The Intersection of Ethical Concepts and Machine Intelligence: A Literature Review

Muhammad Osama Nusrat
Dept of Computing
Fast Nuces
Islamabad, Pakistan
i212169@nu.edu.pk

***Abstract*—The deployment of Artificial Moral Agents (AMAs) in real-world applications presents significant ethical challenges. While AMAs are designed for autonomous decision-making, their integration into society raises concerns regarding accountability, alignment with human moral values, and unintended consequences. This study critically examines three major ethical frameworks—deontology, utilitarianism, and virtue ethics—and evaluates their applicability in AI-driven decision-making. A real-world case study involving an autonomous vehicle crash is analyzed to illustrate the complexities of ethical reasoning in AI systems. The findings highlight that no single ethical framework adequately addresses all challenges, necessitating a hybrid approach that integrates rule-based ethics with adaptive learning mechanisms. The study underscores the need for interdisciplinary collaboration among ethicists, AI researchers, and policymakers to establish ethical guidelines for AI systems. Future research should focus on enhancing transparency, explainability, and fairness in AI ethics to ensure responsible deployment.**

I. Introduction

The rapid advancement of artificial intelligence (AI) has sparked ongoing discussions regarding its implications for human society. While AI systems are currently designed as tools to augment human capabilities, their increasing autonomy raises concerns about the potential for machines to surpass human intelligence. If AI can perform tasks traditionally reserved for humans, critical questions arise: What role will humans play in an AI-driven world? Will AI remain a tool, or could it evolve into an entity that dictates human decisions? These concerns highlight the need to establish ethical frameworks that govern AI behavior and ensure its alignment with human values [1].

Machine ethics is an emerging field that seeks to address these concerns by examining how AI can be designed to make morally sound decisions. Popular culture has long depicted AI in both utopian and dystopian narratives, with films such as *Terminator* introducing the idea of AI systems capable of both beneficial and harmful actions [1]. Beyond fiction, real-world AI applications in military systems, autonomous vehicles, and decision-making software have already demonstrated the ethical dilemmas posed by autonomous machines. One of the earliest attempts to formalize AI ethics was Isaac Asimov's "Three Laws of Robotics," introduced in 1942 [2]. These laws proposed that robots should (i) not harm humans, (ii) obey human commands unless it contradicts the first law, and (iii) protect themselves as long as doing so does not violate the first two laws. While these principles provide a foundational perspective, they remain theoretical constructs rather than practical ethical guidelines.

Several ethical theories have been explored to instill moral reasoning in AI systems, with three primary approaches dominating the field: deontological ethics, utilitarianism, and virtue ethics. Deontological ethics emphasizes duty-based moral obligations, where actions are deemed ethical based on adherence to predefined rules. Utilitarianism, in contrast, assesses actions based on their consequences, advocating for decisions that maximize overall societal benefit. Virtue ethics, rooted in the philosophies of Aristotle and Confucius, suggests that moral behavior is determined by the character and virtues of the decision-maker rather than specific rules or outcomes [1]. Each of these ethical frameworks presents unique challenges when applied to artificial moral agents (AMAs), necessitating further exploration into how AI can navigate ethical dilemmas effectively.

This study provides a critical review of recent developments in the field of machine ethics within the context of artificial intelligence. It examines existing research on ethical decision-making frameworks for AI systems, focusing on deontological, utilitarian, and virtue-based approaches. Furthermore, the study explores various techniques and methodologies employed in the design and implementation of Artificial Moral Agents (AMAs). By analyzing key challenges, ethical dilemmas, and potential risks associated with autonomous AI systems, this paper aims to contribute to the ongoing discourse on responsible AI development. The discussion integrates insights from the literature, highlighting both the ethical concerns and future prospects of machine ethics.

II. Literature Review

The field of machine ethics explores how artificial agents interact with human ethical values. Moor [3] proposed a taxonomy categorizing ethical agents into four types: *ethical impact agents, implicit ethical agents, explicit ethical agents*, and *entirely ethical agents*. Ethical impact agents influence human behavior without explicit ethical reasoning. For instance, a digital watch that helps a student wake up early indirectly reinforces responsibility. Implicit ethical agents,

such as ATMs or washing machines, are programmed to avoid unethical behavior but lack the capacity for independent moral reasoning.

Explicit ethical agents, on the other hand, are programmed to make ethical decisions based on predefined principles. Self-driving cars, for example, navigate roads by making situational judgments aligned with ethical frameworks. However, these systems remain susceptible to manipulation, posing security risks. The concept of entirely ethical agents, which possess human-like moral reasoning, remains hypothetical and represents an aspirational goal for AI research [3].

One key challenge in artificial moral agents is ensuring that they adhere to ethical constraints while making autonomous decisions. Hadfield-Menell et al. [4] conducted an experiment to examine whether an AMA could achieve its objective without causing harm. The results were context-dependent, highlighting the risk of AI overriding human authority when control mechanisms, such as an emergency shutdown switch, are insufficient. Russell [5] further argued that if AI surpasses human intelligence, traditional safety measures may become ineffective, raising concerns about the so-called "off-switch problem."

To guide ethical decision-making in AI, various moral theories have been explored. Tolmeijer et al. [6] compared Kantian ethics, Ross's theory, and utilitarianism to determine their suitability for AMAs. Kantian ethics, which prioritizes duty over consequences, was preferred due to its rule-based structure. However, utilitarianism, which aims to maximize collective well-being, was criticized for its controversial implications in dilemmas such as the trolley problem, where sacrificing an innocent person to save multiple individuals may be considered justifiable.

Research in machine ethics has also led to the development of structured ethical evaluation frameworks. Bjørgen et al. [7] introduced a database of ethical dilemmas to assess AI decision-making capabilities. These dilemmas span multiple domains, allowing researchers to test how AI systems identify, analyze, and resolve ethical conflicts. Charisi et al. [8] provided a comprehensive review of machine ethics methodologies, identifying both opportunities and challenges in developing ethical AI.

The historical evolution of machine ethics has been extensively studied. Dennis and Slavkovik [9] examined various approaches, including early rule-based systems and modern machine learning-based ethical reasoning. The study concluded that developing reliable ethical machine-learning algorithms remains a significant challenge.

Among theoretical models, Kantian machines have been proposed as a potential solution. Powers [10] discussed how Kantian ethics, emphasizing principles like respect for persons and autonomy, could be integrated into AI. While this approach may enable fair and impartial decision-making, challenges persist in encoding the complexities of Kantian moral reasoning into computational models.

Practical tools have also emerged in this domain. Anderson and Anderson [11] introduced GenEth, a system designed to codify ethical dilemmas across multiple frameworks. This tool facilitates ethical decision-making by analyzing potential courses of action. However, ethical theories themselves have limitations. For example, Ross's theory of *prima facie duties* emphasizes moral obligations like truthfulness and respect for elders. In practice, however, conflicting duties—such as the need for confidentiality—complicate ethical decision-making [12].

The development of *Friendly AI*, an approach aimed at aligning AI behavior with human values, has also been explored [13]. As AI systems become more powerful, ensuring that they act in a manner that prioritizes human well-being is critical. Gips [14] discussed the role of virtue-based consequentialism and deontological ethics in AI design. While these theories offer valuable guidelines, no single ethical framework is sufficient to address all ethical challenges in AI development. A hybrid approach that integrates multiple perspectives may be necessary to ensure ethical AI behavior in diverse real-world scenarios.

III. Developments and Implementations in Machine Ethics

A. Turing Test

The study of machine intelligence and independent reasoning emerged in the 1950s when Turing [15] introduced the Turing Test. This test was designed to evaluate whether a machine could exhibit intelligent behavior indistinguishable from that of a human. In the experiment, a human, a machine, and an interrogator engaged in communication via a teletypewriter. The interrogator, unaware of the machine's presence, was tasked with determining which participant was human. The machine's goal was to convincingly imitate human responses, making itself indistinguishable from a human participant. If the interrogator failed to correctly identify the machine, the system was considered to have passed the test, demonstrating a form of artificial intelligence.

Despite its significance, the Turing Test has notable limitations. It equates intelligence with human-like imitation, disregarding alternative forms of cognition that may differ from human reasoning. Furthermore, it assumes that human responses are always rational and intelligent, overlooking the fact that humans can also exhibit errors or irrational behavior. Consequently, a machine that mimics non-intelligent human behavior could still pass the test, leading to an inaccurate assessment of machine intelligence. These shortcomings highlight the need for more advanced and nuanced benchmarks for evaluating AI ethics and cognition.

B. Asimov's First Law

Asimov's First Law of Robotics states that a moral agent must not harm human beings. Additionally, the law dictates that robots must obey their owners while ensuring compliance with the first law and prioritize their own survival as long as these conditions are met. Alvarez et al. [16] implemented an experimental study to assess the feasibility of Asimov's First Law using an AI-controlled robot in a simulated environment.

The experiment utilized an A* search algorithm to determine optimal movement paths for a robot tasked with preventing human agents from entering a designated danger zone, referred to as the "lava tile."

Two experimental scenarios were tested. In the first scenario, a single human agent was protected by the robot, which successfully prevented them from entering the danger zone. However, in the second scenario, where multiple human agents were present, the system prioritized saving the nearest individual, sometimes failing to protect both. Additionally, the experiment revealed that if a human agent altered its trajectory toward the danger zone at the last moment, the robot often lacked the reaction time necessary to intervene effectively.

These findings illustrate the challenges in implementing Asimov's First Law in real-world AI applications. Autonomous systems must handle dynamic and unpredictable human behavior, which complicates rigid rule-based ethical programming. Furthermore, Asimov's Laws are insufficient in military contexts, where autonomous systems such as drone missiles are explicitly designed to cause harm, contradicting the fundamental premise of Asimov's First Law. Additionally, human actors could exploit robotic obedience by commanding robots to engage in unethical activities. These limitations suggest that a more complex and context-aware ethical framework is necessary for ensuring responsible AI decision-making.

C. HERA: Hybrid Ethical Reasoning Agent

HERA is a software library designed for moral decision-making in AI systems. Lindner et al. [17] introduced HERA as a hybrid ethical reasoning agent that incorporates multiple ethical theories to enhance autonomous moral judgment in robots. Unlike purely rule-based systems, HERA integrates learning mechanisms to allow AI systems to adapt their ethical reasoning based on user interactions.

The evaluation of HERA involved presenting it with multiple ethical dilemmas and comparing its responses with human decisions. Ethical theories such as utilitarianism, the principle of double effect, and the Pareto principle were incorporated into the system. The IMMANUEL robot was used as a testbed for the HERA framework, with the goal of identifying which ethical approach was most suitable for each scenario. Results indicated that HERA could effectively analyze ethical dilemmas and select an optimal course of action, demonstrating the potential for AI systems to integrate multiple ethical principles dynamically.

However, challenges remain in ensuring consistency and transparency in AI moral decision-making. Ethical decisions often require contextual awareness, and predefined ethical principles may not always align with human expectations in real-world situations. Future improvements to HERA should focus on refining adaptive learning mechanisms to enhance contextual ethical reasoning.

D. Ethel: Ethical Eldercare Robot

Ethel is an AI-driven robotic system designed to assist elderly individuals by ensuring medication adherence and dietary compliance. Anderson and Anderson [18] developed the Ethel system as an ethical eldercare assistant capable of making moral decisions in caregiving environments.

One of the key challenges in eldercare robotics is handling non-compliance. If a patient refuses to take their medication, the robot must determine an appropriate response that balances ethical considerations with practical caregiving needs. Ethel was programmed with multiple response strategies, allowing it to adjust its actions based on the patient's mood and behavior. This adaptive approach aimed to ensure both ethical sensitivity and patient autonomy.

The ethical considerations surrounding AI-driven medical assistance extend beyond eldercare. AI has the potential to enhance diagnostic accuracy, reduce healthcare costs, and improve treatment efficiency. However, it also introduces risks such as misdiagnosis, algorithmic bias, and potential violations of patient privacy [19]. The integration of autonomous ethical reasoning in AI-assisted healthcare remains an ongoing research challenge, requiring robust regulatory frameworks to ensure ethical compliance.

IV. Artificial Moral Agents and Ethical Approaches

The integration of Artificial Moral Agents (AMAs) into various domains presents both opportunities and challenges. On the positive side, AMAs can enhance productivity, improve decision-making efficiency, and contribute to fields such as healthcare, transportation, and military applications. In strategic environments, AMAs offer advantages by rapidly processing ethical dilemmas and making calculated decisions. However, their deployment also raises ethical concerns, including job displacement, socioeconomic instability, and the risk of reinforcing biases in AI-driven decision-making. Addressing these challenges requires a well-defined ethical framework to ensure that AMAs align with human moral values and societal norms.

A. Approaches to Building Artificial Moral Agents

The development of AMAs follows three primary approaches: top-down, bottom-up, and hybrid methodologies [20, 21, 22]. Each approach has distinct advantages and limitations, influencing how AMAs process ethical decisions.

1) Top-Down Approach: The top-down approach involves programming AMAs with a predefined set of ethical rules. These rules function as explicit constraints, guiding the AI system's decision-making processes. Notable examples of top-down approaches include Asimov's Laws of Robotics and utilitarian principles embedded within AI decision models.

One advantage of the top-down approach is its predictability AMAs strictly follow established ethical guidelines, reducing uncertainty in decision-making. For instance, self-driving vehicles often employ rule-based programming to adhere to traffic laws. However, a key limitation is rule conflict and ambiguity. Asimov's first two laws, for example, may contradict one another in specific situations, making it unclear how an AMA should act. Furthermore, rigid adherence to rules prevents

adaptability; if an AI encounters an ethical dilemma not explicitly addressed in its programming, it may fail to respond effectively.

2) Bottom-Up Approach: Unlike the top-down approach, the bottom-up methodology allows AMAs to learn ethical behavior through experience and adaptation [21]. Inspired by human learning processes, AMAs trained via bottom-up methods refine their ethical decision-making capabilities through exposure to diverse moral scenarios.

The primary advantage of this approach is its flexibility—AMAs develop ethical reasoning dynamically, rather than relying on preprogrammed constraints. This method is particularly useful in human-interacting AI systems, such as social robots and conversational AI, where ethical behavior may vary depending on context.

However, the bottom-up approach presents significant challenges. Ethical unpredictability is a major concern; AMAs may inadvertently learn undesirable behaviors if trained on biased or ethically ambiguous data. For example, an AI interacting with online communities may internalize discriminatory tendencies if exposed to biased inputs. Additionally, the bottom-up approach requires extensive training time, making real-world deployment challenging. AI systems must undergo rigorous testing to ensure they develop a robust ethical framework that generalizes across multiple contexts.

3) Hybrid Approach: The hybrid approach combines elements of both the top-down and bottom-up methodologies, integrating predefined ethical rules with adaptive learning mechanisms [22]. This model seeks to overcome the limitations of each individual approach by allowing AMAs to both follow predefined ethical principles and learn from their experiences.

In a hybrid framework, AMAs may be initialized with a core set of ethical guidelines while progressively refining their decision-making through real-world interactions. This model holds promise for applications requiring both consistency and adaptability, such as autonomous medical diagnostics and ethical decision-making in autonomous weapons systems.

Despite its potential, the hybrid approach remains challenging to implement. One major difficulty is balancing rule-based decision-making with experiential learning. Ethical principles, when hardcoded, may restrict adaptability, whereas excessive reliance on learning-based adaptation risks unintended ethical drift. Furthermore, the integration of multiple ethical frameworks increases computational complexity, requiring sophisticated architectures capable of harmonizing conflicting ethical principles.

B. Challenges and Future Directions

While each approach presents distinct advantages, no single methodology fully addresses the complexity of ethical AI decision-making. Top-down approaches provide clear guidelines but lack adaptability, while bottom-up models offer learning flexibility but introduce unpredictability. Hybrid approaches attempt to bridge these gaps but face implementation challenges related to ethical coherence, explainability, and computational feasibility.

Future research should focus on:

- Developing explainable hybrid models that allow AI to provide transparent justifications for ethical decisions.
- Integrating reinforcement learning with ethical constraints to enhance real-time moral reasoning in AMAs.
- Creating interdisciplinary regulatory frameworks that ensure AMAs align with human-centric ethical standards.
- Exploring AI-human collaboration in ethical decision-making, where AMAs assist rather than replace human moral judgment.

Addressing these challenges will be essential to the responsible deployment of Artificial Moral Agents, ensuring that AI systems act in ways that are both ethically sound and socially beneficial.

V. Discussion

The ethical decision-making process in Artificial Moral Agents (AMAs) is particularly complex in real-world scenarios where autonomous systems must make split-second decisions. A relevant case study is a Tesla Model S crash that occurred in Brașov, Romania, in October 2024. In this incident, a pedestrian tripped due to poor road conditions and fell onto the street. A Tesla vehicle, attempting to avoid hitting the pedestrian, swerved into oncoming traffic, colliding with another vehicle. This raises the critical ethical question: Did Tesla's decision align with ethical reasoning, and which ethical framework would have been the most justifiable in this scenario?

A. Analyzing the Tesla Incident Through Ethical Frameworks

To understand the implications of this decision, we examine the incident using three major ethical theories: deontology, utilitarianism, and virtue ethics.

1) Deontological Perspective: Deontology, founded on the principle of duty-based ethics, asserts that moral actions should adhere to strict ethical rules, regardless of consequences. According to Kantian deontology, the Tesla should follow absolute ethical principles, such as "Do not harm humans" or "Obey traffic laws."

From this perspective, the Tesla's decision to avoid the pedestrian aligns with the deontological principle of preventing direct harm to a human being. The AI system, if programmed with deontological ethics, prioritized a moral duty to avoid harming the pedestrian over potential indirect harm caused by swerving into another vehicle. However, a rigid deontological approach does not account for contextual consequences, meaning that while Tesla followed a moral rule, the resulting harm to the oncoming vehicle challenges the applicability of strict deontological ethics in real-world AI decision-making.

2) Utilitarian Perspective: Utilitarianism evaluates decisions based on their overall consequences, aiming to maximize well-being and minimize harm. A utilitarian AI system would assess the total potential damage to all parties involved and choose the action that results in the least overall harm.

Applying this perspective to the Tesla incident, the AI should have calculated which option would minimize total harm:

- If the Tesla had maintained its path, the pedestrian might have suffered severe injuries or death.
- By swerving, the Tesla protected the pedestrian but caused a collision, injuring the occupants of the other vehicle.

A strictly utilitarian approach would involve risk assessment and probability calculations. If statistical data showed that swerving into traffic had a higher probability of causing multiple injuries, a utilitarian AMA might instead choose to brake forcefully rather than swerve, as this could minimize total casualties. However, the challenge in real-time AI ethics is whether a machine can accurately compute and compare harm in the fraction of a second before a collision.

3) Virtue Ethics Perspective: Virtue ethics, originating from Aristotle, focuses on moral character and virtuous decision-making rather than rigid rules or consequences. A system based on virtue ethics would consider what a morally virtuous driver should do in the given situation, rather than strictly following predefined rules or calculating outcomes.

In this case, a virtuous human driver might weigh responsibility, care, and justice:

- Protecting the pedestrian could be seen as an act of compassion and responsibility.
- However, recklessly endangering another vehicle contradicts the virtue of prudence.
- A virtuous driver would likely opt for controlled braking rather than swerving to minimize the total harm while still acting ethically toward all parties involved.

A virtue-based AI system would ideally balance care for human life, responsibility in decision-making, and practical wisdom in executing maneuvers. However, designing AI to exhibit human-like moral reasoning remains a significant challenge.

B. Evaluating Tesla's Ethical Decision

Tesla's AI, whether or not it was in Autopilot mode, followed an ethical reasoning process that aligns most closely with deontological ethics—it prioritized not harming the pedestrian, even at the risk of harming others. However, from a utilitarian perspective, this decision may not have been optimal, as it transferred the risk from one party to another without minimizing overall harm. A virtue ethics approach might suggest that the Tesla should have prioritized braking instead of swerving, balancing both moral responsibility and harm reduction.

C. Broader Implications for AI Ethics in Autonomous Vehicles

The Tesla incident highlights several broader ethical concerns in AI decision-making:

- Regulatory and legal accountability: Should manufacturers be held liable if an AI-driven vehicle makes an ethically controversial decision?
- Real-time ethical calculations: Can AI accurately assess moral trade-offs within milliseconds, or should it rely on fixed ethical rules?
- AI as moral agents: Should AI attempt to mimic human moral reasoning (virtue ethics), follow strict legal principles (deontology), or optimize for least harm (utilitarianism)?

The optimal ethical approach for autonomous vehicles may involve a hybrid ethical model, integrating deontological safeguards to prevent direct harm, utilitarian calculations to minimize total harm, and virtue-based reasoning to ensure responsible decision-making in uncertain conditions. Future advancements in AI ethics must aim to balance these elements to ensure that autonomous systems make morally justifiable decisions in real-world scenarios.

VI. Conclusion and Future Work

The ethical decision-making capabilities of Artificial Moral Agents (AMAs) remain an ongoing challenge in artificial intelligence. This study analyzed machine ethics through the lenses of deontology, utilitarianism, and virtue ethics, evaluating their applicability in real-world scenarios. A case study of a Tesla vehicle's decision-making process in an accident scenario highlighted the complexities of ethical AI decision-making. The findings indicate that no single ethical framework is sufficient to guide autonomous systems in morally ambiguous situations. While deontological ethics ensures adherence to predefined rules, it lacks adaptability. Utilitarianism seeks to minimize overall harm but struggles with computational feasibility in real-time decision-making. Virtue ethics emphasizes moral character but remains difficult to formalize within AI systems.

Despite significant advancements in ethical AI, unresolved challenges persist. The absence of a universally accepted ethical framework complicates the development of AMAs capable of making morally sound decisions in diverse and unpredictable environments. Furthermore, questions surrounding accountability, explainability, and bias remain pressing concerns. Ethical AI systems must balance the competing priorities of rule-based compliance, consequence-driven calculations, and adaptive ethical reasoning.

Future research should focus on developing a hybrid ethical framework that integrates principles from multiple ethical traditions while ensuring AI systems remain transparent, interpretable, and aligned with human values. One promising direction is the incorporation of reinforcement learning techniques, allowing AMAs to refine their ethical decision-making based on real-world interactions. Additionally, establishing regulatory frameworks and interdisciplinary collaboration among ethicists, AI researchers, legal scholars, and policymakers will be essential to ensuring responsible AI deployment. As artificial intelligence continues to evolve, addressing these ethical challenges will be crucial in fostering trust and ensuring that AI-driven decision-making aligns with societal norms and expectations.

References

[1] M. Coeckelbergh, *AI Ethics*. The MIT Press, 04 2020. [Online]. Available: https://doi.org/10.7551/mitpress/12549.001.0001

[2] I. Asimov, *I. robot*. Narkaling Productions., 1940.

[3] J. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18–21, 2006.

[4] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "The off-switch game," in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[5] S. Russell, "Provably beneficial artificial intelligence," in *Proceedings of the 27th International Conference on Intelligent User Interfaces*, ser. IUI '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3. [Online]. Available: https://doi.org/10.1145/3490099.3519388

[6] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein, "Implementations in machine ethics: A survey," *ACM Comput. Surv.*, vol. 53, no. 6, Dec. 2021. [Online]. Available: https://doi.org/10.1145/3419633

[7] E. P. Bjørgen, S. Madsen, T. S. Bjørknes, F. V. Heimsæter, R. Håvik, M. Linderud, P.-N. Longberg, L. A. Dennis, and M. Slavkovik, "Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 23–29.

[8] V. Charisi, L. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A. F. T. Winfield, and R. Yampolskiy, "Towards moral autonomous systems," 2017. [Online]. Available: https://arxiv.org/abs/1703.04741

[9] L. A. Dennis and M. Slavkovik, "Machines that know right and cannot do wrong: The theory and practice of machine ethics," *IEEE Intelligent Informatics Bulletin*, vol. 19, no. 1, pp. 8–11, 2018.

[10] T. Powers, "Prospects for a kantian machine," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 46–51, 2006.

[11] M. Anderson and S. L. Anderson, "Geneth: a general ethical dilemma analyzer," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 337–357, 2018. [Online]. Available: https://doi.org/10.1515/pjbr-2018-0024

[12] ——, "Machine ethics: Creating an ethical intelligent agent," *AI magazine*, vol. 28, no. 4, pp. 15–15, 2007.

[13] ——, "Robot be good," *Scientific American*, vol. 303, no. 4, pp. 72–77, 2010.

[14] J. Gips, "Toward the ethical robot," in *Android Epistemology*, K. M. Ford, C. N. Glymour, and P. J. Hayes, Eds. MIT Press, 1994, pp. 243–252.

[15] A. M. Turing, "I.—computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, 10 1950. [Online]. Available: https://doi.org/10.1093/mind/LIX.236.433

[16] M. Alvarez, O. Berge, A. Berget, E. Bjorknes, D. V. K. Johnsen, F. Madsen, and M. Slavkovik, "Implementing asimov's first law of robotics," *Norsk IKT-konferanse for forskning og utdanning*, Nov 2017. [Online]. Available: https://www.ntnu.no/ojs/index.php/nikt/article/view/5354

[17] F. Lindner, M. M. Bentzen, and B. Nebel, "The hera approach to morally competent robots," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 6991–6997.

[18] M. Anderson and S. L. Anderson, "Ethel: Toward a principled ethical eldercare system." in *AAAI fall symposium: AI in eldercare: New solutions to old problems*, vol. 2, 2008.

[19] A. F. Winfield, K. Michael, J. Pitt, and V. Evers, "Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue]," *Proceedings of the IEEE*, vol. 107, no. 3, pp. 509–517, 2019.

[20] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," vol. 7, no. 3, pp. 149–155. [Online]. Available: https://doi.org/10.1007/s10676-006-0004-4

[21] W. Wallach, C. Allen, and I. Smit, "Machine morality: bottom-up and top-down approaches for modelling human moral faculties," in *Machine ethics and robot ethics*. Routledge, 2020, pp. 249–266.

[22] F. Song and S. H. F. Yeung, "A pluralist hybrid model for moral AIs," vol. 39, no. 3, pp. 891–900. [Online]. Available: https://doi.org/10.1007/s00146-022-01601-0