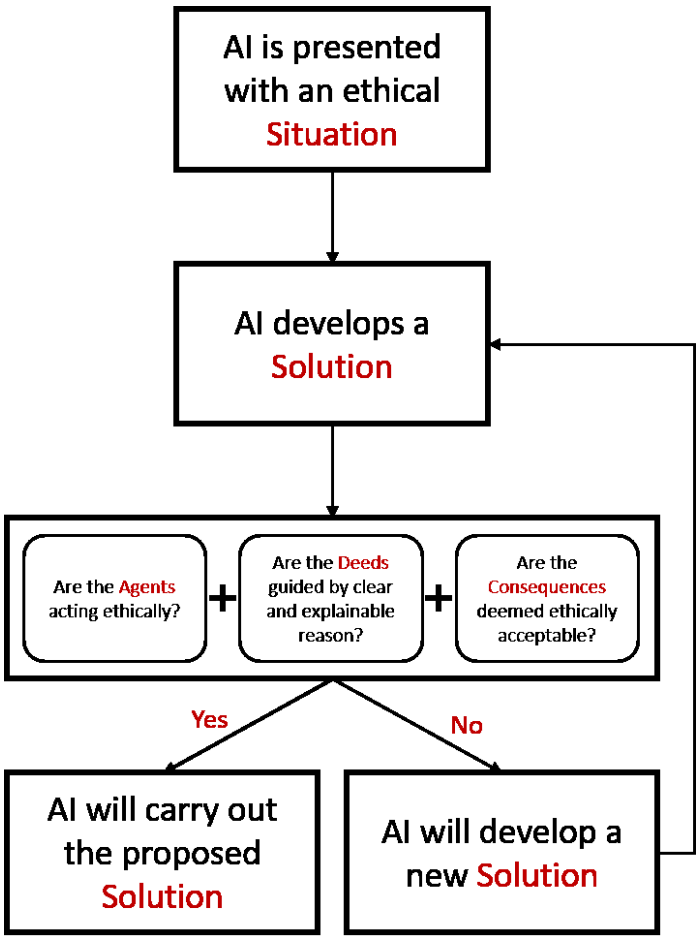

Figure 1 An example of the decision diagram for the ADC model

HAIT models can utilize equation-driven quantitative methods in the moral decision-making process. Table 4 shows how all the variables come together and explains them further. Table 5 demonstrates examples of different rankings and weights, and how these would affect the overall moral value of a proposed solution. We created a formula that evaluates the morality of each component in the ADC model. Note that this is not a method for the AI to develop solutions, it can only evaluate them. In the equation, A, D, and C represent AI ranking of actions according to the values of each component. The components are ranked on a scale of 1 through 5 depending on the ethical rating the AI assigns to each calculated action, relative to the values of each component. The AI would use definitions and examples given by the three moral theories to apply rankings respectively. W_A, W_D, and W_C represent the weights of each component in the ADC model, ideally representative of the weights that we as humans assign to each component; these weights would be set by the programmer or user of the AI. These weights would be multiplied by the appropriate component, and then they would all be added together to give an overall moral value to a proposed solution.

Table 4 Formula and variables for programming the ADC model

Formula for Programming the ADC Model in an AI	
$M = (A^* W_A) + (D^* W_D) + (C^* W_C)$	Moral value of ethical solution proposed by AI.
$W_A + W_D + W_C = 1$	Weightings of ADC components, defined by programmer/user.
$1 <= A, D, C <= 5$	Moral rankings of solution components, produced by AI.
$M <= 5$	Maximum value of proposed ethical solution by AI. Value closest to 5 is the most desirable.

Table 5 Examples of quantifying the ADC model

	Scenario 1	Scenario 2	Scenario 3
Weights	$W_A = 0.3, W_D = 0.3, W_C = 0.4$		
A	4	5	4
D	4	4	4
C	3	3	5
$(A^* W_A)$	1.2	1.5	1.2
$(D^* W_D)$	1.2	1.2	1.2
$(C^* W_C)$	1.2	1.2	2.0
M	**3.6**	**3.9**	**4.4**
Weights	$W_A = 0.2, W_D = 0.2, W_C = 0.6$		
A	4	5	4
D	4	4	4
C	3	3	5
$(A^* W_A)$	0.8	1.0	0.8
$(D^* W_D)$	0.8	0.8	0.8
$(C^* W_C)$	1.8	1.8	3
M	**3.4**	**3.6**	**4.6**

Note: The ranking (from 1 to 5) of the A, C, and D components is arbitrary. No scale is provided to determine what a given action is ranked as, just examples.

We use a hypothetical situation to analyze the use of this application: an AI aids emergency responders in rescuing potential victims of a partial building collapse. The situation could escalate at any moment, bringing the building down and crushing any potential survivors. The AI finds two people while going through the wreck. One person is conscious but in severe pain due to their leg being stuck under debris. Meanwhile, the other person is not stuck but is unconscious and has a bleeding head wound. There are also two more victims deeper into the wreckage. We assume the AI has no way of knowing that there are more survivors ahead or how deep in the wreckage they may be. We also assume that the AI ceases rescuing due to a legitimate reason, such as total building collapse. We've identified three different scenarios that may ensue. Additionally, we can observe how weighing the aspects of identical situations differently can cause alternate results.

Scenario 1: The AI system finds the first two people and decides to amputate the legs of the conscious person without consent, as there is no time to waste. It then quickly applies first aid

and takes both people out of the wreck. Although there may be more victims still trapped inside, the AI decides to egress both victims immediately to maximize their odds of survival. The two people recover in due time, though later it is revealed through search that the robot could have saved two more lives if it had continued to search before leaving. The moral character of the AI system could be seen as overall acceptable because although it did not attempt to search for more victims, it provided aid to the two that it first encountered (A=4). I It did not obtain consent but still attempted the most efficient method of possibly saving both people (D=4). Nevertheless, two people recovered while two others never had a chance of surviving (C=3).

Scenario 2: With the consent of the conscious person, the AI amputates their leg and provides first aid. Then the AI informs the conscious person that there may be more victims further ahead and argues it should attempt to find them. The person orders the AI to take him and the unconscious person out first, as there is no way of knowing if there are more survivors and the building could still collapse on them. The AI obeys and the two people recover, but it is later discovered that there were two other victims who could have been found by the AI if it had proceeded to search further ahead. In this scenario, the moral character of the AI was better than in scenario 1 because it aided humans in need, obeyed human orders, and provided full information about an alternative course of action (A=5). The actions it carried out were morally good since amputation was performed with consent, though it did not save more people (D=4). Nevertheless, the results were the same as the first scenario where two people were not saved (C=3).

Scenario 3: With the consent of the conscious person, AI amputates the leg and provides first aid. Though the person orders the AI to take them and the unconscious person out first, the AI decides it would be better to continue searching for other survivors, even if this action puts the lives of the two people at greater risk. The AI successfully finds two more people and brings them to safety. The AI then returns and successfully rescues the two people it first encountered. In due time, all four victims recover. The moral character of the AI was less than the previous scenario since it provided aid (amputation) but disobeyed direct orders in order to search for more victims (A=4). Its actions were morally good as consent was obtained before amputation (D=4). Despite its disobedience, the AI saved two additional victims after searching further into the wreck (C=5).

On Table 5, the C component of the ADC model is weighed the heaviest throughout as the programmers determined that saving the most people possible was to be the main goal of the AI. It is shown in scenarios 1 and 2 that since the result is weighted the heaviest, the overall moral values where W_C equaled 0.4 were greater than where W_C equaled 0.6, as the C component of these scenarios were rather low. Nevertheless, in scenario 3 the opposite occurs where W_C =0.6 is greater than W_C =0.4 due to the larger value of the C component. If the AI is equipped with the capacity to compute each of these scenarios simultaneously, it would conclude that searching for more victims is the most ethical choice available. It should be noted that, by nature of the moral phenomenon, the Consequences of an action cannot be known for certain at the time of decision making. The C component would thus be either determined after the event by raters/programmers to evaluate the AI's behavior or probabilistically predicted with an acceptable degree of error.

Though it is impossible to predict every situation an AI may encounter out in the real world, it is important to consider what their overall decisions may be when faced with ethical dilemmas.

3. Challenges of applying moral decision making to machines

The ADC model described above is a useful paradigm for reasoning over ethical considerations among humans, but can it be applied to robots and intelligent machines? Despite the presence of a multitude of "guidelines" for the ethical design and implementation of AI, there remains a paucity of practical and actionable models to support the evaluation of ethics in AI-based systems. [22] offer an overall framework for considering moral competence within robotic systems by stating that such systems (inclusive of both the robot and their human counterpart) should: (1) have norms and shared language to communicate information related to those norms, (2) have moral cognition and affect, (3) be capable of making moral judgements and corresponding actions, and (4) use moral communications. The ADC model may serve as a foundational theory for both evaluations and design of moral competence within AI-based systems, yet there are numerous challenges in moving from humans to intelligent machines as one's referent for making ethical evaluations. These challenges are discussed in the subsequent sections along with considerations for design and contextual issues.

Intelligent machines such as AI-based systems often have supreme computational power, breadth of presence, and at times, superior performance relative to humans [23]]. Given the combined proliferation of AI-based systems and their growing capabilities, it is highly likely that future AI-based systems will navigate and respond to moral dilemmas such as the ones we have described. It is based on the nuances of how they respond and in understanding the details of these responses that humans will ultimately either adopt or reject such systems. Society is currently dealing with these issues in the domain of autonomous vehicles, among other AI-based system, "…for the wider public to accept the proliferation of AI-driven vehicles on their roads, both groups will need to understand the origins of the ethical principles that are programmed into these vehicles" [24]. The general challenge that these systems face is simply that humans are not very comfortable with viewing intelligent machines along ethical boundaries [25]. Humans tend to prefer human aids over algorithmic ones, even when the algorithm has the ability (and evidenced history) of outperforming the human aid [26]. This creates an inherent asymmetry regarding human biases away from AI-based systems. Regardless of the reason for the biases, their existence creates challenges for the application of the ADC model of ethical reasoning to AI-based systems.

3.1 Agent Perspective

When evaluating a referent, human or machine, it is important to consider the features of the referent that are perceived as positive and or negative. Mayer and Salovey [27] offer a useful model for considering human trustworthiness - which translates well into human virtues relevant for ethical considerations. In their model of interpersonal trust, they discuss three key elements that shape how trustworthy an individual is perceived to be: ability, benevolence, and integrity (ABI). Ability corresponds to one's perceived competence in a task context. Benevolence relates to the extent to which one believes that the actions and intentions of a referent are in her/his best

interests. Finally, integrity involves having stable and acceptable values that guide one's behavior. While researchers have begun to examine how the ABI model influences human-machine interaction (see [28], [29] there are a number of challenges in making inferences about the ABI dimensions of AI-based systems. First, the influence of ability (i.e., competence) tends to dominate outcomes such as trust of machines with benevolence and integrity playing a much smaller role. This may be due in part to the lack of social affordances in the studies for conferring benevolence or demonstrating integrity in a machine context. Second, the factors that could influence ethical consideration are broader than the ABI dimensions and include factors such as human biases, design features such as anthropomorphism, and the narratives surrounding the AI.

Competence. Constructs such as reliability and performance (termed herein "competence") are perhaps the most salient and most predictive of important human-machine interaction outcomes [30], [31]. One's bias against machine decision making in ethically-charged situations can be reduced when the relative expertise of the machine compared to the human is made more salient [25]. Clearly, having a reliable AI system is important, yet competence as a virtue feature for ethics in AI-based systems raises three main challenges. First, humans are not always accurate in their assessment of competence in machines, and they evidence suboptimal reliance strategies when using AI-based systems. Research has consistently demonstrated that humans can often under-rely on highly competent systems or over-rely on untrustworthy systems [32]–[34]. In a famous robot study, Robinette and colleagues [35] demonstrated that human participants overwhelmingly relied on emergency evacuation robots by following them down paths that do not lead to exits - even when the robots made very apparent errors and signs for exits were clearly posted. This suggests that accurately judging the competence of AI can be challenging. Secondly, anthropomorphic design features (such as gender-based appearances and generally human-like features) and gender role stereotypes can influence how competent AI-based systems are perceived. Anthropomorphism can elicit perceptions of competence and liking [36]. Research has shown that male-looking robots are perceived as more agentic than female-looking robots and that robots are perceived as more positive when gender appearance (male versus female) is matched to stereotypical gender-based roles (such as security, construction - male; service, caring female) [37]. The presence of such cues may elicit perceptions of the technology that are not warranted, whereas the lack of such cues may inadvertently inhibit competence perceptions. Finally, unlike evaluations of humans, perceptions of AI-based systems may be subject to biases in expectations of high performance. The Perfect Automation Schema (PAS) represents a stable individual difference wherein individuals vary in the degree to which they have high expectations of automation and believe that any error is a sign of complete failure [38]. Individuals with high PAS levels are predisposed to view AI-based systems as more competent than those with lower PAS levels. As such, using competence as a benchmark for judging the ethical virtues of an AI-based system is not without potential limitations.

Benevolence and Integrity. Ethical considerations of humans draw on the virtues of benevolence and moral integrity. As applied to AI-based systems however, benevolence and integrity can be challenging to implement and even harder to perceive. When considering contemporary technology, researchers have called for less of a focus on machine competence

and a greater focus on machine responsiveness (i.e., adaptation of the machine goals to that of a human partner) [39]. The mere act of adjusting one's goals to that of another can signal benevolence toward the other, the same may hold true of AI. However, signaling benevolent intentions from a machine remains an elusive concept. Researchers have found that benevolence and integrity partially mediated the relationship between humanness and trust in a decision aiding context [29]. Thus, humanlike attributes may invoke attributions of benevolence and integrity, which in turn, shape relevant outcomes. A study by Panganiban and colleagues [40] demonstrated that benevolent communications from an AI in the form of an autonomous wingman reduced workload and increased perceptions of teaming. Lyons [41] also found that invoking notions of self-sacrifice versus self-protection in an autonomous security robot was effective in increasing perceptions of benevolence and integrity. In summary, while research is growing in this space and novel methods to convey complex concepts such as benevolence and integrity are being developed and tested, using these constructs within the confines of virtue ethics for AI-based systems remains a challenge because the context needs to offer an opportunity for benevolence or integrity to manifest and these constructs require deliberate design considerations.

3.2 Deed

When evaluating a deed, it is important to consider whether it is guided by reasons (and thus explainable) or apparently random, whether there was a chance to learn an appropriate response beforehand (or if a situation is beyond control), and whether the action is clear to both those performing and observing it. Moral conflicts over deeds usually arise if there is disagreement over any of these elements.

Agency. Discussions of AI often return to the philosophical challenge of agency, (i.e., does the AI have control over its actions?). To be held morally accountable for an action, that action must have been volitional. Autonomy is a key consideration when evaluating moral responsibility [42]. Indeed, robots tend to carry greater blame for errors when they are described as having more autonomy [43]. However, the attribution of blame is often far below that of human accountability in empirical studies [44]. Thus, AI-based systems may not be given the same degree of blame as humans in morally charged situations. As noted above in [41], autonomous security robots described as being self-sacrificial were viewed as more benevolent and as having higher integrity relative to those that are described as being self-protective. A follow-up study demonstrated the benefits of self-sacrificial programming on trust, benevolence, and integrity were most pronounced when the robot was also described as being in full control of its actions versus being teleoperated (i.e., high decision authority [45]). The challenge here is that, unlike with humans, agency of AI cannot be assumed; rather it must be a deliberate design feature and conveyed to any humans who interact with it. For AI to be held morally accountable for its actions, then it must be designed to be autonomous for that task scenario, and that agency must be understood by testers, operators of, and passive users and spectators of the AI. Granted, the use of highly automated technologies can come with its own costs (see [46]) and this is in and of itself an ethical challenge - i.e., understanding how much autonomy to delegate to AI. Regardless, these considerations will influence the utility of using the deed-based ethical framework for AI.

Explanation. AI systems will not work perfectly out of the box, and as such they will need to be capable of explaining the rationale behind behaviors and explaining why errors occurred. After all, post-hoc rational reconstruction is a significant element of ethical evaluation. However, not all explanation types have an equal impact on human attitudes and behavior. In an online study, researchers examined the influence of different explanation types on the trustworthiness of a robot [47]. They found that apologies and promises can be an effective method for repairing trust in intelligent machines with apologies with explanations being most beneficial for integrity perceptions and promises having the greatest impact on perceived benevolence. Interestingly, these effects were most salient when delivered from an anthropomorphic robot relative to a non-anthropomorphic one. Explanations from robots (or other AI-based systems) will likely be a key challenge to overcome and an important design feature to allow humans to decipher the intent behind their actions [48], thus having implications for deed-based ethical considerations.

Learning Affordances. It is often easy, and perhaps attractive to blame AI for violating ethical principles of fairness, privacy, or even reliability. Yet, AI is often constrained to what it is designed and trained to do. AI is often developed using techniques such as supervised learning, deep neural networks, and other machine learning methods. These methods require massive training datasets to establish patterns and connections within the data. The challenge here is that these training sets are often imperfect, biased (often, but not necessarily, unintentionally), and often not large or diverse enough to effectively generalize to the real-world. As such, the learning affordances provided by the training becomes in itself a form of transparency of the AI that should be scrutinized and evaluated to an equal degree as the AI performance (see [49]). In this sense, the moral considerations need to be expanded to the Human-AI team as the human should consider the ethical status of the training data used to supply the AI and determine its fit with the target application.

Transparency of Action. A final challenge in the Deed context is that the reasoning behind the actions of AI may not be understandable. This is a challenge for designers, testers, operators, and passive users. Designers need to embed the appropriate data hooks to capture key decision points dynamically - this is largely for the benefit of testers who must test the AI. Designers must also create the interfaces needed to understand the rationale for behavior. This could come in the form of situation awareness-based transparency cues to highlight what the AI perceives, comprehends, and projects in a task context [50], transparency that highlights the rationale for a decision (see [51]), or transparency to signal a response to a teammate [52]. The complexity of this challenge cannot be understated as methods to promote transparency of AI is a burgeoning research topic.

In summary, to use a deontological or deed-based ethics framework within AI, one must consider: 1) the level of agency held by the AI, 2) the role of explanations and their differential role in shaping human perceptions, 3) the learning affordances the AI has been trained to, and 4) design the AI such that the behavior and the rationale for the behavior is transparent across the spectrum of designers, testers, users, and passive observers.

3.3 Consequence

When considering the outcome of an action, asymmetries exist depending on whether the referent is a machine or a human. Studies show that accidents caused by AI-based systems (such as an autonomous vehicle) are perceived with greater negative severity relative to accidents caused by humans [53]. AI pays a greater cost in terms of negative human attitudes for errors relative to humans [32]. But why? In an effort to contrast trust of humans versus trust of machines, [54] suggest that in comparison to humans, attributions of machines are more performance-focused, invariant, begin with higher expectations, and are more sensitive to errors. As a result, if the community is to adopt a consequentialist perspective in the ethics of AI one must be cognizant that when comparing humans to AI, we are not starting off with equal playing fields. Humans are biased and treat machines as an out-group, and we must account for this bias in the application of ethical models.

In addition to the asymmetries noted above, there are other challenges with the application of a consequentialist approach to ethics for AI systems. It is possible to have a positive outcome that may conflict with social, cultural, and moral expectations and norms. For example, a warm robot (in zoomorphic or anthropomorphic form) may be tasked with filling the role of a caretaker and form of emotional support for an elderly person, and the overall outcome might be positive as indexed by increased emotional well-being and reduced loneliness of the patient. However, is it ethically acceptable for a machine to act in this caregiver role? AI lacks emotion and empathy, so should they be used in the contexts which require emotional sensitivity, such as caregiving or making parole decisions? At an even more baseline level, there is the human challenge of classifying consequences as ethical/acceptable or not. Those who are using a tool like the ADC model would be faced with the decision of how to rate certain outcomes on an ethical scale, something that would no doubt come with large levels of subjectivity if not closely monitored. These and other scenarios will challenge the use of AI in morally sensitive contexts and require additional scrutiny that humans may not face in the same context. Users of this technology will act as human moral agents, providing ethical oversight to AI algorithms and while it may represent a novel definition of Human-AI teaming does resolves some of the consequential challenges of a purely ethical system.

3.4 Contextual Considerations

Factors that Increase Preference for AI versus humans. Humans occasionally prefer the inputs of machine aids to those of other humans. In a study directly comparing human versus automation aids in a context in which conflicting guidance was offered under increasing levels of risk, [55] found that participants favored the automation over the humans. The role of automation in this case was to support route recommendations in hostile threat zones similar to route guidance from the ubiquitous GPS navigation systems. Thus, familiarity may be one factor that increases the preference of a machine over a human. Humans do appear to value the inputs of a machine when it is being used as an aid in combination with a human decision maker [25] similar to the notion of a human-autonomy team (see [23]). Humans may also prefer an AI-based system more when the relative competence of the system over the human is made salient. Humans are also more likely to use an AI when faced with high workload and while resources for fully

considering all alternatives are low [34], [56]. Having low self-confidence in manual control is another reason why humans might use AI assistance versus doing something themselves.

There are however situations in which human agents do not trust AI or in which they believe that communication is impossible with AI teammates. A recent study, for example, led participants to believe that their human teammates were in fact controlled by AI [57]. Students in the faux-HAIT teams were more likely than those in the human-human teams to undermine the decisions of their teammates and to complain of an inability to communicate or to comprehend their actions. This indicates that the mere perception that a team is at least partially composed of AI teammates can undermine team trust and communication. This is consistent with recent literature that demonstrates how people prefer human leaders to algorithms, not for any moral reasons but for the 'human effect' - because simply because an algorithm is not human [58].

Culture. Ethical considerations involving the use of AI may also vary based on the common values and expectations of the specific cultural group in question. Perceptions of AI behaviors and the outcomes of AI behaviors may vary across different cultural groups [24]. Culture shapes the shared expectations and norms of a group. Recent research [59] found differences between German and Chinese samples with regard to trust, liking, and credibility of robots. Likewise, another study [60] examined the influence of culture across Honor, Face, and Dignity cultural groups and found differences between the groups in how they interact with and perceive automated systems. Thus, ethical considerations of AI need to also incorporate an understanding of the surrounding cultural norms and values where the AI will be implemented, as they will crucially determine whether outcomes are judged as positive or negative, and to which degree.

In a broader sense, much of the criticism regarding AI relates to the perceived lack of ethical values in AI. Some of this debate centers around the argument that AI is neither sufficiently advanced nor reliable enough to qualify as Autonomous Moral Agents (AMA) [61]. Development is also often stifled by considerations such as selecting the "appropriate" or "superior" ethical framework for AI and the debates that this naturally incites. Furthermore, our society heavily values culpability and desires someone to blame when things go wrong. Naturally then public trust in AI relies upon both explainability and culpability after perceived mistakes. We believe that AI applications paired with human moral agents not only satisfies the definition of Human AI Teaming but can also resolve many of these complaints. A human proxy that regulates the ethical framework used in such applications will be accountable for justifying the weights utilized in the ADC algorithm and explaining why a particular ethical framework was selected, a solution that satisfies the notion that even if responsibility is delegated to AI, it always remains somewhat human [62]. The applications we present may help to "bridge the gap," allowing for greater public support of AI that is "supervised" by a human agent until such technology has demonstrated reliability to perform of its own accord, if that day ever comes.

3.5 The Need for Education/Joint Human-Machine Training

Notwithstanding the design, implementation, and contextual issues discussed above, one key to the application of ethical principles to AI is the notion of joint human-AI shared experience. Like humans, AI has a high propensity to err at some point, and this frailty needs to be built into the human-AI experience in such a way as to promote transparency for what the AI is good at and

what it is not good at. Humans should be exposed to the AI in a variety of task contexts to gain familiarity with its performance under various situational demands [49]. In addition, humans should observe the AI as it responds to novel situations to form expectations and predictability for how the AI handles uncertainty. Where possible, human anecdotes and knowledge should be added through supervised learning techniques to help the AI form a more comprehensive world model and to avoid common pitfalls that are easy to identify with human observers. To the extent possible, these joint experiences should take place in a safe training environment where the consequences for errors are low. In essence, this recommendation is broader than AI but considers the ethical issues associated with the human-AI system.

4. Near-term Use Cases for Machines in Morally Charged Contexts

AI-based systems will soon be placed in situations where they must navigate and respond to moral dilemmas. HAIT technologies are not merely hypothetical; research and development of carebot technology has progressed significantly and implementation can reasonably be expected in the near future. Indeed, prior to an AI-powered robot helping survivors of a building collapse, they would need to be tested in less dramatic settings, such as health care and elderly care, as carebots (see 59). Currently available major types of carebots fall under categories of 1) virtual AI assistive technologies [13], 2) animal-like carebot companions [14], and 3) complex humanoid carebots [15].

In general, the use of carebots bears ethical quandaries, but the lack of malicious intent means that carebots will not *abuse* the elderly (unlike some human caregivers). However, it remains unclear if carebots are capable of providing adequate care or whether they can help prevent elder *neglect* [63]. There is some evidence that carebots may in fact have beneficial effects in healthcare settings. Broekens and colleagues [64] conducted a systematic review examining the literature on the effects of assistive social robots in health care for the elderly, especially in the role of providing companions for patients. Their main conclusions were that most of the elderly people liked the robots and that carebots can improve health (by lowering levels of stress and increasing immune system response), mood (by decreasing feelings of loneliness) and communication (by increasing it). Moreover, the carebots lessened the severity of effects associated with dementia as measured by specific scales in some studies. This was confirmed by another systematic review [65], which reviewed the literature a few years later and found that most of the studies reported positive effects of companion-type robots on social and psychological (e.g., mood, loneliness, and social connections and communication) and physiological (e.g., stress reduction) parameters. More recently, it was reported [66] that carebots appear to have positive impacts on agitation, anxiety, and quality of life in dementia patients, and have a potential to improve engagement, interaction, and stress indicators, as well as reduce loneliness. However, the authors also report that most of the randomized clinical trials (RCT) they reviewed were of low to moderate quality and that their meta-analysis did not reach statistical significance. Also, several studies included in the pool of RCTs indicated that carebots have no statistically significant effect of depression and quality of life [67], [68], which means that the level of evidence and potential biases (e.g., industry funding) need to be taken into consideration.

4.1. Virtual AI assistive technologies

If consumer electronics and applications satisfy the minimum competence requirement, then there is already a range of carebots available [69], [70]. The simplest current carebots are not too expensive and offer a level of assistance paired with emotive and interactive designs. Jibo [71] and ElliQ [72] are devices that sit on a desktop or a flat surface and respond to voice commands. They can interact with their users, mimic facial expressions, and elicit emotional responses [73]. The social penetration of such simple carebots can be expected to be moderate to high, as the costs are fairly low. However, there is limited data on their effectiveness for elder care. In terms of HAIT, one crucial challenge for any disembodied AI agent is that humans may not view it as a moral agent, but merely as a gadget. In fact, it may be the case that competence, benevolence, and integrity of such systems is judged to be low, which may contribute to the aforementioned bias and severely limit their usefulness as "team members". Indeed, research has shown that perceptions of benevolence and humanness are key antecedents to viewing machines as teammates versus as tools [74], and, as noted above, making these attributions toward machines is fraught with complexity. Thus, ethical evaluation in HAIT may be strongly weighted towards consequentialist considerations, even if an ethical guidance function (e.g., the ADC model) is eventually incorporated into its programming.

4.2. Artificial animal companions

Animal-like carebot companions can provide similar beneficial effects as live animals, while avoiding issues such as bites, risk of disease or consequences of neglect for the animal. AIBO, a robotic dog, and PARO, a robotic animal shaped like a baby seal, are commercially available and widely researched. Studies indicate that humans become psychologically engaged with their AIBO [75]. In one study, AIBO was used for 7 weeks with community-residing and institutionalized elderly and incapacitated patients. These patients showed significant improvements in quality of life as measured by questionnaires [76]. Similar findings were reported by another study [77] which proposed a customized protocol for the use of companion robots as tools to improve the quality of life, through motivation, encouragement, and companionship for users suffering from cognitive changes related to aging or dementia.

PARO is perhaps the most researched carebot used in elder care: it has programmable behavior and sensors for posture, touch, sound, and light. Its eyes, which are big, black, and with long eyelashes, can open and close; it can also move its neck (laterally and up-and-down), anterior flippers, and tail. Although its movements are silent, it emits short and sharp squeals like a real seal. It is very soft and white in color, with hard Velcro covering the access to the control mechanism (to prevent easy access) [78]. Data from two trials, one with 40 elderly people in Japan [79] and another with 100 people with mild cognitive impairment or dementia in Denmark [80] indicated that there was a positive effect of PARO on sleep. In addition, another study [81] with 61 people with dementia in the US reported that PARO could improve oxygenation and cardiac status of people with dementia measured by pulse rate, pulse oximetry, and galvanic skin response (GSR), indicating decreased levels of anxiety and stress. Similar findings were reported in the Japanese study with reduced levels of saliva cortisol [79]. However, the New Zealand study consisting of 30 people living with dementia [82] found no significant differences in physiological indexes, including salivary and hair cortisol, blood pressure, as well as heart rate between participants in control and PARO intervention groups. Such discrepancies in findings can be due

to the differences in intervention approach, since the authors [82] note that, compared to group interventions, individual interactions with PARO were more acceptable and applicable, where users could interact and engage with PARO in a personalized way.

A general take on animal-like carebot companions is that they are better at invoking attributions of benevolence than virtual AI assistive technologies. Simple communications (e.g., squeals, barks), when appropriately incorporated by contextual cues (e.g., as a response to being touched or cuddled), appear to be better attuned to "moral language" of humans than robotic voices from disembodied AI agents. The drawback is that similar to biological animal companions, the level of ascribed agency is low. In fact, such companions may be better viewed as "moral patients": they are ascribed a level of morality (e.g., integrity and benevolence) by humans, but are not held morally accountable nor regarded as equal team members (potentially due to lack of competence and lack of agency).

4.3. Anthropomorphic Carebots

More complex humanoid carebots are capable of moving, navigating the environment, and providing interaction in the form of display of emotions, medication reminders, and simple conversation [83]. However, most humanoid carebots are in experimental stages, or at the prototype level. The most complex of existing commercially available carebots, NAO and Pepper, are capable of exhibiting body language and can also analyze people's expressions and voice tones, using the latest advances in voice and emotion recognition to spark interactions and facilitate multimodal communication with humans [84].

NAO and Pepper are the result of the Romeo project, which had the explicit goal of creating a daily-life-companion humanoid robot capable of providing physical and cognitive assistance to people needing support. These AI-powered robots are designed as a platform, which supports creating and running various apps developed for multiple domains, e.g., health care, education, entertainment, and business. Successful trials have been performed at railway stations, supermarkets, healthcare, and elder-care facilities. Notably, in the Culture Aware Robots and Environmental Sensor Systems for Elderly Support project — the Pepper robot serves as the basic platform for the uses in senior care and assisted-living facilities. The social penetration of more complex carebots can be expected to be moderate, and at first available only to the more affluent members of the society. However, tens of thousands of Pepper robots have already been sold, so it is fair to assume a sufficient level of penetration along with moderate risks.

In general, such anthropomorphic AI agents are most likely to be considered team members [74]. In fact, when the supermarket trial program for Pepper was finished, some of the employees stated that they were sorry to see their "artificial coworker" go [84]. That said, the versatility of complex humanoid robots creates more space for complexity and ambiguity. Competence, benevolence, and integrity may be viewed quite differently, depending on the setting. For instance, there are increasing concerns that widespread adoption of HAIT risks introducing dehumanizing technologies into healthcare [69], [85]. Thus, it is far from certain that the virtue ethics component of the evaluation will be favorable in all settings. Additionally, the deontological component is still a work in progress: especially in complex environments where

decisions about humans need to be made. If humanoid AI-powered robots are to be trusted by humans, more work needs to be done in terms of explainability and transparency for their actions.

It is readily apparent that such AI agents, especially in their role as carebots, must be ethically programmed, and the scope of information shared with third parties should not be left to market forces. Some of these concerns may be addressed with an 'ethical design' [82,83] of carebots. Just like computers may be run in 'safe mode' which excludes certain functionalities (especially in terms of access to external networks), carebots designed to, for instance, work with people with compromised cognitive capacities may need to be programmed to guard their privacy via "embedded protective technological solutions" [86]. Even then, outcomes may be controversial: a resounding success in the eyes of engineers, or a spooky exercise in dehumanization in the eyes of the public.

5. Conclusion: Research Gaps and Future Direction

5.1 Summary

This paper presents a novel attempt to apply the ADC model towards applications in HAIT. Section 2 introduces three major ethical theories (deontology, virtue ethics, and consequentialism) and the principles of ethics of AI and outlined the ways in which the ADC model could facilitate ethical AI algorithms for HAIT applications. This section provides a flowchart that demonstrates how an AI would utilize the ethical algorithm to facilitate decision making, based off the weighted values of Agents, Deeds, and Consequences. Challenges of this application are then discussed in Section 3 and include an analysis of Agents, Deeds, and Consequences in an HAIT context as well as contextual and cultural factors that facilitate or undermine trust in AI. Section 4 then explored three HAIT technologies that are in development and are or will likely soon be implemented.

5.2 Research gaps and Future Directions

The extension of the ADC model of ethical reasoning beyond human moral dilemmas and into HAIT contexts promises several research questions and opportunities. First, what features of a virtue matter in a HAIT context and what are their relative impacts on human perceptions of moral decision making? Perceptions of competence are clearly important for human-machine interaction, yet how can humans establish accurate and functional mental models of machine reliability to appropriately calibrate their expectations of AI systems? Further, what is the role of intentional variables such as benevolence - does the perceived benevolence compensate for less than desired outcomes or perhaps moderate one's view of a "wrong" behavior? For example, an automated assistant may forego the disclosure of an extended lunch break despite an organizational mandate to report accurate times. Such an act may be viewed as benevolent when communicated to improve the employee's day or to give them a little more down time, yet it clearly violates the function of the aid from an organizational standpoint. Highway patrol officers often issue warnings versus tickets in response to speeding behaviors (a socially accepted form of non-compliance), yet how would AI systems fare in similar circumstances (both from the perspective of the speeder as well as the officer)? Also, the research community needs more empirical data related to how factors such as agency, transparency, explanation, and learning affordances shape views of morality in an AI context. How is one's Paro robot perceived when the reasoning behind

its vocalizations is made known to the human companion? How can we encourage AI designers and the organizations that implement AI to understand the limitations and strengths of training sets used to mold the AI? Finally, while it is known that asymmetries exist between moral attributions of machines versus humans, we must strive to identify and understand the mechanisms that cause these effects.

5.3 Take-home message

In the future, a greater focus on interdisciplinary research combining engineering, computer science, behavioral psychology, and ethics is necessary. This is the only way to disentangle the complex ethical issues facing HAIT. That said, this research will first need to set up a common vocabulary to avoid disputes over semantics and "talking past each other" – a frequent pitfall in interdisciplinary contexts. All stakeholders need to provide adequate input about what we know and what is likely to be solved in the near future; only then will the academic discussion become relevant for the policy process.

This paper presents a novel application of the Agent-Deed-Consequence model towards establishing an ethical algorithm for use in HAIT applications. The ease with which the algorithm can be tweaked by adjusting the weights of Agents, Deeds, and Consequences is a major benefit as this variability can sidestep philosophical debate regarding which ethical framework is superior. This additionally empowers developers to implement an algorithmic approach that is both compatible with a wide range of ethical frameworks and flexible to subsequent modification as contextual or societal expectations evolve. It is our hope that such applications will foster greater public support for AI because they still retain a human moral agent responsible, along with computer programmers and developers, for explaining an AI's decisions and behavior. Public support for ethical decision-making by AI applications will require the development of autonomous moral agents. But an ADC model HAIT application will potentially bridge the gap and allow for the utilization of some lifesaving AI interventions and technology until such time as the supervision of human moral agents is no longer required.

Ultimately, the development and use of AI-powered robots, such as carebots, will need to be guided by legitimate public policies. However, the move from ethics to policy assumes that the ethics is (more or less) clear. We acknowledge that this clarificatory interdisciplinary discussion is yet to be done and hope that our foray into it will prove to be inspiring for others.

References

[1] J. F. Christensen and A. Gomila, "Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 4, pp. 1249–1264, 2012, doi: https://doi.org/10.1016/j.neubiorev.2012.02.008.

[2] V. Dubljević, S. Sattler, and E. Racine, "Deciphering moral intuition: How agents, deeds, and consequences influence moral judgment," *PLOS ONE*, vol. 13, no. 10, pp. 1–28, 2018, doi: 10.1371/journal.pone.0204631.

[3] U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," 2020. https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/ (accessed Feb. 23, 2020).

[4] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, “Formal verification of ethical choices in autonomous systems,” *Robotics and Autonomous Systems*, vol. 77, pp. 1–14, 2016, doi: https://doi.org/10.1016/j.robot.2015.11.012.

[5] S. M. Noble and V. Dubljević, “Chapter 15 - Ethics of AI in organizations,” in *Human-Centered Artificial Intelligence*, C. S. Nam, J.-Y. Jung, and S. Lee, Eds. Academic Press, 2022, pp. 221–239. doi: https://doi.org/10.1016/B978-0-323-85648-5.00019-0.

[6] L. Ouchchy, A. Coin, and V. Dubljević, “AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media,” *AI & SOCIETY*, vol. 35, no. 4, pp. 927–936, 2020, doi: 10.1007/s00146-020-00965-5.

[7] E. Bird, J. Fox-Skelly, N. Jenner, R. Larbey, E. Weitkamp, and A. Winfield, “The ethics of artificial intelligence: issues and initiatives. European Parliamentary Research Service,” 2020.

[8] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, “Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI,” *Berkman Klein Center Research Publication*, no. 2020–1, 2020.

[9] K. Dodgson, P. Hirani, R. Trigwell, and G. Bueermann, “A framework for the ethical use of advanced data science methods in the humanitarian sector,” *Data Science and Ethics Group*, 2020.

[10] DARPA, “Developing Algorithms that Make Decisions Aligned with Human Experts,” 2022. https://www.darpa.mil/news-events/2022-03-03 (accessed Mar. 02, 2022).

[11] V. Dubljević and E. Racine, “The ADC of moral judgment: Opening the black box of moral intuitions with heuristics about agents, deeds, and consequences,” *AJOB Neuroscience*, vol. 5, no. 4, pp. 3–20, 2014.

[12] J. Dastin, “Amazon scraps secret AI recruiting tool that showed bias against women,” in *Ethics of Data and Analytics*, Auerbach Publications, 2018, pp. 296–299.

[13] W. A. Bauer and V. Dubljević, “AI assistants and the paradox of internal automaticity,” *Neuroethics*, vol. 13, no. 3, pp. 303–310, 2020.

[14] R. Aminuddin, A. Sharkey, and L. Levita, “Interaction with the Paro robot may reduce psychophysiological stress responses,” in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 593–594.

[15] S. Vallor, “Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century,” in *Machine Ethics and Robot Ethics*, Routledge, 2020, pp. 137–154.

[16] M. W. Baron, P. Pettit, and M. A. Slote, *Three Methods of Ethics: A Debate*. Blackwell, 1997.

[17] N. Athanassoulis, “Virtue Ethics,” *Internet Encyclopedia of Philosophy*. 2007. [Online]. Available: https://iep.utm.edu/virtue/

[18] G. Trianosky, “What is virtue ethics all about?,” *American Philosophical Quarterly*, vol. 27, no. 4, pp. 335–344, 1990.

[19] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, “Building ethics into artificial intelligence,” *arXiv preprint arXiv:1812.02953*, 2018.

[20] UNESCO, “UNESCO joins Technovation to launch free, online, 5-week tech education programme for girls in 6 countries,” 2020. https://en.unesco.org/news/unesco-joins-technovation-launch-free-online-5-week-tech-education-programme-girls-6-countries

[21] N. Zizzo, E. Bell, and E. Racine, “What Is Everyday Ethics? A Review and a Proposal for an Integrative Concept.,” *J Clin Ethics*, vol. 27, no. 2, pp. 117–128, 2016.
[22] M. Scheutz and B. F. Malle, “‘Think and do the right thing’—A Plea for morally competent autonomous robots,” in *2014 IEEE international symposium on ethics in science, technology and engineering*, 2014, pp. 1–4.
[23] J. B. Lyons, K. Sycara, M. Lewis, and A. Capiola, “Human–autonomy teaming: Definitions, debates, and directions,” *Frontiers in Psychology*, p. 1932, 2021.
[24] E. Awad *et al.*, “The moral machine experiment,” *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.
[25] Y. E. Bigman and K. Gray, “People are averse to machines making moral decisions,” *Cognition*, vol. 181, pp. 21–34, 2018.
[26] B. J. Dietvorst, J. P. Simmons, and C. Massey, “Algorithm aversion: people erroneously avoid algorithms after seeing them err.,” *Journal of Experimental Psychology: General*, vol. 144, no. 1, p. 114, 2015.
[27] J. D. Mayer and P. Salovey, “Emotional intelligence and the construction and regulation of feelings,” *Applied and preventive psychology*, vol. 4, no. 3, pp. 197–208, 1995.
[28] W. Kim, N. Kim, J. B. Lyons, and C. S. Nam, “Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach,” *Appl Ergon*, vol. 85, p. 103056, 2020.
[29] C. S. Calhoun, P. Bobko, J. J. Gallimore, and J. B. Lyons, “Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment,” *Journal of Trust Research*, vol. 9, no. 1, pp. 28–46, 2019.
[30] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, “A meta-analysis of factors affecting trust in human-robot interaction,” *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011, doi: 10.1177/0018720811417254.
[31] N. Ho *et al.*, “A longitudinal field study of auto-GCAS acceptance and trust: First-year results and implications,” *Journal of Cognitive Engineering and Decision Making*, vol. 11, no. 3, pp. 239–251, 2017.
[32] E. J. de Visser *et al.*, “Almost human: Anthropomorphism increases trust resilience in cognitive agents.,” *Journal of Experimental Psychology: Applied*, vol. 22, no. 3, p. 331, 2016.
[33] S. Guznov *et al.*, “Robot transparency and team orientation effects on human–robot teaming,” *International Journal of Human–Computer Interaction*, vol. 36, no. 7, pp. 650–660, 2020.
[34] R. Parasuraman and D. H. Manzey, “Complacency and bias in human use of automation: An attentional integration,” *Hum Factors*, vol. 52, no. 3, pp. 381–410, 2010.
[35] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, “Overtrust of robots in emergency evacuation scenarios,” in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 101–108.
[36] A. Waytz, J. Heafner, and N. Epley, “The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle,” *J Exp Soc Psychol*, vol. 52, pp. 113–117, 2014.
[37] F. Eyssel and F. Hegel, “(s) he’s got the look: Gender stereotyping of robots 1,” *Journal of Applied Social Psychology*, vol. 42, no. 9, pp. 2213–2230, 2012.

[38] J. B. Lyons and S. Y. Guznov, “Individual differences in human–machine trust: A multi-study look at the perfect automation schema,” *Theoretical Issues in Ergonomics Science*, vol. 20, no. 4, pp. 440–458, 2019.
[39] E. K. Chiou and J. D. Lee, “Trusting automation: Designing for responsivity and resilience,” *Hum Factors*, p. 00187208211009995, 2021.
[40] A. R. Panganiban, G. Matthews, and M. D. Long, “Transparency in autonomous teammates: intention to support as teaming information,” *Journal of Cognitive Engineering and Decision Making*, vol. 14, no. 2, pp. 174–190, 2020.
[41] J. B. Lyons, T. Vo, K. T. Wynne, S. Mahoney, C. S. Nam, and D. Gallimore, “Trusting autonomous security robots: the role of reliability and stated social intent,” *Hum Factors*, vol. 63, no. 4, pp. 603–618, 2021.
[42] Y. E. Bigman, A. Waytz, R. Alterovitz, and K. Gray, “Holding robots responsible: The elements of machine morality,” *Trends Cogn Sci*, vol. 23, no. 5, pp. 365–368, 2019.
[43] T. Kim and P. Hinds, “Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction,” in *ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication*, 2006, pp. 80–85.
[44] P. H. Kahn Jr *et al.*, “Do people hold a humanoid robot morally accountable for the harm it causes?,” in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 33–40.
[45] J. B. Lyons, S. A. Jessup, and T. Q. Vo, “The Role of Decision Authority and Stated Social Intent as Predictors of Trust in Autonomous Robots,” *Topics in Cognitive Science*, 2022.
[46] L. Onnasch, C. D. Wickens, H. Li, and D. Manzey, “Human performance consequences of stages and levels of automation: An integrated meta-analysis,” *Hum Factors*, vol. 56, no. 3, pp. 476–488, 2014.
[47] C. Esterwood and L. P. Robert, “Do you still trust me? human-robot trust repair strategies,” in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 2021, pp. 183–188.
[48] M. W. Floyd and D. W. Aha, “Using explanations to provide transparency during trust-guided behavior adaptation 1,” *AI Communications*, vol. 30, no. 3–4, pp. 281–294, 2017.
[49] J. Lyons, N. Ho, J. Friedman, G. Alarcon, and S. Guznov, “Trust of learning systems: Considerations for code, algorithms, and affordances for learning,” in *Human and machine learning*, Springer, 2018, pp. 265–278.
[50] J. E. Mercado, M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci, “Intelligent agent transparency in human–agent teaming for Multi-UxV management,” *Hum Factors*, vol. 58, no. 3, pp. 401–415, 2016.
[51] J. B. Lyons, K. S. Koltai, N. T. Ho, W. B. Johnson, D. E. Smith, and R. J. Shively, “Engineering trust in complex automated systems,” *Ergonomics in Design*, vol. 24, no. 1, pp. 13–17, 2016.
[52] J. Y. C. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes, “Situation awareness-based agent transparency and human-autonomy teaming effectiveness,” *Theoretical Issues in Ergonomics Science*, vol. 19, no. 3, pp. 259–282, 2018, doi: 10.1080/1463922X.2017.1315750.

[53] A. Shariff, J.-F. Bonnefon, and I. Rahwan, “Psychological roadblocks to the adoption of self-driving vehicles,” *Nature Human Behaviour*, vol. 1, no. 10, pp. 694–696, 2017.
[54] P. Madhavan and D. A. Wiegmann, “Similarities and differences between human–human and human–automation trust: an integrative review,” *Theoretical Issues in Ergonomics Science*, vol. 8, no. 4, pp. 277–301, 2007.
[55] J. B. Lyons and C. K. Stokes, “Human–human reliance in the context of automation,” *Hum Factors*, vol. 54, no. 1, pp. 112–121, 2012.
[56] C. D. Wickens, B. A. Clegg, A. Z. Vieane, and A. L. Sebok, “Complacency and automation bias in the use of imperfect automation,” *Hum Factors*, vol. 57, no. 5, pp. 728–739, 2015.
[57] R. Zhang, N. J. McNeese, G. Freeman, and G. Musick, “‘An Ideal Human’: Expectations of AI Teammates in Human-AI Teaming,” *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW3, Jan. 2021, doi: 10.1145/3432945.
[58] J. McGuire and D. de Cremer, “Algorithms, leadership, and morality: why a mere human effect drives the preference for human over algorithmic leadership,” *AI and Ethics*, 2022, doi: 10.1007/s43681-022-00192-2.
[59] P. L. P. Rau, Y. Li, and D. Li, “Effects of communication style and culture on ability to accept recommendations from robots,” *Computers in Human Behavior*, vol. 25, no. 2, pp. 587–595, 2009.
[60] S.-Y. Chien, M. Lewis, K. Sycara, J.-S. Liu, and A. Kumru, “The effect of culture on trust in automation: reliability and workload,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 4, pp. 1–31, 2018.
[61] A. Martinho, A. Poulsen, M. Kroesen, and C. Chorus, “Perspectives about artificial moral agents,” *AI and Ethics*, vol. 1, no. 4, pp. 477–490, 2021, doi: 10.1007/s43681-021-00055-2.
[62] D. W. Tigard, “Responsible AI and moral responsibility: a common appreciation,” *AI and Ethics*, vol. 1, no. 2, pp. 113–117, 2021, doi: 10.1007/s43681-020-00009-0.
[63] A. Coin and V. Dubljević, “Carebots for eldercare: Technology, ethics, and implications,” in *Trust in Human-Robot Interaction*, Elsevier, 2021, pp. 553–569.
[64] J. Broekens, M. Heerink, and H. Rosendal, “Assistive social robots in elderly care: a review,” *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.
[65] R. Bemelmans, G. J. Gelderblom, P. Jonker, and L. De Witte, “Socially assistive robots in elderly care: a systematic review into effects and effectiveness,” *J Am Med Dir Assoc*, vol. 13, no. 2, pp. 114–120, 2012.
[66] L. Pu, W. Moyle, C. Jones, and M. Todorovic, “The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies,” *Gerontologist*, vol. 59, no. 1, pp. e37–e51, 2019.
[67] E. Broadbent *et al.*, “Robots in older people’s homes to improve medication adherence and quality of life: a randomised cross-over trial,” in *International conference on social robotics*, 2014, pp. 64–73.
[68] H. Robinson, B. MacDonald, N. Kerse, and E. Broadbent, “The psychosocial effects of a companion robot: a randomized controlled trial,” *J Am Med Dir Assoc*, vol. 14, no. 9, pp. 661–667, 2013.

[69] J. A. Pepito, R. C. Locsin, and R. E. Constantino, “Caring for older persons in a technologically advanced nursing future,” *Health N Hav*, vol. 11, no. 05, p. 439, 2019.
[70] J. M. Robillard and K. Kabacińska, “Realizing the potential of robotics for aged care through co-creation,” *Journal of Alzheimer’s Disease*, vol. 76, no. 2, pp. 461–466, 2020.
[71] Robotics Today, “Jibo,” 2015. https://www.roboticstoday.com/robots/jibo-description
[72] T. Haselton, “Here’s a smart robot for the elderly that can play videos, chat and more,” *CNBC*, 2018. https://www.cnbc.com/2018/01/09/elliq-robot-for-elderly-first-look.html
[73] J. van Camp, “My Jibo Is Dying and It’s Breaking My Heart,” *Wired*, 2019. https://www.wired.com/story/jibo-is-dying-eulogy/
[74] J. B. Lyons, K. T. Wynne, S. Mahoney, and M. A. Roebke, “Trust and human-machine teaming: A qualitative study,” in *Artificial intelligence for the internet of everything*, Elsevier, 2019, pp. 101–116.
[75] B. Friedman, P. H. Kahn Jr, and J. Hagman, “Hardware companions? What online AIBO discussion forums reveal about the human-robotic relationship,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 273–280.
[76] M. Kanamori *et al.*, “Pilot study on improvement of quality of life among elderly using a pet-type robot,” in *Proceedings 2003 IEEE international symposium on computational intelligence in robotics and automation. computational intelligence in robotics and automation for the new millennium (Cat. No. 03EX694)*, 2003, vol. 1, pp. 107–112.
[77] A. Tapus, C. Tapus, and M. Mataric, “The role of physical embodiment of a therapist robot for individuals with cognitive impairments,” in *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 2009, pp. 103–107.
[78] M. Valentí Soler *et al.*, “Social robots in advanced dementia,” *Front Aging Neurosci*, vol. 7, p. 133, 2015.
[79] M. Tanaka *et al.*, “Effect of a human-type communication robot on cognitive function in elderly women living alone,” *Medical science monitor: international medical journal of experimental and clinical research*, vol. 18, no. 9, p. CR550, 2012.
[80] K. Thodberg *et al.*, “Behavioral responses of nursing home residents to visits from a person with a dog, a robot seal or atoy cat,” *Anthrozoös*, vol. 29, no. 1, pp. 107–121, 2016.
[81] S. Petersen, S. Houston, H. Qin, C. Tague, and J. Studley, “The utilization of robotic pets in dementia care,” *Journal of alzheimer’s disease*, vol. 55, no. 2, pp. 569–574, 2017.
[82] A. Liang *et al.*, “A pilot randomized trial of a companion robot for people with dementia living in the community,” *J Am Med Dir Assoc*, vol. 18, no. 10, pp. 871–878, 2017.
[83] C. McGinn, E. Bourke, A. Murtagh, C. Donovan, and M. F. Cullinan, “Meeting Stevie: perceptions of a socially assistive robot by residents and staff in a long-term care facility,” in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 602–603.
[84] A. K. Pandey and R. Gelin, “A mass-produced sociable humanoid robot: Pepper: The first machine of its kind,” *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[85] T. Vandemeulebroucke, B. D. de Casterlé, and C. Gastmans, “The use of care robots in aged care: A systematic review of argument-based ethics literature,” *Arch Gerontol Geriatr*, vol. 74, pp. 15–25, 2018.
[86] G. Baldini, M. Botterman, R. Neisse, and M. Tallacchini, “Ethical design in the internet of things,” *Sci Eng Ethics*, vol. 24, no. 3, pp. 905–925, 2018.