Artificial Moral Agents: Should Machines Take Ethical Responsibility?

Patrick Reiter[1], Utku Norman[2,*], Nora Weinberger[2] and Barbara Bruno[1]

***Abstract*— Robots and AI systems are increasingly being developed for and deployed in contexts where decision-making entails moral implications. Examples include autonomous vehicles navigating the ethical dilemmas of traffic, healthcare robots tasked with ensuring patient safety and well-being, assistive robots upholding the autonomy and dignity of the elderly and people with disabilities, and social robots guiding children through educational experiences while assisting their emotional and cognitive development. The transition from passive tools to autonomous entities with moral decision-making capabilities has ignited extensive debate about the ethical responsibilities of Artificial Moral Agents (AMAs). This paper synthesizes recent advancements in AMA research, addressing evolving debates on their ethical feasibility and societal integration. We assess key arguments against and for AMAs, highlighting impacts on moral responsibility, cultural perspectives, and stakeholder trust. Our analysis reveals that while AMAs remain a subject of theoretical debate, their integration into ethically sensitive contexts is increasingly proposed, necessitating clearer governance strategies. Given the recent AI advancements and increased deployment of robotics in high-stakes settings, this synthesis is timely and highlights the urgency of addressing these ethical challenges. By analyzing recent advancements and diverse perspectives, we aim to provide a concise but comprehensive understanding of the complexities involved in empowering machines with moral decision-making capabilities.**

I. Introduction

With the increasing autonomy of robots and AI systems, moral decision-making extends beyond human actors to the very design and deployment of intelligent systems. Whether in self-driving cars, assistive healthcare robotics, or algorithmic governance, the question of Artificial Moral Agents (AMAs) has become central to the ethics of robotics. As autonomous robots become increasingly integrated into various professional domains, assuming roles traditionally associated with human ethical responsibilities, the question of robotic moral agency becomes unavoidable. Should robots be allowed to make moral decisions and take responsibility? If not, who takes responsibility for their actions?

The rapid advancement of Artificial Intelligence (AI), especially with Large Language Models (LLMs) like ChatGPT, has accelerated the rise of AI technologies at an unprecedented pace [1]. Widespread access to these technologies has made AI more prevalent in various sectors [2], transforming how we interact with and utilize these systems [3], [4]. This progress prompts us to consider how much 'power' we are willing to delegate to these AI systems. Will AI remain restricted to narrowly defined tasks, or are we advancing toward the creation of AMAs capable of independent moral reasoning [5]?

This paper argues that while AMAs have the potential to significantly benefit society, notably expanding the range of scenarios in which robots can be deployed, such benefits depend on addressing the theoretical and ethical challenges they present. This involves solving open questions in the field of ethics and understanding the implications of integrating moral decision-making capabilities into AI systems.

Building on the influential work of Formosa and Ryan [6], who examined arguments against and for AMAs, this survey incorporates newer publications to offer updated perspectives and insights. We begin by defining key terms in Sec. II and outlining our systematized review methodology in Sec. III. We then examine two central arguments against and two in favor of AMAs in Sec. IV and V, respectively. Unlike [6], which favors breadth over depth, this paper focuses on a few key reasons, providing a more in-depth analysis.

In Sec. VI, we synthesize our findings and discuss how AMA decisions can impact different stakeholders involved in their development, usage, and the consequences of their actions. We reflect on how AMAs might influence societal perceptions of moral agency, including potential shifts in human moral development, ethical expectations in social systems, and the definition of moral failure in technological contexts. Furthermore, we consider the cultural variability of moral agency, discussing how different societies might respond to AMAs based on existing ethical frameworks and societal norms. Finally, in Sec. VII, we outline the practical and ethical implications of our findings and propose directions for future research and policy development.

II. Artificial Moral Agents (AMAs)

A recent survey defines an AMA as "a virtual agent (software) or physical agent (robot) capable of engaging in moral behavior or at least of avoiding immoral behavior" [9].

The term *artificial* refers to the non-biological origin of these agents, which are manufactured from pre-existing materials, in contrast with *natural* agents like humans [16].

For an agent to be considered *moral*, it must engage in behaviors evaluated under moral standards and be held accountable for its actions [17]. This involves a commitment to adhere to ethical norms, deserving of praise or blame [16]. However, this moral capacity is complex and intertwined with the agent's ability to act autonomously and adaptively within its environment [18]. Autonomy in moral reasoning

*This work was funded by the Baden-Württemberg Ministry of Science, Research and Art (MWK), using funds from the state digitalization strategy digital@bw. Corresponding author's email: utku.norman@kit.edu

[1]Socially Assistive Robotics with Artificial Intelligence (SARAI) Lab, Karlsruhe Institute of Technology (KIT), Germany

[2]Institute for Technology Assessment and Systems Analysis (ITAS), Karlsruhe Institute of Technology (KIT), Germany

TABLE I

SUMMARY OF KEY FACTS FROM SURVEYED PAPERS ON AMAS (N = 10). "FOCUS" INDICATES WHETHER THE WORK INVOLVES EMBODIED APPLICATIONS (PHYSICAL), THEORETICAL DISCUSSIONS (CONCEPTUAL), OR A COMBINATION OF BOTH (BOTH).

Paper	Key Reasons For AMAs	Key Reasons Against AMAs	Focus	Scenario of Use
Formosa and Ryan (2021) [6]	Essential for ethical integration of AI in social contexts; potential societal benefits	Technical and ethical challenges, including responsibility gaps	Both	Healthcare, autonomous vehicles
Sharkey (2017) [7]	Not specified	Robots lack moral agency; humans must remain responsible	Physical	Caregiving, policing, decision-making
van Wynsberghe and Robbins (2019) [8]	Not specified	Criticizes inevitability and complexity claims; emphasizes safety	Both	Service robots, military applications
Cervantes et al. (2020) [9]	Proposes ethical frameworks for safe AI-human interaction.	Challenges of mirroring human judgment in complex situations	Both	Ethical agent development
Bonnefon et al. (2024) [10]	Improves decision-making in high-stakes situations, reducing human bias and errors	Public resistance due to perceived lack of empathy, fairness concerns, and extreme performance expectations	Both	Medicine, law, autonomous vehicles
Hallamaa and Kalliokoski (2020) [11]	Enhances efficiency and safety in critical sectors like surgery and transportation	AI lacks moral agency; accountability gaps, human disempowerment	Both	AI-human decision systems
Nyholm (2018) [12]	Provides consistent ethical decision-making for crash scenarios	Raises concerns about fairness, public trust, and moral responsibility.	Physical	Self-driving cars in traffic accidents
Sullivan and Fosso Wamba (2022) [13]	Not specified	Emphasizes responsibility on developers and companies for AI harms	Both	Industries, stakeholder responsibility
Tigard (2021a) [14]	Nuanced understanding of 'responsibility' in AI ethics; distinguishes types of responsibility	Critiques vague usage of 'responsible AI'; calls for clear ethical frameworks	Conceptual	AI ethics, structured responsibility
Tigard (2021b) [15]	Not specified	Accountability issues and moral responsibility gaps in AI	Both	Warfare, healthcare

is required for AMAs, as their behavior must "occur without the direct real-time input from a human user" [8].

An *agent* is distinguished from a *tool*. Sullins argues that if we consider virtual and physical agents (henceforth referred to as "AI" and "robots") merely as tools, moral responsibility would always reside with the user of the tool. Consequently, it would be meaningless to speak of AMAs [18].

Three criteria qualify an agent as an AMA: interactivity, autonomy, and adaptability [17]. Interactivity means that the agent is capable of interacting with its environment and interpreting situations based on the inputs it receives. Autonomy, as already described, denotes the lack of human intervention [8]. Adaptability means that the agent can handle multiple and variable situations of various complexity. Within this framework, morality is determined by the agent's actions respecting a threshold defined by relevant observables in these variables, which guides the evaluation of the agent's behavior as moral or not in complex scenarios [17].

Understanding what makes an agent *artificial* and *moral* sets the stage for exploring how researchers have studied these agents. The next section explains how this study reviewed the literature on AMAs.

III. METHODOLOGY

This study presents a *systematized* literature review of recent work on AMAs, conducted by a single reviewer. While it does not follow the formal protocols of a systematic review, it adopts a structured and transparent approach by applying a defined search strategy and consistent selection criteria. This aligns with the definition of a systematized review as outlined by Grant and Booth [19]. The review employed the *snowball method*, as described by Kai O. Arras [20], in which an initial set of relevant papers is identified and expanded by tracing citations and references.

The initial search was conducted on Google Scholar using the key phrases: "AI moral agency", "robot moral agency" and "artificial moral agency". Papers were selected based on i) their explicit engagement with the concept of moral agency, for instance by including an explicit definition, and ii) their philosophical or ethical analysis of whether machines should be granted moral decision-making capabilities, including those that argue *for or against* the development of AMAs. The review process focused on works that contribute to the *theoretical and normative discourse* surrounding AMAs.

To maintain this focus, we prioritized literature addressing conceptual, ethical, and philosophical dimensions, rather than technical implementations, such as building or programming AMAs, or empirical user studies, like comparing AMA decisions to human decisions using scenarios like the trolley problem as in [21]. This allowed for a coherent exploration of the moral and societal implications of AMAs as discussed in academic discourse. However, we acknowledge that not including user studies may introduce bias by omitting empirical data on public perceptions and reactions to AMAs. This may limit insights into how public perceptions shape the ethical acceptance of AMAs and their integration into society. Future research could complement this study by incorporating empirical findings to provide a more comprehensive view.

The iterative search process yielded 13 relevant papers, including both foundational and recent contributions. The three foundational works [16]–[18] primarily define what constitutes an AMA, without delving deeply into arguments against or for them. The ten more recent papers, all published between 2017 and 2024, explore the disadvantages and

advantages of developing AMAs for a variety of contexts. For a summary, see Table I.

Interestingly, albeit perhaps not surprisingly, only one paper by Formosa and Ryan [6] openly advocates for the advancement of AMA-related technology, while two papers clearly argue against their development and spread [7], [8]. The remaining seven papers [9]–[15] can be considered as taking a neutral stance, examining the potential dangers alongside the potential benefits and uses of moral agents, without advocating for or against them. This diverse selection of papers serves as our basis for evaluating the benefits and drawbacks of developing and using AMAs.

Our paper differs by synthesizing findings from the literature to explore the broader implications of AMAs. We reflect on how AMAs might influence societal perceptions of moral agency, ethical expectations, and cultural variability. This approach provides a nuanced perspective that goes beyond merely examining potential dangers and benefits.

Having briefly recalled what constitutes an AMA in Sec. II, we can now examine the arguments against and for their use highlighted in the 10 selected articles. While Table I provides a list of all key arguments raised in the reviewed articles, the following sections specifically focus on the two most frequent arguments, respectively against and for AMAs.

IV. Reasons Against AMAs

A. Impossibility to Teach Correct Ethics to a Machine

Moral dilemmas often lack clear-cut solutions, particularly when every possible choice entails ethical trade-offs. For example, in the trolley problem, one can argue for both flipping the switch to have fewer people die or not acting at all to avoid causing death directly. Nyholm provides an example of an applied trolley problem in the context of self-driving cars, where the system must decide in unavoidable crash scenarios [12]. While human responses may be instinctive, AMAs typically have sufficient processing time to analyze data and make calculated decisions. However, Sharkey is skeptical about the possibility of teaching AMAs the 'correct' decision, as machines do not truly care about human lives, making their decisions merely *as if* they did [7]. Sharkey thus contends that this suggests the terms 'ethical' or 'moral' cannot meaningfully apply to AMAs decisions.

Similarly, Himma argues that a moral decision require free choice and deliberation [16]. As such, van Wynsberghe and Robbins point out the "impossibility of finding universal agreement concerning the ethical theory used to program a machine" [8]. Since we cannot teach AMAs free will or universally 'correct ethical answers', their existence may deem impossible. Conversely, Formosa and Ryan argue that the previously mentioned points make it harder, but not impossible, to build AMAs [6]. They contend that disagreements in moral decision-making are not a valid reason against creating AMAs, as these disagreements also occur with human decisions. Thus, the advantage of using AMAs as moral decision-makers should be measured by their consequences, not the correctness of the decisions. Tigard even considers a "consequential-based form of punishment or reward" [15] for the decisions made by AMAs.

Beside the above discussion on the 'absolute' correctness of an AMA's decisions, authors also discuss the 'perceived' correctness of moral decisions and how this might depend on who makes them. Bonnefon et al. argue that delegating moral decision-making to robots or AI can lead to psychological distance from the potential victims of a decision, which can be seen as a negative consequence of using AMAs, even if there is no discrepancy between the decisions made by humans and AMAs [10]. Furthermore, while it may be impossible to teach 'correct morality', it certainly is possible to intentionally teach unethical behavior to an AMA. Bonnefon et al. warn that AMAs may be misused by people with malicious intentions, especially given this possibility of distancing oneself from an AMA's decisions. In contrast, as long as AMAs do not exist, the connection between a moral decision and its responsible human is much harder to deny.

Despite these challenges, ongoing research explores computational models of moral reasoning, such as top-down ethical programming, bottom-up learning models, and hybrid approaches [22]. However, these remain limited in capturing the complexity of human ethical reasoning and cultural variation. The difficulty of encoding human-like moral reasoning into machines suggests that these models should act only as aids in ethical reasoning [23]. There are barely any options for adding cultural influences or societal preferences, and almost all systems assume the user cannot influence output [24]. Prototypes are generally tested in controlled environments, unlike real scenarios with unexpected moral dilemmas, indicating a long way to go before AI can replace human judgment in complex situations [9].

Meanwhile, psychological studies on AI moral judgments reveal that people often react similarly to mistakes made by machines and humans, experiencing a spectrum of negative emotions like anger and blame [10]. There is evidence that shows that when AI-inflicted harm is perceived as intentional, people often blame not only developers and companies, but also the AI itself [13]. The acceptance of AI is highly context-dependent [25], raising the question: would humans ever accept moral judgment from an AI?

B. The Question of Responsibility

The most discussed topic regarding AMAs, among our selected papers, is whether they can be held accountable for their decisions. Sullivan and Fosso Wamba identify three entities that can be held accountable: companies, as they are the representatives of the AI system they adopted; developers, as they manufactured and programmed the AMA; and lastly, the AMA itself, as it is the entity directly responsible for making a certain decision [13]. Responsibility also depends on whether a decision is perceived as intentional or accidental. People assign more blame to entities making intentional decisions with bad consequences, especially AI [13], as their decisions are usually 'truly' calculated.

Tigard defines three types of responsibility: normative (behaving in socially acceptable ways), possessive (having

a duty), and descriptive (being worthy of responses based on actions) [14]. While most papers only argue based on normative responsibility, Tigard notes that to truly hold an AMA responsible, we must accept that AI can have obligations and be held answerable for failing to meet expectations. Hallamaa and Kalliokoski argue that AMAs cannot possess reflective self-control and thus are unable to take moral responsibility [11]. This implies that responsibility still resides with other parties, and the AI is simply used as a tool.

Sharkey suggests that while an AMA can appear responsible, the company or developers are actually accountable [7]. This allows the AMA to learn from experiences and be rewarded or punished based on their decisions. However, for 'bad' moral decisions, people prefer retribution against individuals rather than non-living objects [12]. Nyholm argues that it does not make sense for AMAs to exist as anything other than tools at our disposal, as their autonomy could lead to consequences for others who had little to do with the decision-making process of the AMA [12]. Conversely, Hallamaa and Kalliokoski warns that holding AMAs fully responsible could make society overly dependent on AI, potentially reducing irrational human behavior but also leading to a loss of human control in the way society works [11].

Linking the question of responsibility to the notion of autonomy, Sharkey adds that while the decision-making process may be autonomous, human intervention is inevitable at some point during the creation or training of the AMA. She thus argues that the responsibility of those individuals cannot be offloaded onto the machine under any circumstances, as it is impossible for the output of the AMA to be completely independent of human intervention. She thus defines them as "moral entities", rather than "moral agents" [7].

Furthermore, current legal frameworks, such as the EU AI Act and various U.S. policy proposals, struggle to address AMA responsibility often due to a lack clear guidelines on attributing liability for AMA decisions that lead to negative consequences [26]. One potential regulatory solution is the implementation of "liability models", which may hold the programmers or companies accountable for the actions of their AMAs [27]. Future policy efforts may need to consider AI-specific liability insurance models or corporate responsibility mechanisms to prevent AMAs from serving as moral 'scapegoats' for decision-making failures. Clear guidelines and regulatory solutions are essential to effectively address the complex issue of AMA responsibility.

V. Reasons For AMAs

A. Inevitability

The integration of AMAs into morally complex domains, such as healthcare and autonomous vehicles, is often justified by their potential to handle tasks requiring rapid, data-driven decision-making with ethical implications [6], [8]. For instance, AI-assisted triage systems in healthcare help nurses determine the severity of patients' conditions quickly and accurately [28]. Judicial decision support systems, like COMPAS, assist judges in evaluating complex evidence and ensuring consistency [29]. Autonomous vehicles, such as those developed by Waymo, operate in morally complex environments where they must make real-time decisions to ensure passenger and pedestrian safety [30]. The success of these integrations underscores the need for AMAs, as they introduce efficiencies and reduce human error in ethically charged decisions [5]. To function effectively in these contexts, AI systems must evolve into AMAs, ensuring that decisions align with ethical standards and minimize harm.

However, this notion of inevitability is often criticized for its technologically deterministic perspective, overlooking governance structures, societal resistance, and cultural constraints [31]. Historical examples, such as prohibitions on lethal autonomous weapons, show that technological trajectories can be redirected by policy interventions, ethical concerns, and public discourse [32], [33]. Thus, the emergence of AMAs may not be as inevitable as some suggest, but rather subject to ongoing socio-political negotiations.

Currently, systems operated by multinational corporations like Google, Facebook, and X (formerly Twitter) significantly impact the economy, politics, and people's lives. Although these algorithms were not designed for moral decision-making, their consequences are profound enough to warrant consideration of AMAs to manage them. This is crucial given the opacity of these algorithms, even to their creators [11]. As AI continues to be implemented in new services, such as Google Search, the need for AMAs seems imminent.

Nevertheless, the inevitability of AMAs can still be questioned. Stringent regulations and ethical guidelines could limit and guide their deployment. For example, policies could ensure that human oversight remains a critical component in morally sensitive decisions, thereby preventing full autonomy of AMAs [6]. Ultimately, the future of AMAs depends not just on technological feasibility, but on societal and regulatory choices that will shape their trajectory.

B. The Prevention of Harm and Immoral Use

When thinking of the prevention of harm in connection with robots or AI systems, the first thing that comes to mind for many, are the Three Laws of Robotics, devised by science-fiction writer Isaac Asimov [34]. It is important to note that Asimov's Laws were never meant to be a real ethical framework, but rather a fictional device highlighting moral tensions [35]. Although inevitability alone does not justify the adoption of AMAs, it may serve as a compelling argument against actively opposing their development.

Cervantes et al. present a scenario in which a robot, acting as an AMA, takes care of an elderly person attacked by an armed thief in a park [9]. The robot faces two options: inaction, leading to harm for the elderly person, or action, disarming the thief, which might cause harm to the thief instead. While Asimov's first law does not specify which option to choose, the second option poses less risk and aligns more with the perceived 'correct' decision. In comparison, if no physical AMA is present or a human caretaker accompanies the elderly person, it is difficult to imagine a scenario where no human would come to harm.

In the context of self-driving cars, the potential harm minimization of AMAs can be further explored. Nyholm argues that while it is not clear what the exact "ethic settings" of self-driving cars should be, they have the ability to find the option that would cause the least harm to humans in the case of an unavoidable accident [12]. The same cannot be said for a human driver, who would not have the time to weigh the different possible options and outcomes against each other and thus come to an ethically sound decision.

To provide a more grounded approach, these considerations link with existing machine ethics models, such as value alignment theory or AI safety research. Value alignment theory emphasizes ensuring that AI systems behave consistently with human values and ethical principles [36]. This theory supports the idea that AMAs should align with societal norms and moral values, thereby minimizing harm in various scenarios. AI safety research focuses on identifying, measuring, and mitigating the risks of advanced AI systems to harness their potential benefits while preventing misuse [36]. By incorporating AI safety principles, we can ensure that AMAs are equipped to handle complex ethical dilemmas, such as the scenarios presented by [9] and [12]. This integration helps create a robust framework for the ethical deployment of AMAs, ensuring they can make decisions that minimize harm and align with human values.

Van Wynsberghe and Robbins present another interesting theoretical case [8]: A drunk woman enters her car to flee from domestic violence; however, the car performs an automatic breathalyzer test and only starts when the alcohol level is zero or low. Formosa and Ryan argue that in any case, an AMA that could adapt to and judge the situation would be better than simply having a "dumb car" [6]. While there is no obvious correct or false answer to whether or not the car should be able to start, having an AMA decide could prove to be safer overall, as it is able to consider all given information and come to a conclusion that will minimize the overall harm, without having to rely on emotional decisions.

VI. Discussion

When weighing the reasons against and for the further development of AMAs, it is challenging to determine a clear winner. Both sides bring up valid points, necessitating further evaluation to reach a conclusion. Indeed, only a continuous revision of the debate, constantly informed by and informing technological breakthroughs and conceptual and societal discourse, can lead to the answer.

A. Assigning Responsibility in AMA Decision-Making

Our results highlight the complexities in teaching correct ethics to machines and assigning responsibility. When assigning responsibility for AMA actions, there are four options: First, developers could be held accountable for creating the AMA in a way that led to the decision. However, this might be unfair in cases like self-driving cars that will crash regardless of the decision made, as developers cannot influence the actual crash situation. Second, the owner who chose to use the AMA and transferred their decision-making power to it could be held responsible. Since they would have been responsible without the AMA, using one should not allow them to circumvent responsibility but rather 'enhance' their decision-making. Third, consider the damage caused by an AMA's decision as no natural agent's fault, thus no person should be held legally liable. This might be seen as unfair to victims who had no influence over the decision but must deal with its consequences. Lastly, the state could pay for damages caused by an AMA's decision, financed by an 'AMA tax'. This would be feasible only when AMAs are widely spread and fully integrated into society and produce enough benefits to justify the tax and its amount. These options emphasize the need for a balanced approach to responsibility, ensuring fairness while promoting innovation.

Call to Action: Develop legal and ethical frameworks that clearly define responsibility for AMA decisions, ensuring accountability while carefully managing the pace and direction of technological integration.

B. Preventing Normative Lock-In and Promoting Moral Pluralism

A major risk associated with AMAs is the possibility of normative lock-in, where certain ethical perspectives become hard-coded into AI systems and constrain moral decision-making for generations. Unlike human morality, which evolves through cultural shifts and social debate, AMAs could institutionalize fixed moral hierarchies, potentially reinforcing dominant ideologies while marginalizing alternative ethical perspectives. This raises the critical question of whose morality is being programmed into AMAs and whether moral pluralism can be accommodated within algorithmic frameworks. A practical step to address this issue could be to conduct interdisciplinary workshops bringing together ethicists, technologists, and policymakers to explore ways to incorporate moral pluralism into AMA design. For example, [24] emphasize the importance of collaboration between ethicists and computer scientists to avoid overly abstract theories or faulty implementations. Developing guidelines for ethical programming that accommodate diverse moral perspectives as collaborative efforts can prevent ethical rigidity and promote inclusivity in AI ethics.

Call to Action: Foster interdisciplinary collaboration to design AMAs that reflect moral pluralism and remain adaptable to evolving ethical norms.

C. Shifting Perceptions of Moral Agency in the age of AMAs

The integration of AMAs could profoundly reshape societal conceptions of moral agency. As AMAs assume responsibilities traditionally held by humans, societal views of moral development and ethical behavior may shift. Delegating moral decisions to machines could redefine ethical expectations and moral failure within social structures. Initiating longitudinal studies to observe AMAs' impact on societal views on moral agency could inform policies and educational programs to prepare society for these changes.

Call to Action: Initiate long-term studies to monitor how AMAs reshape societal views on moral agency and guide adaptive policy-making.

D. Building Trust Through Transparency and Governance

Beyond technical performance, the social acceptability of AMAs depends on how trust is built and maintained in human-machine interactions. Trust is not just about reliability; it is about perceived legitimacy—who develops AMAs, who governs them, and how transparent their decision-making processes are. Without strong institutional safeguards and meaningful user involvement, AMAs risk being viewed as untrustworthy 'black boxes' rather than legitimate moral agents. Public resistance to self-driving cars, AI judges, and algorithmic hiring decisions suggests that trust in autonomous decision-making is deeply tied to issues of accountability and transparency. Building public trust in AMAs may require measures such as transparent AI decision audits, participatory ethics frameworks, and co-design approaches involving diverse stakeholders. Developing and implementing transparency standards for AMA decision-making processes, such as creating certification programs for AMAs that meet certain transparency and accountability criteria, could be a practical step to build public trust.

Call to Action: Establish transparency standards and participatory frameworks to ensure AMAs are trusted and perceived as legitimate moral agents.

E. Cultural Perspectives on Ethical AI Integration

The response to AMAs will likely vary across different cultures, influenced by existing ethical frameworks and societal norms. For example, in Japan, there is generally a higher level of trust in technology and robots, which may lead to greater acceptance of AMAs. In contrast, the EU emphasizes stringent data privacy and ethical standards, potentially leading to more cautious adoption. The U.S. often focuses on innovation and economic efficiency, which might prioritize rapid deployment over ethical considerations. A joint analysis of attitudes towards technology, ethical principles and theories, and the status of the discourse on AMAs in different countries would be instrumental in identifying and addressing cultural differences, accommodating multiple moral perspectives, and ensuring ethical pluralism.

Call to Action: Conduct cross-cultural research to design AMAs that respect and reflect global ethical diversity.

F. Power, Control, and the Political Economy of AMAs

The deployment of AMAs is not just an ethical challenge but also a question of power—who controls these systems, and whose interests do they serve? If major tech companies develop AMAs, they may embed corporate priorities into ethical decision-making frameworks, subtly prioritizing economic efficiency over moral nuance. Furthermore, AMAs could be strategically used by organizations to deflect responsibility, creating a 'moral buffer' between decision-makers and accountability. This raises concerns that AMAs could entrench power asymmetries rather than democratize ethical reasoning. Establishing regulatory frameworks to ensure that the development and deployment of AMAs are aligned with public interest rather than corporate priorities is essential. This could include setting up independent oversight bodies to monitor and evaluate the ethical implications of AMAs. Researchers are encouraged to contribute to this effort by developing and advocating for such frameworks.

Call to Action: Advocate for regulatory frameworks that ensure AMAs serve democratic values and do not entrench corporate power.

VII. Conclusions

Should machines take ethical responsibility? This paper has explored this question by examining what constitutes an Artificial Moral Agent (AMA) and evaluating two key arguments both for and against their further development. While the idea of machines bearing ethical responsibility is conceptually provocative and technologically ambitious, our analysis reveals that the answer is far from straightforward. The challenges of assigning responsibility, avoiding ethical rigidity, and ensuring cultural and societal alignment suggest that full moral autonomy for machines remains ethically and practically problematic.

As no definitive resolution exists, further research is essential to delineate the risks posed by AMAs and explore viable mitigation strategies. Future work should focus on developing robust frameworks for assigning responsibility and accountability for AMA decisions, exploring legal and ethical guidelines across different contexts and cultures, and investigating the potential for misuse of AMAs. Given the unresolved challenges of responsibility gaps and ethical rigidity, AMAs may be best suited as assistive ethical tools rather than fully autonomous moral agents. However, further research is needed to assess whether this approach sufficiently mitigates risks while enabling meaningful integration. This approach aligns with responsible AI principles [37], ensuring that AMAs enhance, rather than replace, human ethical reasoning.

The development and deployment of AMAs have significant practical implications. Policymakers must establish ethical and legal frameworks needed to govern the use of AMAs. Developers should prioritize transparency and accountability in the design and implementation of AMAs. Users need to be informed and educated about the potential risks and benefits of relying on AMAs for moral decision-making. To address these challenges, we propose several concrete next steps: conducting empirical research to understand the real-world impact of AMAs on decision-making processes and outcomes; developing comprehensive regulatory frameworks that address the ethical, legal, and social implications of AMAs; fostering interdisciplinary collaborations to explore the ethical dimensions of AMAs; and implementing educational programs to raise awareness about the ethical implications of AMAs among developers, users, and policymakers.

The rise of AMAs raises profound questions about moral agency, responsibility, and societal impact. As AI technology advances, it is imperative to address these questions thoughtfully and comprehensively to ensure that AMAs are integrated into society in a way that aligns with our ethical values and societal norms. This approach allows us

to harness the benefits of AMAs while minimizing the risks and ensuring they contribute positively to human well-being.

REFERENCES

[1] D. Ghosh, R. Ghosh, S. Roy Chowdhury, and B. Ganguly, "AI-exposure and labour market: a systematic literature review on estimations, validations, and perceptions," *Manag. Rev. Q.*, Jan. 2024. doi: 10.1007/s11301-023-00393-x

[2] R. Kabalisa and J. Altmann, "AI technologies and motives for AI adoption by countries and firms: a systematic literature review," in *Economics of Grids, Clouds, Systems, and Services*, K. Tserpes, J. Altmann, J. Á. Bañares, O. Agmon Ben-Yehuda, K. Djemame, V. Stankovski, and B. Tuffin, Eds., vol. 13072. Cham: Springer, 2021. doi: 10.1007/978-3-030-92916-9_4 pp. 39–51.

[3] G. Secundo, C. Spilotro, J. Gast, and V. Corvello, "The transformative power of artificial intelligence within innovation ecosystems: a review and a conceptual framework," *Rev. Manag. Sci.*, Nov. 2024. doi: 10.1007/s11846-024-00828-z

[4] S. Rawas, "AI: the future of humanity," *Discover Artif. Intell.*, vol. 4, no. 25, Mar. 2024. doi: 10.1007/s44163-024-00118-3

[5] A. Martinho, A. Poulsen, M. Kroesen, and C. Chorus, "Perspectives about artificial moral agents," *AI Ethics*, vol. 1, no. 4, pp. 477–490, Nov. 2021. doi: 10.1007/s43681-021-00055-2

[6] P. Formosa and M. Ryan, "Making moral machines: why we need artificial moral agents," *AI Soc.*, vol. 36, no. 3, pp. 839–851, Sep. 2021. doi: 10.1007/s00146-020-01089-6

[7] A. Sharkey, "Can robots be responsible moral agents? And why should we care?" *Connection Science*, vol. 29, no. 3, pp. 210–216, Jul. 2017. doi: 10.1080/09540091.2017.1313815

[8] A. van Wynsberghe and S. Robbins, "Critiquing the reasons for making artificial moral agents," *Sci. Eng. Ethics*, vol. 25, no. 3, pp. 719–735, Jun. 2019. doi: 10.1007/s11948-018-0030-8

[9] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos, "Artificial moral agents: a survey of the current status," *Sci. Eng. Ethics*, vol. 26, no. 2, pp. 501–532, Apr. 2020. doi: 10.1007/s11948-019-00151-x

[10] J.-F. Bonnefon, I. Rahwan, and A. Shariff, "The moral psychology of artificial intelligence," *Annu. Rev. Psychol.*, vol. 75, no. 1, pp. 653–675, Jan. 2024. doi: 10.1146/annurev-psych-030123-113559

[11] J. Hallamaa and T. Kalliokoski, "How AI systems challenge the conditions of moral agency?" in *Culture and Computing*, M. Rauterberg, Ed., vol. 12215. Cham: Springer, 2020. doi: 10.1007/978-3-030-50267-6_5 pp. 54–64.

[12] S. Nyholm, "The ethics of crashes with self-driving cars: A roadmap, I," *Philosophy Compass*, vol. 13, no. 7, p. e12507, Jul. 2018. doi: 10.1111/phc3.12507

[13] Y. W. Sullivan and S. Fosso Wamba, "Moral judgments in the age of artificial intelligence," *J. Bus. Ethics*, vol. 178, no. 4, pp. 917–943, Jul. 2022. doi: 10.1007/s10551-022-05053-w

[14] D. W. Tigard, "Responsible AI and moral responsibility: a common appreciation," *AI Ethics*, vol. 1, no. 2, pp. 113–117, May 2021. doi: 10.1007/s43681-020-00009-0

[15] ——, "Artificial moral responsibility: how we can and cannot hold machines responsible," *Camb. Q. Healthc. Ethics*, vol. 30, no. 3, pp. 435–447, Jul. 2021. doi: 10.1017/S0963180120000985

[16] K. E. Himma, "Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?" *Ethics Inf. Technol.*, vol. 11, no. 1, pp. 19–29, Mar. 2009. doi: 10.1007/s10676-008-9167-5

[17] L. Floridi and J. Sanders, "On the morality of artificial agents," *Minds and Machines*, vol. 14, no. 3, pp. 349–379, Aug. 2004. doi: 10.1023/B:MIND.0000035461.63578.9d

[18] J. P. Sullins, "When is a robot a moral agent," *Int. J. Inf. Ethics*, vol. 6, no. 12, pp. 23–30, 2006. doi: 10.29173/irie136

[19] M. J. Grant and A. Booth, "A typology of reviews: an analysis of 14 review types and associated methodologies," *Health Information & Libraries Journal*, vol. 26, no. 2, pp. 91–108, Jun. 2009. doi: 10.1111/j.1471-1842.2009.00848.x

[20] K. O. Arras, "How to conduct a literature survey and how to read a paper – a tutorial," in *Social Robotics Seminar*. Kai Arras Homepage, 2015, accessed: 2025–05-09. [Online]. Available: http://srl.informatik.uni-freiburg.de/teachingdir/ss15/SemSS15-tutorial.pdf

[21] S. Bruers and J. Braeckman, "A review and systematization of the trolley problem," *Philosophia*, vol. 42, no. 2, pp. 251–269, Jun. 2014. doi: 10.1007/s11406-013-9507-5

[22] C. Allen, I. Smit, and W. Wallach, "Artificial morality: top-down, bottom-up, and hybrid approaches," *Ethics Inf. Technol.*, vol. 7, no. 3, pp. 149–155, Sep. 2005. doi: 10.1007/s10676-006-0004-4

[23] B. McLaren, "Computational models of ethical reasoning: challenges, initial steps, and future directions," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 29–37, Jul. 2006. doi: 10.1109/MIS.2006.67

[24] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein, "Implementations in machine ethics: a survey," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1–38, Nov. 2021. doi: 10.1145/3419633

[25] H. Yamamoto and T. Suzuki, "Exploring condition in which people accept AI over human judgements on justified defection," *Scientific Reports*, vol. 15, no. 1, p. 3339, Jan. 2025. doi: 10.1038/s41598-025-87170-w

[26] J. K. C. Kingston, "Artificial intelligence and legal liability," in *Research and Development in Intelligent Systems XXXIII*, M. Bramer and M. Petridis, Eds. Cham: Springer, 2016. doi: 10.1007/978-3-319-47175-4_20 pp. 269–279.

[27] A. Bertolini and G. Aiello, "Robot companions: A legal and ethical analysis," *Inf. Soc.*, vol. 34, no. 3, pp. 130–140, May 2018. doi: 10.1080/01972243.2018.1444249

[28] H. Von Gerich, H. Moen, L. J. Block, C. H. Chu, H. DeForest, M. Hobensack, M. Michalowski, J. Mitchell, R. Nibber, M. A. Olalia, L. Pruinelli, C. E. Ronquillo, M. Topaz, and L.-M. Peltonen, "Artificial intelligence -based technologies in nursing: a scoping literature review of the evidence," *Int. J. Nurs. Stud.*, vol. 127, p. 104153, Mar. 2022. doi: 10.1016/j.ijnurstu.2021.104153

[29] Z. Xu, "Human judges in the era of artificial intelligence: challenges and opportunities," *Appl. Artif. Intell.*, vol. 36, no. 1, p. 2013652, Dec. 2022. doi: 10.1080/08839514.2021.2013652

[30] M. Hashim and P. Dehraj, "A survey on autonomous vehicles," in *Machine Intelligence and Data Science Applications*, V. Skala, T. P. Singh, T. Choudhury, R. Tomar, and M. Abul Bashar, Eds., vol. 132. Singapore: Springer Nature, 2022. doi: 10.1007/978-981-19-2347-0_21 pp. 277–292.

[31] A. Dafoe, "On technological determinism: a typology, scope conditions, and a mechanism," *Sci. Technol. Hum. Values*, vol. 40, no. 6, pp. 1047–1076, Nov. 2015. doi: 10.1177/0162243915579283

[32] G. E. Marchant, B. R. Allenby, and J. R. Herkert, Eds., *The growing gap between emerging technologies and legal-ethical oversight: the pacing problem*, ser. The International Library of Ethics, Law and Technology. Dordrecht: Springer, 2011, vol. 7. doi: 10.1007/978-94-007-1356-7

[33] B. C. Stahl, J. Timmermans, and C. Flick, "Ethics of emerging information and communication technologies: on the implementation of responsible research and innovation," *Sci. Public Policy*, vol. 44, no. 3, p. scw069, Sep. 2016. doi: 10.1093/scipol/scw069

[34] I. Asimov, "Runaround," *Astounding Science Fiction*, Mar. 1942, Publisher: Street & Smith.

[35] S. Chesterman, "Chapter 8: From ethics to law: why, when, and how to regulate AI," in *Handbook on the Ethics of Artificial Intelligence*, D. J. Gunkel, Ed., 2024. doi: 10.4337/9781803926728.00013

[36] I. Gabriel and V. Ghazavi, "The challenge of value alignment: from fairer algorithms to AI safety," in *Oxford Handbook of Digital Ethics*, C. Véliz, Ed. Oxford University Press, Dec. 2023, pp. 336—-355. doi: 10.1093/oxfordhb/9780198857815.013.18

[37] P. Akbarighatar, "Operationalizing responsible AI principles through responsible AI capabilities," *AI Ethics*, Jul. 2024. doi: 10.1007/s43681-024-00524-4