Beyond Kant and Bentham: How Ethical Theories are being used in Artificial Moral Agents

John Zoshak
Human Centered Design and Engineering University of Washington, United States
jzoshak@uw.edu

Kristin Dew
Human Centered Design and Engineering University of Washington, United States
kndew@uw.edu

ABSTRACT

From robots to autonomous vehicles, the design of artificial moral agents (AMAs) is an area of growing interest to HCI, and rising concern for the implications of deploying moral machines raises questions about which ethical frameworks are being used in AMAs. We performed a literature review to identify and synthesize key themes in how ethical theories have been applied to AMAs. We reviewed 53 papers and performed a thematic analysis to describe and synthesize the current conversation across HCI. We found many describe the value of ethical theories and implement them as technical contributions, but very few conduct empirical studies in real settings. Furthermore, we found AMA development is dominated by two ethical theories: deontology and consequentialism. We argue that the focus on deontology and consequentialism risks creating AMAs based on a narrow set of Western ethical values and concepts at the expense of other forms of moral reasoning.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**;

KEYWORDS

Additional Keywords and Phrases Ethics, Philosophy, Artificial Moral Agents, Literature Survey, Deontology, Consequentialism, Virtue Ethics, Confucianism

ACM Reference Format:
John Zoshak and Kristin Dew. 2021. Beyond Kant and Bentham: How Ethical Theories are being used in Artificial Moral Agents. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3411764.3445102

1 INTRODUCTION

Autonomous machines are with us today – automated cars are on our roads, elder care robots are being tested in nursing homes, and complex financial transactions are conducted by sophisticated AI at a speed humans can't approach. Such emergent forms of autonomy come with questions of moral delegation and decision making: what kinds of morality do we want from machines driving on our roads, assisting in care of our family members, conducting financial transactions, or perhaps acting as sentient advisors in hybrid human-machine systems? There are numerous approaches to addressing this question, from professional ethics, to concerns about specific data ethics issues like transparency and privacy. One promising thread of research explores how we might work to implement normative ethical theories borrowed from philosophy in the form of machine agents; however, this line of work remains scattered across diverse HCI, engineering and ethics conversations, and remains a nascent conversation.

In this paper we survey and synthesize disparate threads of work across HCI to provide scholars and practitioners with an understanding of how normative ethical theories are being implemented in artificial moral agents (AMAs). Jose-Antonio Cervantes, et al. in their paper Artificial Moral Agents: A Survey of the Current Status [28] describe an AMA as a "virtual agent (software) or physical agent (robot) capable of engaging in moral behavior or at least of avoiding immoral behavior. This moral behavior may be based on ethical theories such as teleological ethics, deontology, and virtue ethics, but not necessarily." AMAs include the autonomous vehicles on our streets [73], the eldercare robots being tested in assisted living complexes [5], and even robotic weapons systems under development [93]. "Ethical theories" refers to normative ethical frameworks, which are defined as arriving at moral standards that regulate right and wrong conduct [51]. Generally, normative ethical theories are broken down into three strains: deontology, consequentialism, and virtue ethics [84]. However, this is an oversimplification and other philosophical theories hold normative principles, such as Confucianism [82], existentialism [90], and indigenous theories [66], which we discuss below. While our paper is concerned with research that explicitly references ethical theory(s), we are not arguing that in order to qualify as an AMA one *must* use an ethical theory.

"AMA" is a term that is not uniformly used throughout this disparate conversation, however we think it is a useful one for describing the range of entities discussed in the papers we surveyed. The papers in our corpus use many terms, including: "robot [105]," "artificial general intelligence [65]," a "system [9]," and "ethical autonomous agents [21]." We use the term AMA throughout this paper because the definition includes both software and physical robots, and the focus on "agency" as opposed to merely "intelligence" is especially salient for examining normative ethical theories as applied to AMAs because of normative ethics' focus on right conduct.

To better understand how HCI approaches building AMAs with normative ethical theories, we performed a literature search and qualitative review of works published on machine ethics in the ACM

CHI '21, May 08–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445102

Digital Library and IEEE Xplore. We built a corpus of papers that have a primary contribution of advancing understanding of ethical theories as applied to AMAs, then classified papers based on the ethical theory(s) they use. We found they took different approaches to working with ethical theories but were mostly limited to making analytical arguments and contributing technical implementation knowledge, with very limited empirical study of AMA development and deployment in real settings.

We found that the conversation is currently dominated by deontology and consequentialism, as well as theories that combine aspects of both such as Ross' Ethical Theory or the Doctrine of Double Effect. However, there exists nascent conversations leveraging other theories, and we highlight papers that explore Confucian [96, 99, 105, 106], existentialist [89], and indigenous ethics [65] as applied to AMAs. We argue the current dominance of deontology and consequentialism could lead to overly narrow development of rule and consequence-based AMAs, while neglecting the potential of other strains of moral philosophy for grappling with ethics cases. This dominance also has the potential to encode normative ethics into AMAs in a way that reinforces and scales Western philosophical hegemony over other forms of ethical reasoning. For example, Confucian ethics ask a different set of questions focused on societal roles, e.g. how might interacting with an AMA provide an opportunity for human moral development [96, 99, 105, 106]? We argue there is immense opportunity to explore ethical theories outside of deontology and consequentialism for articulating through AMAs. We also found a shortage of empirical work, which is critical to understanding how ethical decision making by AMAs unfolds in actual cases. Much work thus far has been dedicated to assessing the value of particular ethical theories or working on the difficult problem of implementing them into AMAs, but comparatively little on how these machines might act alongside humans in real-life instances. We argue there is potential for rich interdisciplinary work between philosophy and HCI, particularly through empirical work that can inform analysis and real-world implementation practices. For example, how might an AMA explain its actions based on a theory of care ethics? Could an AMA rebuke a human for falling into a trap of existentialist bad faith? There exists a significant opportunity to explore these questions outside of deontology and consequentialist frames.

2 RELATED WORK

2.1 Centering Values And Ethics In Sociotechnical Systems

The HCI community has long contributed to the design and implementation of systems aligned with human values, ethics, and priorities, from participatory design [24, 35] to value-sensitive design [39] to recent turns to critical theory that center more situated notions of justice [14, 16, 74]. Value-sensitive design in particular has been a foundational approach for the field to acknowledge and shape how ethical values and tensions may be embedded in and arise through sociotechnical systems [39, 40]. Though the distinctions between values and ethics have been debated for some time, HCI scholars and practitioners have demonstrated a shared commitment to understanding how diverse stakeholders' moral sensibilities are leveraged during the creation of – as well as distributed and emergent through – new sociotechnical systems [79].

The HCI community has repeatedly taken a central role in understanding how ethical decision-making can go on to impact interaction experiences in computing domains. Shilton, in surveying approaches to considering values and ethics HCI design processes, describes the field as being in a long established but recently invigorated "ethical turn," evidenced in part by a rapid rise in the number of papers mentioning the term "ethics" in the past two decades [79]. HCI scholars and practitioners have led the way in charting out ethical questions raised in a range of computing domains from data ethics concerns like privacy and autonomy [23, 36] to human-robot interaction concerns like how to negotiate ethical questions amongst distributed communicating agents [13].

Such work takes on new urgency in light of both documented and potential harmful impacts such as racially biased sentencing [15, 77] and autonomous vehicle deaths [73]. At stake in this conversation is how we address questions of ethical agency, risk, and responsibility in systems that may exceed designers', developers', and policymakers' abilities to understand them completely. For example, the distinction between explicit and implicit ethical agency shapes legal liability for decisions made and real harms accomplished through and alongside "moral machines" [34, 53].

While HCI work on research [71, 72], professional [59], pedagogical [61, 68], and data handling ethics [88, 104] has been robust, how people articulate and encode ethical theories into AMAs is a core question that HCI is well positioned to examine yet has been underexplored. In this paper we synthesize and extend this emergent conversation by drawing together the various formal approaches to building moral agents across the ACM and IEEE literature and closely adjacent HCI communities to highlight how a dominant focus on consequentialism and deontology constrains the construction of moral agents in Western rule and consequence-based logics, and how the preponderance of analytical argumentation and technical implementations highlights a need for other ways to build empirical understandings of AMAs.

2.2 Approaches to Encoding Ethical Reasoning In Artificial Moral Agents

Trying to develop new AMAs that align with human capacities and values requires both a closer look at the ethical theories being leveraged and encoded, as well as an examination of how ethical theories emerging from centuries of religious and philosophical debate might resist the materiality of encoding [33, 95]. The debate on how to approach the creation of moral agents is far from settled, and favored approaches tend to align with disciplinary norms and boundaries [79].

Allen, Smit and Wallach [3] characterize the main approaches to creating AMAs as top-down, bottom-up, and hybrid; top-down approaches rely on rule-based logics, abstract principles, and decision procedures to arrive at a desired action, and tend to align with philosophical traditions such as consequentialism and deontology. Bottom-up approaches provide an AMA with environments and instances to learn preferable behaviors and infer core values to arrive at moral decisions and actions, while hybrid approaches combine rule-based and bottom-up techniques [3]. Approaches also diverge in terms of whether the system has implicit or explicit ethical agency, and to what extent they model and take into account the actions of other ethical agents in situ [30]; each brings challenges in articulating and implementing a morality model that can respond in messy, real-life scenarios and be a meaningful augmentation of human judgment [28, 30].

In their recent review of the technical models and implementation techniques used in creating AMAs, Cervantes et al. found that general intelligence systems that could replace human judgment are far from fruition, but that a range of prototypes function in controlled settings on limited, well bounded cases [28]. These authors propose a robust taxonomy of AMAs by strategies (top-down, bottom-up, hybrid), implementation criteria, and whether or not their ethical agency is explicit (machines programmed to act ethically by design) or implicit (machines programmed to reflect on their actions and resist acting on or violating ethical expectations [34]) to support technical implementation decisions [28]. However, scholarship from the CHI community has revealed that even once built, there are significant limitations to the benchmarking and testing strategies employed in determining an AMA is "ethical" due to an overreliance on contrived ethical dilemmas [69].

Meanwhile, critical scholars have pushed back against the idea that machines can or should be effectively imbued with moral decision-making capabilities due to the interpretive flexibility [20] of computational systems and their attendant abstractions [55, 100]. Researchers and practitioners alike might instead focus on empirical studies of practice to uncover value tensions and levers to better understand how ethics gets worked out in situ [80, 96]. Studies of practice also provide an opening to situate machine ethics within specific historical conditions of dis/empowerment. This turns the debate away from universal moral frameworks and leverages instead situated notions of "justice" that reframe AMAs as the means through which various moral actors negotiate political and ethical outcomes; in this view, accomplishing abstract values like "fairness" in an ethical machine matters little if the machine's purpose is used for structurally discriminatory purposes [16, 57].

To gain traction into such debates between top-down and bottom-up, implicit and explicit, and universalist and situated approaches to encoding ethical reasoning in AMAs, we see a need to step back to examine the moral theories and philosophical traditions that have been leveraged thus far in building AMAs. HCI has long borrowed from other intellectual traditions in the humanities, including critical race theory – [74], feminist theory [14], postcoloniality [52], and disability studies [16] – to clarify and catalyze alternative approaches to developing sociotechnical systems. In the pages that follow, we seek to identify not only what moral philosophical frameworks are in use but synthesize what they offer AMAs as ethical systems. As such, our literature review sheds light on strengths, gaps and opportunities for HCI scholars and practitioners to devise alternative approaches to developing ethically informed agents.

3 METHODS

Our work began by constructing a corpus of papers on machine ethics by searching the ACM Guide to Computing Literature using the terms "machine ethics." We started with ACM literature because it contains a range of nascent HCI conversations on this topic that cross-cuts subdisciplines, and because it covers interdisciplinary work as well as engineering papers. Because our interest was in papers that explicitly use ethical theories in their contribution to understanding and implementing AMAs, we kept our search term broader than our research interest to avoid missing relevant articles that may use other terminology, particularly as field terminology changes quickly and relevant work can appear under numerous terms and subdisciplines of HCI; we quickly realized this conversation is emergent and far from consolidated. This resulted in 54 papers of which 11 were included in our corpus. To ensure we were covering the whole conversation we broadened our search to any paper using the keyword "ethics" from 2014 onward. This added an additional 9 papers. Finally, we manually reviewed the references of included works and added another 29 papers. To further ensure we had adequately captured the conversation across disciplines, we also performed a search in IEEE for "machine ethics." This search returned an additional two papers, and we pulled two more manually from their citations for a total of four additional papers. Our final corpus comprised 53 papers from 1995-2020.

Because our interest is to understand how ethical theories are being used and incorporated into AMAs we included papers based on the following criteria: 1) The primary contribution analyzed, described, or otherwise advanced our understanding of one or more specified normative ethical theories in the context of AMAs; or 2) the primary contribution explored implementing one or more specified normative ethical theories in the context of AMAs. We excluded papers that referred to ethical theories generally, that merely mentioned a non-specific ethical theory in their proposed implementation, or that examined general data ethics values like "fairness", "transparency," and so on.

Drawing on DiSalvo et al., [32] we then developed a coding rubric based on our research questions, which we refined as we read through each paper's key claims, methods, and contribution statements. This process resulted in research sub-questions to help describe and synthesize the corpus as a conversation characterized by their contributions, methodologies, and ethical frameworks. Specifically, we asked: 1) What ethical theory did the researchers use? And 2) What are the authors doing with the ethical theory? More specifically, what is the primary contribution of the paper? Was the paper an analytical argument, an empirical study, an implementation case, or another methodology? The latter classification schema emerged thematically when examining question 1 above.

For each paper in our corpus we wrote answers to the questions our rubric asked. As we conducted our analysis we discussed repeated themes and issues, gaps we noticed, and areas where there seemed to be both disagreement and agreement. Based on our review we identified three key themes for discussion: 1) The dominance of deontology and consequentialism in the conversation; 2) Other philosophical theories receive significantly less attention, but could prove insightful to the conversation; and 3) Paper contributions primarily take the form of analytical arguments and implementation proposals, but few are empirical examinations of AMAs in real-life settings and practices.

4 FINDINGS

To find out what specific approaches are driving work on AMAs and ethical theory, we reviewed each paper to denote which ethical framework or theory was leveraged. Our main finding is that deontology and consequentialism, and related theories, are currently dominating the conversation; 46/53 papers reference theories that fall under this umbrella. In the sections that follow, we examine the ethical theories in use before turning our attention to how they are being used.

4.1 Deontology, Consequentialism and Related Theories

Deontology and consequentialism make up two of the three dominant paradigms for normative ethics in Western moral philosophy [84], and this dominance was apparent in our survey. Deontology is a type of ethics that is based on rules (e.g. "it is unethical to steal") [86]. Deontology is closely associated with Immanuel Kant and his formulations of the categorical imperative [56]. Kant had several formulations of the categorical imperative but perhaps the most well-known is the first which states that you are to "act only in accordance with that maxim through which you can at the same time will that it become a universal law [85]." Consequentialism stands in contrast to deontology because it focuses on the outcomes of behavior [81] (e.g. it is ethical to steal a loaf of bread if the store owner can absorb the cost with little trouble and the thief avoids starving to death). One of the most prevalent branches of consequentialism is utilitarianism, which has its 19th century origins in the thinking of Jeremy Bentham and John Stuart Mill [17, 67]. Classic utilitarianism is usually summarized as "an act is morally right if and only if that act maximizes the good, that is, if and only if the total amount of good for all minus the total amount of bad for all is greater than this net amount for any incompatible act available to the agent on that occasion [81]." We've included papers that use Ross' ethical theory [87] (RET) (a theory in part created in response to Kant's deontology and Bentham's utilitarianism) and the doctrine of double effect (DDE) [83] (a theory rooted in Catholic theology that could be described deontological because of its similarities to rational agent-centered deontology[1]) as part of the deontology and consequentialist threads because they foreground both rules and consequences in their formulations.

4.1.1 Deontology. 18 papers in our corpus referenced deontology [2, 4, 19, 26, 29, 31, 42, 47–49, 54, 75, 76, 90, 94, 98, 100, 101], 11 were implementation papers [19, 26, 29, 31, 49, 54, 75, 76, 98, 100, 101], six were analytical argument papers [2, 4, 42, 47, 48, 90], and one was an empirical paper [94]. The deontological papers showed a few common themes and tensions. Because deontology is focused on rules, one of the shared problems frequently highlighted is what happens when the rules conflict [2, 4, 42, 48, 76]? As Powers explains [76], the categorical imperative, by itself, is insufficient, "Obviously, the simple account of mere consistency won't do. It must be buttressed by adding other facts, principles, or maxims, in comparison with which the machine can test the target maxim for contradiction."

There are a few approaches aimed at accounting for more than just the categorical imperative. One paper [49] adds to the formalization of the categorical imperative by acknowledging a consideration of autonomy to their formulation. The formalization specifically accounts for situations where informed consent to violate autonomy has been given, or when violating autonomy norms is appropriate given the current situation. For example, "...suppose there is a car coming, I shout a warning that you cannot hear, and I then pull you out of the path of the car. I prevent you from crossing the street, but there is no violation of autonomy, because my action does not interfere with your action[2]." They also add an element of consequentialism to their formalization, echoing the combination found in RET. Alternatively, another paper [101] argues that the only agents suited to resolving conflicts between duties and resolving ethical dilemmas are public bodies to which the public has given consent. That is, the focus should be on *rightful* machines not *ethical* machines – machines that follow the agreed upon laws – and that deontology is suited towards programming rightful machines.

Another theme in the deontological papers was that deontology lends itself well to being programmed. Although this is seen as a positive attribute in some papers [49, 98, 101], Tonkens [90] argues that although we *could* implement Kantian ethics into AMAs it does not follow that *we should.* Tonkens argues the creation of a Kantian AMA is contradictory and should not be pursued. In this view, because AMAs are *programmed* to be ethical they themselves cannot be said to be moral agents. "[A]ccording to Kant part of being a moral agent means 'the capacity to master one's inclinations when they rebel against the [moral] law,' hence the ability to freely commit actions that are not moral." This view directly pushes against the notion that building Kantian AMAs is possible because, in a Kantian view, ethical agency is asymmetrical [62], i.e. it is a capacity limited to humans because we are rational and free. However, the paper leaves open the possibility of other ethical theories being suitable for implementation into AMAs.

4.1.2 Consequentialism. 18 papers in the corpus directly referenced consequentialism [2, 4, 11, 25, 26, 29, 31, 41, 42, 46, 47, 50, 58, 59, 63, 64, 75, 94]; nine were implementation studies [11, 26, 29, 31, 41, 58, 63, 64, 75], seven were analytical arguments and essays [2, 4, 42, 46, 47, 50, 59], and two were empirical studies. [25, 94]. The major theme across all consequentialist papers was the question of how to calculate utility.[3] One approach [58], in the context of autonomous cars, was to use data from the Moral Machine project, a project crowdsourcing data from various iterations of the Trolley Problem, and assigning value to human lives that way. The authors then used this data for their utilitarian model. This was the only consequentialism implementation paper to try basing the utility calculus on empirical data.

[1] Agent-centered deontology holds that risking/causing evil is distinct from any intention to achieve it. This matches neatly with the Doctrine of Double Effect's second principle: "The agent may not positively will the bad effect but may permit it. If he could attain the good effect without the bad effect he should do so. The bad effect is sometimes said to be indirectly voluntary." However, the Doctrine of Double Effect also explicitly takes into account consequences in its fourth principle: "The good effect must be sufficiently desirable to compensate for the allowing of the bad effect." It is in both these senses that we include the Doctrine of Double Effect in the deontology and consequentialism threads. [83]

[2] In this use of Kantian theory, actions are rational, and getting hit by a car would frustrate the desired action.

[3] Utility calculation is generally referring to act consequentialism which is the claim that "an act is morally right if and only if that act maximizes the good, that is, if and only if the total amount of good for all minus the total amount of bad for all is greater than this net amount for any incompatible act available to the agent on that occasion." [81]

However, one paper [50] warned of the risk of seemingly "neutral" utilitarian criteria to reinforce existing biases (a criticism echoed in the existentialist paper discussed below) and suggested augmenting utilitarian calculus with rational choice theory from economics. Using hiring algorithms as an example, it argues that "...static utility-based conception of optimal hiring, wherein algorithms predict and hire the "good" workers out of a candidate pool, is ill-suited for understanding the dynamics of complex social processes and as a result, the societal obligations to which AI tools may be bound." Instead, utilitarian calculus might be augmented by looking more holistically at the dynamics of the labor market system including attributes such as "... workers' investment opportunities prior to entering the labor market, to their tenure within the market as they interact with various firms and cycle through different jobs." This work shows that the field is going to have to continue to look for ways to solve utility calculus problems without reinforcing existing biases and power structures through AMAs.

4.1.3 Doctrine of Double Effect & Ross' Ethical Theory. Eight papers in our corpus reference the DDE [18, 22, 26, 44, 45, 63, 64, 70]. The DDE comes from Catholic theology and states that for any action that has negative side-effects, in order for the action to be ethical: 1) the action is morally neutral[4]; 2) the net good consequences outweigh the bad consequences by a large amount; and 3) some of the good consequences are intended and none of the bad ones are [44]. The DDE papers exclusively address implementation, and a common argument for using DDE, shared among several papers, is that the DDE seems to be intuitively grasped by many lay people. [18, 44, 45]. One paper [45] even argues that DDE "rises above" philosophical debates about which theory is preferred because "empirical studies have found that DDE plays a prominent role in an ordinary person's ethical decisions and judgments" and that DDE "plays a central role in many legal systems." This echoes the argument above [101] that focusing on *rightful* machines cuts the Gordian Knot of ethical debate by using public institutions to decide what is right.

Eight papers reference Ross' ethical theory (RET) [5–11, 78]. RET is the formulation of several duties that mixes both deontological and consequentialist aspects [10]. It is deontological because some duties, like fidelity (honor promises) and gratitude (return favors) refer to rules someone should follow; it is consequentialist because some duties like nonmaleficence (act to cause the least harm) and beneficence (act to bring about the most good) take into account consequences [10]. Seven of the papers are a series of work led by AMA researchers Michael Anderson and Susan Leigh Anderson which lays the foundations for applying RET to AMAs and iterates over several implementations [5–11]. They first argue that RET is superior to deontology and consequentialism because it combines both, and "...single-principle, absolute-duty ethical theories. . . are unacceptable because they don't appreciate the complexity of ethical decision making and the tensions that arise from different ethical obligations pulling us in different directions [10]." They then follow this insight through several implementation focused papers for AMAs for biomedical ethical advice [10], eldercare [5, 7–9], and a general ethical dilemma analyzer [6]. This line of research culminates in a RET influenced program that is ready for implementation in a Nao robot[5] for testing in eldercare settings [5]. For example, the proposed robot can remind a patient to take their medicine while following the prima facie duties[6] inspired by RET. The other paper that used RET proposed an "ethical casino" implementation in the online video game Second Life and partially based their implementation on the prior work of Anderson and Anderson [78].

4.2 Ethical Traditions Beyond Deontology and Consequentialism

Only ten papers out of 53 in our corpus explored an ethical theory outside of the deontology and consequentialist universe [4, 42, 43, 60, 65, 89, 96, 99, 105, 106]. Four of them [96, 99, 105, 106] were put out by a group of authors exploring Confucian role-based ethics for use in machine ethics; four [4, 42, 43, 60] referenced Virtue ethics with one implementation paper [43]; and three were analytical arguments [4, 42, 60], though in these analytical arguments virtue ethics was examined alongside deontology and consequentialism. One paper explored existentialism [89] and another explored Lakota and Hawaiian approaches to ethics [65]. These papers also share a common theme of pointing out the prevailing deontology and consequentialist dominance of the conversation.

4.2.1 Confucianism. The authors of these papers conceive of Confucianism as a kind of role-based ethics with the ultimate goal of living well within our social roles and practicing the moral responsibilities prescribed by these social roles. Confucianism puts a greater emphasis on roles and relationships, and is therefore more interested in exploring questions related to the roles AMAs will play in our society, as opposed to deontology and consequentialism. This sentiment is best expressed by authors Zhu et al. [106]: "It is the interaction or relationship between robots and their human partners that makes the existence of these robots. In this sense, we suggest that roboticists should not only leverage the traditional, dominant approaches to developing AMAs that focus on integrating rule-based morality, but also consider an alternative approach to designing morally competent robots based on the role responsibilities prescribed by the relationships robots have with human teammates in specific use contexts."

They also note that Confucian ethics open new fields of inquiry in machine ethics by asking questions like "what kind of person is the human teammate becoming through interaction with the robot?" [96, 105] a question deontology and consequentialism, with a focus on rules and moral calculus, are not as well equipped to ask. This focus on the impact of AMAs on their human partners is especially highlighted in Wen et al.'s work [96] on moral rebukes, where they studied how rebukes to unethical commands affected their human partners. Even though this paper found deontology

[4] This is also sometimes formulated as "the act itself must be morally good or at least indifferent." [83]

[5] A type of commercially available programmable robot. See: https://www.softbankrobotics.com/emea/en/nao

[6] The authors provide a situation the robot is prepared for: it notices a patient is immobile while reminding them of their medications. "Although warn and notify both satisfy the duty of maximizing prevention of immobility, they also violate the duty of maximizing respect for autonomy. In the current situation, the principle deems this violation sufficient to not choose either action. . . In this case, the principle determines that honoring commitments supersedes preventing immobility and continues to perform the remind action." [5]

was slightly better at generating reflection in the human partner; Confucianism's focus on relationships prompted the authors to ask this question in the first place. That is, studying different ethical theories and AMAs may generate entirely different research questions. In this case, Confucianism's focus on roles and relationships prompted the authors to ask a question that may be overlooked in a deontology and/or consequentialist paradigm. A line of inquiry based in Confucianism expanded the field of inquiry.

4.2.2 Virtue Ethics. Virtue ethics was mentioned by four of our papers [4, 42, 43, 60], including the only implementation paper [43] that referenced a theory other than deontology or consequentialism. Virtue ethics traces its origins to Plato and Aristotle and turns its analysis away from moral action and outcome and towards the embodied actor and their virtue or moral character [84]. Virtue ethics also focuses on role models or exemplars to instill virtues in others [43]. This shift of focus to the actor's moral development process (whether human, machine, or human-machine assemblage) can be contrasted with deontology and consequentialism's focus on rules and consequences – moral inputs and outputs – and could prove similarly generative for asking questions about how AMAs interact with society and morally develop.

Virtue ethics was the only ethical theory outside of deontology and consequentialism to be included in an implementation-focused paper. Govindarajulu et al., [43] in addition to offering an implementation pathway for virtue ethics, explore why virtue ethics may be preferable to deontology or consequentialism. Drawing from other work they explain virtues have four conditions: stability, consistency, explanatory power, and predictive power.[7] They argue that these qualities could make a virtue ethics-driven AMA more comprehensible to humans compared with deontological or consequentialist AMAs.

On the other hand, one paper [60], in contrast to most of the papers in our corpus, sees virtue ethics, deontology, and consequentialism ideally working in concert, rather than directly contrasting them with each other. Deontology could provide the basic ruleset, consequentialism could provide a fast feedback loop with its focus on the consequences of actions, and virtue ethics could provide a longer feedback loop for moral agent development [60]. Virtue ethics performs a similar function in philosophy [84], turning the lens of inquiry from behavior to the actor, and this focus could prove useful to the field and remains underexplored. For example, what would a virtue-ethics based explanation of behavior look like? Would humans find such an explanation sufficient?

4.2.3 Indigenous Ethics. One paper [65] explored indigenous perspectives on ethics, analyzing Lakota and Hawaiian philosophies and using them as a lens to examine AMAs. This paper had similar themes to the Confucian papers, for example the Hawaiian concept of pono as an ethical approach which has "no scope [to be reduced] to prioritize the individual over a *relationship*, particularly engaging in those different from ourselves" [emphasis ours]. Echoes of the Confucian emphasis on relationships can also be seen in Lakota culture. The paper uses the Lakotan ontology of everything in the universe having a soul to explain how "situational animism" might apply to AMAs. It specifically proposes something like a Lakotan stone ceremony[8] at particular stages of development to "observe and catalogue how a particular AI interacts or connects with humans and its immediate environment." From that we could "begin to construct relational frameworks to protect and empower humans." The emphasis on relationships with non-human entities is something that deontology and consequentialism do not provide in their account of ethics. Lakota and Hawaiian ethics offer another lens for looking at the relationships between humans and AMAs, and again this area is underexplored relative to the questions that deontology and consequentialism ask.

4.2.4 Existentialism. One paper [89] focused on existentialism. The paper steps back from a narrow scope on AMAs in specific situations and instead argues we should be focused on the people and companies who build AMAs through the existentialist lens of "bad faith". Bad faith "...refers to the individuals, be it industry leaders, lobbyists, or developers, who hide themselves behind the curtains of the others' expectations or the disguise of the corporate environment in order to evade the anguish and the risk of realizing the burden of responsibility that they pragmatically shoulder." One can find examples of bad faith in vague justifications such as "we did this for business reasons" since that offsets the responsibility from specific decision-makers onto an abstract collective concept. Existentialism has been criticized for not having a normative take on ethics. Without a view on normative ethics existentialism may not have much to offer to the field of ethical theories and AMAs. However, this paper takes on that criticism by citing French philosopher and feminist scholar Simone de Beauvoir and arguing the normative impulse in existentialist ethics is focused on maximizing the freedom of others and helping them break out of existentialist bad faith. "The designer, the CEO, the lobbyist or the activist have one thing through which to filtrate their available options when confronting ethical decisions: which option will eventually magnify the freedom of others affected by my choice?" Existentialism provides a lens through which to see the entire AMA ecosystem. That is, existentialism shines a light not just on AMA behavior, but on the decisions that lead to the creation of AMAs themselves. Deontology and consequentialism, by their nature of focusing on rules and outcomes ignore these larger issues.

While deontology and consequentialism certainly have value to provide to the field, this dominance may be leading to blind spots. As noted, Confucianism and Lakota and Hawaiian theories offer a focus on relationships that deontology and consequentialism do not consider. Existentialism also shows promise for its potential to examine the higher-level systems and structures that deontology and consequentialism, with their narrower focus on actions, may miss.

[7] Virtues are stable because when someone possesses a virtue they will retain that virtue. Virtues are consistent because if someone possesses a virtue that is sensitive to reason r then they will respond to r in most contexts. Virtues have explanatory power because referring to their virtue will sometimes help explain the actor's behavior. Finally, Virtues have predictive power in that if someone has a "high fidelity" virtue it will enable near perfect predictive power of their behavior and even low-fidelity virtue will enable weak predictions of their behavior. [43].

[8] This draws on the Lakota concept of "wakan" or anything that could not be understood. Even stones could have wakan if they did something unusual such as locomotion. These stones were then accorded status and special ceremonies were held for establishing a relationship and beginning a relational process of exchange which developed into trust and reciprocity. Crucially the stone was only considered animate when it was closely bound to the life of a person. [65]

4.3 How are Ethical Theories Being Used in AMAs? Technical Implementations, Analytical Arguments, and Empirical Investigations

To find out how ethical theories are being used to build AMAs, we reviewed the included works' contributions and predominant methodological approaches. We found three core approaches to using specific ethical theories for building AMAs: 1) Technical implementations, meaning the paper is contributing to our engineering knowledge of how an ethical theory could be implemented through particular software and hardware configurations; 2) Analytical arguments, meaning the paper is contributing to our understand of how an ethical theory might be used in AMAs and the issues it brings up; and 3) Empirical studies, the paper is an empirical study of how the AMA works in the real world. 15 papers were focused on analysis [2, 4, 42, 46–48, 50, 59, 60, 65, 89, 90, 99, 105, 106], 35 papers were focused on implementation [5–12, 18, 19, 21, 22, 26, 27, 29–31, 34, 41, 43–45, 49, 54, 58, 63, 64, 70, 75, 76, 78, 98, 100, 101, 103], and only four papers were focused on empirical work [6, 25, 94, 96]. We describe and unpack each approach below.

4.3.1 Technical Implementation. 35/53 papers are focused on programming ethical theories in AMAs. They publish examples of logic and/or methods for implementing one or more ethical theories in an AMA. Many of these papers offer formalizations that could leverage either or both deontological or consequentialist theories [12, 27, 30, 49, 103]. As one paper [49] notes: "Utilitarianism is normally conceived of as a consequentialist theory but it can be formulated deontologically as well. . .the principle can require that an agent take actions that it can rationally believe maximize what it regards as utility."

The trolley problem (or slightly modified versions of it) features in many implementation papers [22, 26, 44, 45, 58, 70, 75] as it is a well-known philosophical dilemma. These papers are using the trolley problem to validate their implementation of a particular ethical theory. However, one paper [101] criticized the use of the trolley problem as "there will likely never be an ethical consensus as to their correct resolution, and even if one could be achieved, it would be largely irrelevant to the task at hand." The paper argues it would be irrelevant because, again, the focus should be on what is *rightful.* "...[I]t is public law that should determine when makers or users of semi-autonomous machines such as self-driving cars are liable or culpable for the machine's decisions, and law must conform to principles of justice, not the partial ethical preferences of one group or another."

Only one implementation focused paper worked on an ethical theory outside of the deontology and consequentialism universe, and that was Govindarajulu et al.'s work on virtue ethics [43]. This means that in our corpus there are no implementation papers that focus on non-Western theories of ethics; the current conversation around implementing ethical theories into AMAs is completely dominated by Western ethical theories.

4.3.2 Analytical Argument Work. 15/53 papers offer analytical arguments, meaning these papers are taking an ethical theory and either examining it to ask questions about how it might be used in creating AMAs, identifying strengths the theory provides, or pointing out its weaknesses. Papers outside of deontology and consequentialism generally took the form of introducing the ethical theory, and explaining some of the advantages of the theory in contrast to deontology and consequentialism. For example, when introducing Confucian philosophy [99], Williams et al. explicitly note both that: "[Confucian ethics] is uniquely well suited to robotics due to its focus on adherence to hierarchically structured relational roles (which will necessarily govern how robots will fit into their unique socio-technical niche within human society); and "a discussion of Confucian ethics may be uniquely informative for the HRI community due to the countervailing focus in the community on Western ethical theories, especially norm-based theories such as deontology." This is especially interesting in light of Tonken's argument that Kantian machines are a contradiction (discussed above) [90]. Confucian ethics may be suited to answering the challenge offered by Tonkens; in a Confucian world, AMAs don't necessarily need free will to be seen as moral agents, since they will be moral by fitting in their unique socio-technical niche.

Analytical argument papers in deontology and consequentialism largely pointed to advantages of deontology and consequentialism or identified flaws. For example, in a criticism of utilitarianism, one paper [46] argues that utilitarianism doesn't respect individual autonomy: "We can reject utilitarianism not on the grounds that it requires too much of an artificial agent but rather that it ignores the individual identity and rights of the human subject affected by the agent." In other words, people aren't interchangeable and pure utilitarianism fails to acknowledge that. Alternatively, after pointing out flaws in pure deontological and consequentialist approaches, Greene et al. [47] hypothesize that we can use insights from evolutionary psychology to make a more useful combined deontology, consequentialism, and virtue ethics framework "...the brain seems to make both types of judgment (deontological and consequentialist) and then makes a higher order judgment about which lower-order judgment to trust, which may be viewed as a kind of wisdom (reflecting virtue or good character). In contrast to the philosophies outside of deontology and consequentialism these analytical argument papers rarely mention other ethical theories. The deontology and consequentialism universe seems unaware of its own dominance.

4.3.3 Empirical Work. Only 4/53 papers conducted studies on how AMAs might act as ethical agents in real-world settings alongside humans. Wen et al. used Confucianism to ask a unique question about AMAs and generated insights into how deontology and Confucianism may provoke reflection in their human partners [96]. Wachter and Lindner and Bonnefon et al. both explored questions around whether consumers may ever accept utilitarian autonomous vehicles [25, 94]. Finally, Anderson and Anderson were able to provide evidence that an approach based on RET was able to generate satisfactory resolutions to ethical issues in a way that was accepted by professional ethicists the vast majority of the time [6]. Even if these studies were repeated for different ethical theories, the field would benefit from understanding moral encounters with real people and practices.

Specifically, Wen et al. compared participants' reactions to rebukes [96] for unethical commands. One rebuke used Confucian ethics[9] as a framework and the other used deontology. The paper found that both styles of rebuke were effective, but counter to the expectations of the authors, the deontological approach provoked more self-reflection from their participants; the authors further hypothesized that this may be due to deontology being a more familiar concept to their Western participants. However, this study highlights that using a Confucian lens prompted a unique research question from the authors. Given the potential for AMAs to have to react to problematic commands, studying rebukes is an important line of inquiry that Confucianism's focus on roles and relationships unlocked; there is a rich opportunity for further research in rebukes as a tactic for inviting ongoing human reviews, resistances, and correctives to AMAs.

Meanwhile, Wächter and Lindner [94], using a combined deontology and utilitarian framework, analyzed blame attribution when an AMA makes decisions based on the trolley problem. The paper found that utilitarian decisions by a robot were ascribed more blame over several dilemmas when compared to deontologically based decisions. This suggests that even if we were able to implement utilitarianism into an AMA a large segment of society may not accept its justifications. Similarly, Bonnefon et al. [25] found comfort with the concept of utilitarian autonomous vehicles, though this finding was somewhat tempered by hesitation in participants to buy vehicles that would sacrifice the owner to save lives: Would anyone willingly buy a self-sacrificing utilitarian autonomous vehicle?

Finally, Anderson and Anderson [6] tested their ethical advice generator with what the authors referred to as their "ethical Turing test" on a panel of ethicists. The authors found out that the ethicists were in agreement with their ethical advice generator in the vast majority of cases, showing potential utility for this type of AMA in the field.

Looking across the different types of papers, we note opportunities for empirical work to inform implementation studies in particular. Analytical argument papers generally argued for the advantages and disadvantages of particular theories, and implementation papers worked to articulate and materialize these theories in AMAs; meanwhile empirical studies have an underexplored role to play in these conversations because they could add nuance and direction to AMA research agenda. For example, many implementation and analytical argument papers explore utilitarianism, but Wächter and Lindner's study suggests that society may not accept pure-utilitarian decisions by an AMA [94]. Empirical studies can point the way to what approaches work in real settings where ethical decisions and practices unfold, and guide further analysis and implementation work.

5 DISCUSSION

5.1 Moving Beyond Deontology & Consequentialism

Deontology and consequentialism are normative ethical theories that have existed for centuries and have demonstrated value in the overall discussion of analyzing, studying, and implementing ethical theories into AMAs. Specifically, they were the first theories to be implemented into code, and have shown the way for the formalization of other ethical theories [43]. However, the near-exclusive focus on deontology, consequentialism, and related theories could lead to significant limitations in how the field approaches the creation of AMAs.

The existentialist critique [89] points out that the focus deontology and consequentialism have on rules and outcomes respectively could be obscuring a broader focus on the ethical values of who is creating AMAs and the economic systems that are producing them. Having the lens of existentialist bad faith could lead to more ethical decision-making processes in system design; one can imagine professional ethics codes calling for colleagues to hold each other accountable when bad-faith reasons are used to drive the day-to-day work of building AMAs (e.g. "we have to do this for business reasons" or "we must meet this deadline"). This type of analysis operates at a completely different level than the focus of the deontology and consequentialist papers surveyed, from a narrow focus on the machine itself, to the individuals, organizations, communities and institutional structures through which AMAs are instantiated. However, we think even holding to the narrower focus of AMA behavior, existentialism could provide insight into responses to unethical commands. To the extent that the normative aspect of existentialism focuses on freeing others from existentialist bad faith, one might configure AMAs to use existentialist rebukes in their daily work alongside humans. Empirical work that explores existentialist rebukes to unethical commands, similar to the Confucian rebukes explored by Zhu et al., [105] may be a fruitful next step for exploring existentialism and AMAs.

The use of deontology and consequentialism may also be crowding out a more robust examination of the relationships between humans and ethically informed machines that HCI is well situated to address with questions such as "What does a good machine/human relationship look like?" and "What new forms of personhood and ethical decision making might emerge through interacting with this machine?" Confucian, Lakota, and Hawaiian ethics [65, 96, 99, 105, 106] have centered these types of questions and remain underexplored in the field. Feminist care ethics, an approach emergent in HCI, but not represented in our corpus[10], seems like another ripe area for exploration given its similarly situated and relational focus on acting in ways that "maintain, continue, and repair our world [37]." These theories provide a way forward for imagining how AMAs would integrate (or not) into broader networks of ethical actors, focusing not just on the specific moral tasks we ask them to perform, but as to their larger *role* in ethical decision-making processes distributed across a range of actors over time. Such questions cannot be fully addressed using solely deontology and consequentialism as guiding theories, and align with growing calls for transparency and justice in the development of technical systems [38].

Even if we focus narrowly on Western philosophical traditions, virtue ethics warrants further examination due to its focus on the development of moral character over many instances. Even the one paper [43] exploring implementing virtue ethics revealed alignment with consequentialism and deontology when the authors argued: "it has not been clear that there exists a version of virtue ethics

[9] A Confucian rebuke would sound like "I'm not going to punch Sam because they are my friend, and good friends don't punch each other." Whereas a deontological rebuke would be "I'm not punching Sam because punching is wrong." [96]

[10] However, care ethics has been leveraged as a lens in other areas of HCI, see [91].

rigorous enough to be a target for machine ethics. . ..” One could similarly argue that virtue ethics *is* rigorous and rather it is ethics implementation models and techniques that currently lack the rigor and flexibility to implement any ethical theory beyond deontology and consequentialism.

Leveraging virtue ethics could bring a new focus to AMA development by shifting it away from the specific actions undertaken by humans or machines, and instead turning it around to the agent and their moral development. Kuipers [60] already foreshadows a potential benefit of virtue ethics to the field by making the argument that virtue ethics could serve as a long-term feedback loop for AMAs. However, we also think, given virtue ethics' focus on role models, there is an interesting opportunity for more empirical study. One could imagine, using a combined virtue ethics/Confucian lens, a study that looks at virtue-based justifications an AMA gives for behaviors and examines the self-reflection that prompts in participants. Could AMAs serve as effective virtue role-models? These are the types of questions one could ask using a virtue ethics framework, that can't be asked in an exclusively deontology and consequentialism world.

5.1.1 Fit to Purpose Over Programmability. A handful of papers [49, 90, 98, 101] explicitly discuss how deontology and consequentialism are amenable to being programmed into AMAs. While ease of implementation is certainly an advantage of deontology and consequentialism, alignment with current computing practices and materialities should not be the main criteria for ethical explorations with AMAs. Machines are certainly well equipped to follow rules and conduct utility calculus; however, it doesn't necessarily follow that implementation concerns should be driving which ethical theories are explored. Ultimately, we agree with the argument put forward by Tonkens [90], the standard for ethical theories' implementation should not be “is it easy?”

It is important that we not look at moral philosophies outside of deontology and consequentialism as “less rigorous” [43] or worth exploring simply because it is not clear how to implement them. Rather, the field has an opportunity to examine other philosophical traditions and become more adept in articulating and implementing the appropriate logics and principles not just *in* but *through* AMAs. If those who research and build AMAs default to using only ethical theory that is most amenable to programming over what is appropriate to the setting, people and problems at hand, we risk missing out on valuable perspectives and insights, such as those outlined by Confucian, indigenous, care and existentialist ethical theories, and may introduce only a narrow subset of Western ethical reasoning into AMAs as a result.

5.1.2 Making Space for non-Western Ethical Theories. The reliance on deontology and consequentialism highlights a need for other kinds of implementation, particularly for non-Western ethical frameworks. This is a serious gap that needs to be closed if AMAs are going to benefit from and fit within different ethical traditions. The stakes are high for this emergent field of inquiry and implementation, as ethical theories are culturally and historically situated. AMAs have the potential to materialize, circulate and scale the ethical logics and scripts they embody [1, 62]. These scripts have power to shape and constrain human behavior in ways that we may not be aware of, and the script creators (in our case AMA developers and designers) wield that power. Adding a post-colonial lens to this argument further raises the stakes. As Irani et al. note [52] “Colonial relationships may have dissolved, and yet the history of global dynamics of power, wealth, economic strength, and political influence shape contemporary cultural encounters.” In the case of AMAs, there is a risk that developers and companies may inadvertently encode Western ethical practices into AMAs and circulate them through market relationships around the world. HCI is well equipped to interrogate such power to shape ethical action and ensure that a greater diversity of ethical theories are informing AMA development.

Furthermore, in pursuing implementations of only deontology and consequentialism, our moral focus narrows and looks at the individual actions and consequences of human and machine agents. By incorporating Confucian, Lakota, and Hawaiian ethics [65, 96, 99, 105, 106] we gain other lenses for looking at the broader societal ties and relationships amongst humans and AMAs. With existentialist ethics we turn our gaze to the systems and structures that produce AMAs, and ask how to transcend existentialist bad faith. With virtue ethics, the focus shifts to the moral development of the agents themselves, and may gain long term feedback loops that lead to more reciprocal and ongoing engagements between humans and AMAs. Care ethics, not found in our corpus, but emergent in HCI, may open up questions on how AMAs relate to others through situated responsibilities, or how AMAs may perform care by identifying and addressing neglected sites, experiences, and objects. We wonder, what might other traditions of moral philosophy and ethical reasoning not addressed here offer our ability to enact ethical forms of decision-making with machines?

5.2 Opportunities for Empirical Examinations

HCI is a field that is well positioned to conduct human-centered empirical studies on how AMAs may be actors in ethical decision-making processes, bridging both technical and philosophical conversations and developing more robust ethical theory through empirical investigations. For example, empirical studies could lead the way as to which ethical theories are fit to implement in certain settings by better aligning them with real-life decision-making practices and in situ moral judgement processes. For example, several papers are focused on the complexity of calculating utility, however one empirical paper suggests [94] that humans may not accept a purely utilitarian explanation for a given action. In light of this, focusing on improving utility calculus may not be the most fruitful path for implementation inquiries. In short, HCI has the potential to add power to the phrase “AI makes philosophy honest,[11]‘ by testing and revising ethical theories in the real world. The dialogue between moral philosophy and HCI scholars and practitioners pursuing AMAs needs to be nurtured.

6 CONCLUSION

This paper describes and synthesizes disparate threads of HCI work on how normative ethical theories might be implemented in AMAs.

[11] Attributed to Daniel Dennett at a talk at the International Computers and Philosophy Conference, Laval, France in 2006, retrieved from: https://philosophynow.org/issues/72/How_Machines_Can_Advance_Ethics#:~:text=As%20Daniel%20Dennett%20stated%20in%20a%20talk%20at,theory.%20No%20single-principle%20action-based%20theory%20is%20generally%20accepted.‘

We categorize papers by what ethical theory(s) they were leveraging, and find they were largely offering analytical arguments, working towards implementation, or empirically investigating how the AMA may act in real world settings alongside humans. The vast majority of the work in this conversation is centered around two ethical paradigms, deontology and consequentialism. We also note a relative lack of empirical work, with most papers focused on offering analytical arguments or technical implementation proposals. We argue that these trends in the development of AMAs are at risk of encoding Western ethical reasoning into AMAs to the exclusion of valuable insights and approaches from other ethical traditions. There is also much room for further empirical work which can serve as a guide as to which ethical theories may be helpful in real settings and practices. Addressing these challenges will be critical to developing AMAs that can participate alongside people in messy, real-world ethical decision-making processes but is a task the HCI community is well positioned to tackle.

REFERENCES

[1] Madeline Akrich. 1992. The De-Scription of Technical Objects. In Bijker, WE. Law, J.(eds) Shaping technology/building society, MIT Press, Cambridge.

[2] Fahad Alaieri and André Vellino. 2016. Ethical Decision Making in Robots: Autonomy, Trust and Responsibility. In: Agah A., Cabibihan JJ., Howard A., Salichs M., He H. (eds) Social Robotics. ICSR 2016. Lecture Notes in Computer Science, vol 9979. Springer, Cham. https://doi.org/10.1007/978-3-319-47437-3_16

[3] Colin Allen, Iva Smit, and Wendell Wallach. 2005. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. Ethics and Inf. Technol. 7, 3 (September 2005), 149–155. DOI: https://doi.org/10.1007/s10676-006-0004-4

[4] Colin Allen, Gary Varner, and Jason Zinse. 2000. Prolegomena to any future artificial moral agent. J. Exp. Theor. Artif. Intell.. 12. 251-261. 10.1080/09528130050111428

[5] Michael Anderson, Susan Leigh Anderson and Vincent Berenz. 2017. A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm. In: AAAI 2016 Workshop on AI, Ethics & Society

[6] Michael Anderson and Susan Leigh Anderson. 2014. GenEth: a general ethical dilemma analyzer. Paladyn, Journal of Behavioral Robotics, 9, 337 - 357.

[7] Michael Anderson and Susan Leigh Anderson. 2014. Toward Ethical Intelligent Autonomous Healthcare Agents: A Case-Supported Principle-Based Behavior Paradigm. AISB 2014 - 50th Annual Convention of the AISB.

[8] Michael Anderson and Susan Leigh Anderson. 2008. ETHEL: Toward a Principled Ethical Eldercare Robot. In: Procs. AAAI Fall 2008 Symposium on AI in Eldercare (2008).

[9] Susan Leigh Anderson and Michael Anderson. 2011. A prima facie duty approach to machine ethics and its application to elder care. In Proceedings of the 12th AAAI Conference on Human-Robot Interaction in Elder Care (AAAIWS’11-12). AAAI Press, 2–7.

[10] Michael Anderson, Susan Leigh Anderson and Chris Armen. 2006. "An Approach to Computing Ethics," in IEEE Intelligent Systems, vol. 21, no. 4, pp. 56-63, July-Aug. 2006, doi: 10.1109/MIS.2006.64.

[11] Michael Anderson, Susan Leigh Anderson, and Chris Armen. 2005. Towards machine ethics: Implementing two action-based ethical theories. AAAI Fall Symposium – Technical Report 1-7.

[12] Konstantine Arkoudas, Selmer Bringsjord, and Paul Bello. 2005. Toward ethical robots via mechanized deontic logic. AAAI Fall Symposium - Technical Report.

[13] Thomas Arnold and Matthias Scheutz. 2017. Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI ’17). Association for Computing Machinery, New York, NY, USA, 445–452. DOI: https://doi.org/10.1145/2909824.3020255

[14] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’10). Association for Computing Machinery, New York, NY, USA, 1301–1310. DOI:https://doi-org.offcampus.lib.washington.edu/10.1145/1753326.1753521

[15] Ruha Benjamin. 2019. *Race after Technology :* Abolitionist Tools for the New Jim Code. Cambridge, UK ; Medford, MA: Polity

[16] Cynthia L. Bennett and Os Keyes. 2020. What is the point of fairness? disability, AI and the complexity of justice. SIGACCESS Access. Comput., 125, Article 5 (October 2019), 1 pages. DOI:https://doi-org.offcampus.lib.washington.edu/10.1145/3386296.3386301

[17] Jeremy Bentham. 2007. An Introduction to the Principles of Morals and Legislation. Dover Press. (originally published in 1789, of course).

[18] Martin Bentzen The Principle of Double Effect Applied to Ethical Dilemmas of Social Robots. Robophilosophy/TRANSOR (2016). doi: 10.3233/978-1-61499-708-5-268

[19] Martin Bentzen and Felix Lindner. 2018. A Formalization of Kant’s Second Formulation of the Categorical Imperative. In the proceedings of AIES 2018.

[20] Wiebe E Bijker, Thomas P Hughes, and Trevor Pinch, (Eds.). 2012. *The social construction of technological systems: New directions in the sociology and history of technology.* MIT press.

[21] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2017. A Declarative Modular Framework for Representing and Applying Ethical Principles. In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS ’17). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 96–104.

[22] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2015. Modelling Moral Reasoning and Ethical Responsibility with Logic Programming. In: Davis M., Fehnker A., McIver A., Voronkov A. (eds) Logic for Programming, Artificial Intelligence, and Reasoning. LPAR 2015. Lecture Notes in Computer Science, vol 9450. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-48899-7_37

[23] Adil Bilal, Stephen Wingreen, and Ravishankar Sharma. 2020. Virtue Ethics as a Solution to the Privacy Paradox and Trust in Emerging Technologies. In Proceedings of the 2020 The 3rd International Conference on Information Science and System (ICISS 2020). Association for Computing Machinery, New York, NY, USA, 224–228. DOI: https://doi.org/10.1145/3388176.3388196

[24] Susanne Bødker, Pelle Ehn, Dan Sjögren, and Yngve Sundblad. 2000. Co-operative Design—perspectives on 20 years with ‘the Scandinavian IT Design Model’. In proceedings of NordiCHI (Vol. 2000, pp. 22-24).

[25] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2015. Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?

[26] Vincent Bonnemains, Claire Saurel, and Catherine Tessier. 2018. Embedded ethics: some technical and ethical challenges. Ethics and Inf. Technol. 20, 1 (March 2018), 41–58.

[27] Selmer Bringsjord, Konstantine Arkoudas and Paul Bello, "Toward a General Logicist Methodology for Engineering Ethically Correct Robots," in IEEE Intelligent Systems, vol. 21, no. 4, pp. 38-44, July-Aug. 2006, doi: 10.1109/MIS.2006.82.

[28] José-Antonio Cervantes, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. 2020. Artificial Moral Agents: A Survey of the Current Status. Sci Eng Ethics. 2020;26(2):501-532. doi:10.1007/s11948-019-00151-x

[29] José-Antonio Cervantes, Luis-Felipe Rodríguez, Sonia López, Félix Ramos and Francisco Robles. 2016. Autonomous Agents and Ethical Decision-Making. Cogn Comput 8, 278–296 (2016). https://doi.org/10.1007/s12559-015-9362-8

[30] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. 2016. Ethical Judgment of Agents’ Behaviors in Multi-Agent Systems. In Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS ’16). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1106–1114.

[31] Morteza Dehghani, Emmett Tomai, and Matthew Klenk. 2008. An Integrated Reasoning Approach to Moral Decision-Making. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008. 1280-1286. 10.1017/CBO9780511978036.024.

[32] Carl DiSalvo, Phoebe Sengers, and Hrönn Brynjarsdóttir. 2010. Mapping the landscape of sustainable HCI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’10). Association for Computing Machinery, New York, NY, USA, 1975–1984. DOI:https://doi-org.offcampus.lib.washington.edu/10.1145/1753326.1753625

[33] Paul Dourish. 2017. The Stuff of Bits. MIT Press.

[34] Sjur Dyrkolbotn, Truls Pedersen, and Marija Slavkovik. 2018. On the Distinction between Implicit and Explicit Ethical Agency. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’18). Association for Computing Machinery, New York, NY, USA, 74–80. DOI:https://doi.org/10.1145/3278721.3278769

[35] Pelle Ehn. 1993. Scandinavian design: On participation and skill. Participatory design: Principles and practices, 41, 77.

[36] Pablo G. Esteban, Daniel Hernández García, Hee Rin Lee, Pauline Chevalier, Paul Baxter, and Cindy Bethel. 2018. Social Robots in Therapy: Focusing on Autonomy and Ethical Challenges. In Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI ’18). Association for Computing Machinery, New York, NY, USA, 391–392. DOI: https://doi.org/10.1145/3173386.3173562

[37] Bernice Fisher and Joan C. Tronto (1990). Toward a Feminist Theory of Caring. In Circles of Care: Work and identity in women’s lives, pp. 35-62.

[38] Sarah Fox, Jill Dimond, Lilly Irani, Tad Hirsch, Michael Muller, and Shaowen Bardzell. 2017. Social Justice and Design: Power and oppression in collaborative systems. In Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW ’17 Companion). Association for Computing Machinery, New York, NY, USA, 117–122. DOI: https://doi-

org.offcampus.lib.washington.edu/10.1145/3022198.3022201

[39] Batya Friedman and David Hendry. 2019. Value Sensitive Deisgn. The MIT Press.

[40] Batya Friedman and Helen Nissenbaum. 1997. Software agents and user autonomy. In Proceedings of the first international conference on Autonomous agents (AGENTS '97). Association for Computing Machinery, New York, NY, USA, 466–469. DOI: https://doi-org.offcampus.lib.washington.edu/10.1145/267658.267772

[41] Jean-Gabriel Ganascia. 2015. Non-monotonic Resolution of Conflicts for Ethical Reasoning. In: Trappl R. (eds) A Construction Manual for Robots' Ethical Systems. Cognitive Technologies. Springer, Cham. https://doi.org/10.1007/978-3-319-21548-8_6

[42] James Gips. 1995. Towards the Ethical Robot, in K. Ford, C. Glymour and P. Hayes, ed., Android Epistemology, Cambridge MA: MIT Press, pp. 243–252.

[43] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Vasanth Sarathy. 2019. Toward the Engineering of Virtuous Machines. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, New York, NY, USA, 29–35. DOI:https://doi.org/10.1145/3306618.3314256

[44] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Matthew Peveler 2019. Beyond the Doctrine of Double Effect: A Formal Model of True Self-sacrifice. In: Aldinhas Ferreira M., Silva Sequeira J., Singh Virk G., Tokhi M., E. Kadar E. (eds) Robotics and Well-Being. Intelligent Systems, Control and Automation: Science and Engineering, vol 95. Springer, Cham. https://doi.org/10.1007/978-3-030-12524-0_5

[45] Naveen Sundar Govindarajulu, and Selmer Bringsjord. 2017. On Automating the Doctrine of Double Effect. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI 2017), Melbourne, Australia, preprint available at this https://arxiv.org/abs/1703.08922

[46] Christopher Grau. 2006. "There Is No "I" in "Robot": Robots and Utilitarianism" in IEEE Intelligent Systems, vol. 21, no. 04, pp. 52-55, 2006. doi: 10.1109/MIS.2006.81

[47] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Williams. 2016. Embedding Ethical Principles in Collective Decision Support Systems. In the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence.

[48] Marcello Guarini. 2012. Conative Dimensions of Machine Ethics: A Defense of Duty. IEEE Trans. Affect. Comput. 3, 4 (January 2012), 434–442. DOI: https://doi-org.offcampus.lib.washington.edu/10.1109/T-AFFC.2012.27

[49] John N. Hooker and Tae Wan N. Kim. 2018. Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, New York, NY, USA, 130–136. DOI:https://doi.org/10.1145/3278721.3278753

[50] Lily Hu. 2018. Justice Beyond Utility in Artificial Intelligence. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, New York, NY, USA, 368–369. DOI:https://doi.org/10.1145/3278721.3278798

[51] Internet Encyclopedia of Philosophy. 2020. Ethics. Retrieved from: https://iep.utm.edu/ethics/

[52] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. Postcolonial computing: a lens on design and development. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10). Association for Computing Machinery, New York, NY, USA, 1311–1320. DOI:https://doi-org.offcampus.lib.washington.edu/10.1145/1753326.1753522

[53] Michael James Heron and Pauline Belford. 2015. Fuzzy ethics: or how I learned to stop worrying and love the bot. SIGCAS Comput. Soc. 45, 4 (November 2015), 4–6. DOI: https://doi.org/10.1145/2856428.2856429

[54] Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, New York, NY, USA, 37–44. DOI:https://doi-org.offcampus.lib.washington.edu/10.1145/3306618.3314267

[55] Deborah G. Johnson and Keith W. Miller. 2008. Un-making artificial moral agents. Ethics and Inf. Technol. 10, 2–3 (September 2008), 123–133. DOI:https://doi.org/10.1007/s10676-008-9174-6

[56] Imanuel Kant. 2012. *The Moral Law: Groundwork of the Metaphysic of Morals*. Routledge Classics. Routledge (originally published in 1785, of course).

[57] Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). Association for Computing Machinery, New York, NY, USA, Paper alt06, 1–11. DOI: https://doi.org/10.1145/3290607.3310433

[58] Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Joshua B. Tenenbaum, and Iyad Rahwan. 2018. A Computational Model of Commonsense Moral 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, 197–203. DOI:https://doi.org/10.1145/3278721.3278770

[59] Wilhelm E. J. Klein. 2016. Robots make ethics honest: and vice versa. SIGCAS Comput. Soc. 45, 3 (September 2015), 261–269. DOI: https://doi.org/10.1145/2874239.2874276

[60] Benjamin Kuipers. 2018. How can we trust a robot? Commun. ACM 61, 3 (March 2018), 86–95. DOI:https://doi.org/10.1145/3173087

[61] Liz Lane. 2016. Ethics of Activist Design: Designing a Technical Communication Pedagogy Grounded in Civic Engagement Activism. In Proceedings of the 34th ACM International Conference on the Design of Communication (SIGDOC '16). Association for Computing Machinery, New York, NY, USA, Article 54, 1. DOI: https://doi.org/10.1145/2987592.2987655

[62] Bruno LaTour. 1992. Where are the Missing Masses, sociology of a few mundane artefacts. In Shaping Technology-Building Society. Studies in Sociotechnical Change, Wiebe Bijker and John Law (editors), MIT Press, Cambridge Mass. pp. 225-259, 1992 [new expanded and revised version of article (35). Republication in the reader Johnson, Deborah J., and Jameson M Wetmore, eds. Technology and Society, Building Our Sociotechnical Future. Cambridge, Mass: MIT Press, 2008 pp. 151-180]

[63] Felix Lindner and Martin Mose Bentzen. 2017. The Hybrid Ethical Reasoning Agent IMMANUEL. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17). Association for Computing Machinery, New York, NY, USA, 187–188. DOI:https://doi.org/10.1145/3029798.3038404

[64] F. Lindner, M. M. Bentzen and B. Nebel, "The HERA approach to morally competent robots," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, 2017, pp. 6991-6997, doi: 10.1109/IROS.2017.8206625.

[65] Suvradip Maitra. 2020. Artificial Intelligence and Indigenous Perspectives: Protecting and Empowering Intelligent Human Beings. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 320–326. DOI:https://doi.org/10.1145/3375627.3375845

[66] Donald McMillan and Barry Brown. 2019. Against Ethical AI. In Proceedings of the Halfway to the Future Symposium 2019 (HTTF 2019). Association for Computing Machinery, New York, NY, USA, Article 9, 1–3. DOI:https://doi.org/10.1145/3363384.3363393

[67] John Stuart Mill. 2002. Utilitarianism. Hackett Publishing Company, Inc.; Second Edition. (originally published in 1863, of course).

[68] Darakhshan Mir, Iris Howley, Janet Davis, Evan Peck, and Deborah Tatar. 2019. Make and Take an Ethics Module: Ethics Across the CS Curriculum. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19). Association for Computing Machinery, New York, NY, USA, 1239. DOI: https://doi.org/10.1145/3287324.3287543

[69] Alexander G. Mirnig and Alexander Meschtscherjakov. 2019. Trolled by the Trolley Problem: On What Matters for Ethical Decision Making in Automated Vehicles. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 509, 1–10. DOI: https://doi.org/10.1145/3290605.3300739

[70] Luís Moniz Pereira and Ari Saptawijaya. 2007. Modelling morality with prospective logic. In Proceedings of the aritficial intelligence 13th Portuguese conference on Progress in artificial intelligence (EPIA'07). Springer-Verlag, Berlin, Heidelberg, 99–111.

[71] Alistair Morrison, Donald McMillan, and Matthew Chalmers. 2014. Improving consent in large scale mobile HCI through personalised representations of data. In Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI '14). Association for Computing Machinery, New York, NY, USA, 471–480. DOI: https://doi.org/10.1145/2639189.2639239

[72] Cosmin Munteanu, Heather Molyneaux, Wendy Moncur, Mario Romero, Susan O'Donnell, and John Vines. 2015. Situational Ethics: Re-thinking Approaches to Formal Ethics Requirements for Human-Computer Interaction. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). Association for Computing Machinery, New York, NY, USA, 105–114. DOI: https://doi.org/10.1145/2702123.2702481

[73] New York Times. 2018. Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. Retrieved from: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html#:~:text=The%20company%20quickly%20suspended%20testing%20in%20Tempe%20as,Newly%20released%20video%20offers%20clues%20about%20what%20happened

[74] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–16. DOI: https://doi-org.offcampus.lib.washington.edu/10.1145/3313831.3376392

[75] James E. Pickering, Patricia Ashman, Alex Gilbert, Dobrila Petrovic, Kevin Warwick and Keith. J. Burnham. 2018. Model-to-Decision Approach for Autonomous Vehicle Convoy Collision Ethics. *2018 UKACC 12th International Conference on Control (CONTROL)*, Sheffield, 2018, pp. 301-308, doi: 10.1109/CONTROL.2018.8516846.

[76] Thomas M. Powers. 2006. Prospects for a Kantian Machine, in IEEE Intelligent Systems, vol. 21, no. 4, pp. 46-51, July-Aug. 2006, doi: 10.1109/MIS.2006.77.

[77] ProPublica. 2016. How We Analyzed the COMPAS Recidivism Algorithm. Retrieved from: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[78] Anna Vartapetiance Samlasi and Lee Gillam. 2009. Machine Ethics for Metaverse Gambling: No Stake in a $24m Market?, *2009* Conference in Games and Virtual Worlds for Serious Applications. Coventry pp. 209-212, doi: 10.1109/VS-GAMES.2009.39.

[79] Katie Shilton. 2018. Values and Ethics in Human-Computer Interaction. Now Publishing.

[80] Katie Shilton and Jes A. Koepfler. 2013. Making space for values: communication & values levers in a virtual team. In Proceedings of the 6th International Conference on Communities and Technologies (C&T '13). Association for Computing Machinery, New York, NY, USA, 110–119. DOI: https://doi.org/10.1145/2482991.2482993

[81] Stanford Encyclopedia of Philosophy. 2019. Consequentialism. Retrieved from: https://plato.stanford.edu/entries/consequentialism/

[82] Stanford Encyclopedia of Philosophy. 2018. Chinese Ethics. Retrieved from: https://plato.stanford.edu/entries/ethics-chinese/

[83] Stanford Encyclopedia of Philosopjy. 2018. Doctrine of Double Effect. Retrieved from: https://plato.stanford.edu/entries/double-effect/

[84] Stanford Encyclopedia of Philosophy. 2016. Virtue Ethics. Retrieved from: https://plato.stanford.edu/entries/ethics-virtue/

[85] Stanford Encyclopedia of Philosophy. 2016. Kant's Moral Philosophy. Retrieved from: https://plato.stanford.edu/entries/kant-moral/

[86] Stanford Encyclopedia of Philosophy. 2016. Deontological Ethics. Retrieved from: https://plato.stanford.edu/entries/ethics-deontological/

[87] Stanford Encyclopedia of Philosophy. 2012. William David Ross. Retrieved from: https://plato.stanford.edu/entries/william-david-ross/

[88] Julia Stoyanovich, Bill Howe, Serge Abiteboul, Gerome Miklau, Arnaud Sahuguet, and Gerhard Weikum. 2017. Fides: Towards a Platform for Responsible Data Science. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM '17). Association for Computing Machinery, New York, NY, USA, Article 26, 1–6. DOI: https://doi.org/10.1145/3085504.3085530

[89] Petros Terzis. 2020. Onward for the freedom of others: marching beyond the AI ethics. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 220–229. DOI:https://doi.org/10.1145/3351095.3373152

[90] Ryan Tonkens. 2009. A Challenge for Machine Ethics. Minds Mach. 19, 3 (August 2009), 421–438. DOI:https://doi.org/10.1007/s11023-009-9159-1

[91] Austin L. Toombs, Shaowen Bardzell, and Jeffrey Bardzell. 2015. The Proper Care and Feeding of Hackerspaces: Care Ethics and Cultures of Making. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). Association for Computing Machinery, New York, NY, USA, 629–638. DOI: https://doi-org.offcampus.lib.washington.edu/10.1145/2702123.2702522

[92] Peter-Paul Verbeek. 2006. Materializing morality: Design ethics and technological mediation. Science, Technology, & Human Values, 31(3), 361-380.

[93] Ilse Verdiesen. 2018. The Design of Human Oversight in Autonomous Weapon Systems. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, New York, NY, USA, 388–389. DOI: https://doi.org/10.1145/3278721.3278785

[94] Laura Wächter and Felix Lindner. 2018. An Explorative Comparison of Blame Attributions to Companion Robots Across Various Moral Dilemmas. In Proceedings of the 6th International Conference on Human-Agent Interaction (HAI '18). Association for Computing Machinery, New York, NY, USA, 269–276. DOI:https://doi.org/10.1145/3284432.3284463

[95] Wendell Wallach and Colin Allen. 2009. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.

[96] Ruchen Wen, Blake Jackson, Tom Williams, and Qin Zhu. Towards a role ethics approach to command rejection. In HRI Workshop on the Dark Side of Human-Robot Interaction. 2019.

[97] 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 (AIES '19). Association for Computing Machinery, New York, NY, USA, 195–200. DOI: https://doi.org/10.1145/3306618.3314289

[98] Vincent Wiegel and Jan van den Berg. 2009. Combining Moral Theory, Modal Logic and Mas to Create Well-Behaving Artificial Agents. Int J of Soc Robotics 1, 233–242 (2009). https://doi.org/10.1007/s12369-009-0023-5

[99] Tom Williams, Qin Zhu, Ruchen Wen, and Ewart J. de Visser. 2020. The Confucian Matador: Three Defenses Against the Mechanical Bull. In Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20). Association for Computing Machinery, New York, NY, USA, 25–33. DOI:https://doi.org/10.1145/3371382.3380740

[100] Ava Thomas Wright. 2020. A Deontic Logic for Programming Rightful Machines. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 392. DOI:https://doi.org/10.1145/3375627.3375867

[101] 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. DOI: https://doi-org.offcampus.lib.washington.edu/10.1145/3306618.3314261

[102] Austin L. Toombs, Shaowen Bardzell, and Jeffrey Bardzell. 2015. The Proper Care and Feeding of Hackerspaces: Care Ethics and Cultures of Making. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). Association for Computing Machinery, New York, NY, USA, 629–638. DOI: https://doi-org.offcampus.lib.washington.edu/10.1145/2702123.2702522

[103] Levent Yilmaz. 2017. Verification and validation of ethical decision-making in autonomous systems. In Proceedings of the Symposium on Modeling and Simulation of Complexity in Intelligent, Adaptive and Autonomous Systems (MSCIAAS '17). Society for Computer Simulation International, San Diego, CA, USA, Article 1, 1–12.

[104] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. 2019. Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 191–200. DOI: https://doi.org/10.1145/3287560.3287577

[105] Qin Zhu, Tom Williams, Blake Jackson, and Ruchen Wen. 2020. Blame-Laden Moral Rebukes and the Morally Competent Robot: A Confucian Ethical Perspective. Sci Eng Ethics (2020). https://doi.org/10.1007/s11948-020-00246-w

[106] Qin Zhu, Tom Williams, and Ruchen Wen. 2019. Confucian robot ethics. In D. Wittkower (Ed.), 2019 Computer Ethics - Philosophical Enquiry (CEPE) Proceedings, (11 pp.). doi: 10.25884/5qbh-m581

A APPENDICES

Final Corpus, Categorization, and Context

Article	Reference Number	Theories Used	Implementation / Analysis / Empirical?	Context
Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic	49	D&C Universe Deontology	Implementation	Healthcare: Taking medication/elder care
The Confucian Matador: Three Defenses Against the Mechanical Bull	98	Other Philosophies	Analysis	Abstract analysis: moral rebukes

An Explorative Comparison of Blame Attributions to Companion Robots Across Various Moral Dilemmas	93	D&C Universe Deontology Utilitarianism	Empirical	Various moral dilemmas: trolley problem, stealing, child care, elder care
How can we trust a robot?	60	D&C Universe Virtue Ethics	Analysis	Autonomous vehicles
Artificial Intelligence and Indigenous Perspectives: Protecting and Empowering Intelligent Human Beings	65	Other Philosophies	Analysis	General AI
Onward for the freedom of others: marching beyond the AI ethics	89	Other Philosophies	Analysis	General AI
Justice Beyond Utility in Artificial Intelligence	50	D&C Universe Utilitarianism	Analysis	Job Hiring
Toward the Engineering of Virtuous Machines	43	Other Philosophies Virtue Ethics	Implementation	Example: Selling items and honesty
A Deontic Logic for Programming Rightful Machines	99	D&C Universe Deontology	Implementation	General AI
A Declarative Modular Framework for Representing and Applying Ethical Principles	21	D&C Universe	Implementation	Healthcare: Experimental Drug
Rightful Machines and Dilemmas	100	D&C Universe Deontology	Implementation	Trolley Problem / Autonomous Vehicles
Robots make ethics honest: and vice versa	59	D&C Universe Utilitarianism	Analysis	General AI
Verification and validation of ethical decision-making in autonomous systems	102	D&C Universe	Implementation	Weapons systems
Ethical Judgment of Agent's Behavior's in Multi-Agent Systems	30	D&C Universe	Implementation	General AI
A computation Model of Commonsense Moral Decision Making	58	D&C Universe Utilitarianism	Implementation	Autonomous Vehicles
Prospects for a Kantian Machine	76	D&C Universe Deontology	Implementation	General AI
Toward a General Logicist Methodology for Engineering Ethically Correct Robots	27	D&C Universe	Implementation	General AI with healthcare as an example.
Towards A Role Ethics Approach to Command Rejection	94	Other Philosophies	Empirical	General AI with cheating on an exam the example.
Blame-Laden Moral Rebukes and the Morally Competent Robot: A Confucian Ethical Perspective	104	Other Philosophies	Analysis	General AI
Confucian Robot Ethics	105	Other Philosophies	Analysis	General AI
The Principle of Double Effect Applied to Ethical Dilemmas of Social Robots	18	D&C Universe DDC	Implementation	Trolley Problem or Variation

BEYOND THE DOCTRINE OF DOUBLE EFFECT: A FORMAL MODEL OF TRUE SELF-SACRIFICE	44	D&C Universe DDC	Implementation	General AI with Trolley problem as an example
Conative Dimensions of Machine Ethics: A Defense of Duty	48	D&C Universe Deontology	Analysis	General AI with a healthcare (life support) and donation hypotheticals.
There Is No "I" in "Robot": Robots and Utilitarianism	46	D&C Universe Utilitarianism	Analysis	General AI
Prolegomena to any future artificial moral agent	4	Consequentialism D&C Universe Deontology Utilitarianism Virtue Ethics	Analysis	General AI
Towards the Ethical Robot	42	D&C Universe Deontology Utilitarianism Virtue Ethics	Analysis	General AI
Ethical Decision Making in Robots: Autonomy,Trust and Responsibility	2	Consequentialism D&C Universe Deontology	Analysis	General AI
Modelling Moral Reasoning and Ethical Responsibility with Logic Programming	22	D&C Universe DDC	Implementation	Trolley Problem
Non-monotonic Resolution of Conflicts for Ethical Reasoning	41	D&C Universe Utilitarianism	Implementation	Lying Dilemma, this situation presents a simple classical conflict: the agents have two possible actions to accomplish—lying or telling the truth—among which they have to choose one. Usually, it is considered that lying is bad and telling the truth good, which would naturally lead to tell the truth, but, in some circumstances, telling the truth may have such dramatic consequences that it looks better to lie
Toward ethical robots via mechanized deontic logic	12	D&C Universe	Implementation	General AI
Combining Moral Theory, Modal Logic and Mas to Create Well-Behaving Artificial Agents	97	D&C Universe Deontology	Implementation	Healthcare case study, but general AI context.
Modelling morality with prospective logic	70	D&C Universe DDC	Implementation	Trolley Problem
Semantics Derived Automatically From Language Corpora Contain Human-like Moral Choices	54	D&C Universe Deontology	Implementation	General AI
On Automating the Doctrine of Double Effect	45	D&C Universe DDC	Implementation	Trolley Problem
An Approach to Computing Ethics	10	D&C Universe Ross' Ethical Theory	Implementation	Healthcare / Advice machine

Toward a Principled Ethical Eldercare Robot	8	D&C Universe Ross' Ethical Theory	Implementation	Healthcare / Elder Care
A Prima Facie Duty Approach to Machine Ethics and its Application to Elder Care	9	D&C Universe Ross' Ethical Theory	Implementation	Healthcare / Elder Care
GenEth: A General Ethical Dilemma Analyzer	6	D&C Universe Ross' Ethical Theory	Empirical Implementation	Ethical dilemma analyzer
Toward Ethical Intelligent Autonomous Healthcare Agents: A Case-Supported Principle-Based Behavior Paradigm	7	D&C Universe Ross' Ethical Theory	Implementation	Healthcare/Eldercare
A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm	5	D&C Universe Ross' Ethical Theory	Implementation	Healthcare/Eldercare
The Hybrid Ethical Reasoning Agent IMMANUEL	63	D&C Universe DDC Utilitarianism	Implementation	Ethical dilemma analyzer
The HERA approach to morally competent robots	64	D&C Universe DDC Utilitarianism	Implementation	Ethical dilemma analyzer
On the Distinction between Implicit and Explicit Ethical Agency	34	D&C Universe	Implementation	General AI
A Challenge for Machine Ethics	90	D&C Universe Deontology	Analysis	General AI
A Formalization of Kant's Second Formulation of the Categorical Imperative	19	D&C Universe Deontology	Implementation	Gives examples, but in the context of General AI
Embedded ethics: some technical and ethical challenges	26	D&C Universe DDC Deontology Utilitarianism	Implementation	Trolley Problem
"Autonomous agents and ethical decision-making."	29	D&C Universe Deontology Utilitarianism	Implementation	Gives examples, but in the context of General AI
An Integrated Reasoning Approach to Moral Decision-Making	31	D&C Universe Deontology Utilitarianism	Implementation	General AI
"Embedding ethical principles in collective decision support systems."	47	D&C Universe Deontology Utilitarianism	Analysis	General AI
Machine Ethics for Metaverse Gambling: No Stake in a $24m Market?	78	D&C Universe Ross' Ethical Theory	Implementation	online gambling
Model-to-Decision Approach for Autonomous Vehicle Convoy Collision Ethics	75	D&C Universe Deontology Utilitarianism	Implementation	autonomous vehicles
Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?	25	D&C Universe Utilitarianism	Empirical	autonomous vehicles
Towards Machine Ethics: Implementing Two Action-Based Ethical Theories	11	D&C Universe Ross' Ethical Theory Utilitarianism	Implementation	advice machine