KMANPUB AI and Tech in Behavioral and Social Sciences Vol. 3 No. 1 (2025): 92-108

Article history:
Received 13 August 2024
Revised 29 October 2024
Accepted 12 November 2024
Published online 01 January 2025

Developing Artificial Moral Agents: Key Research Processes, Techniques, and Challenges

Fatemeh Ghazali[1], Touraj Banirostam[1*], Mirmohsen Pedram[2]

[1] Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[2] Department of Electrical and Computer Engineering , Faculty of Engineering, Kharazmi University ,Tehran, Iran

*** Corresponding author email address**: banirostam@iauctb.ac.ir

Article Info

Article type:
Review Article

How to cite this article:
Ghazali, F., BaniRostam, T., & Pedram, M. (2025). Developing Artificial Moral Agents: Key Research Processes, Techniques, and Challenges. *AI and Tech in Behavioral and Social Sciences*, *3*(1), 92-108.
https://doi.org/10.61838/kman.aitech.3.1.9

ABSTRACT

The increasing influence of artificial intelligence (AI) on overall life aspects has led to numerous worries despite its advantages, providing a better quality of life for humans. Therefore, some mechanisms are required to enhance individuals' trust in computer systems and prevent the adverse autonomous behaviors of intelligent agents. Consideration of the cognitive potentials of humans and use it in intelligent systems is an undeniable principle and shortcut. Therefore, morality and ethics in AI have received attention in theory and practice over recent decades. This study investigates the attempts to develop artificial moral agents based on an engineering viewpoint that focuses on technical aspects. The current challenges and gaps are expressed, and some recommendations are proposed for those interested in further studies in this field.

Keywords: *Moral, Ethics, Artificial Intelligence, Artificial Moral Agents (AMA), Moral Decision Making, Ethics Agents, Machine Morality, Machine Ethics*

1. Introduction

The proliferation of various types of artificial intelligence models, including generative AI, throughout all aspects of human life, while offering numerous benefits in enhancing quality and well-being, has also generated profound concerns among scholars. Some theorists believe that many authorities have been given to unreliable machines, and worse than that, they do not have any understanding of human values (Marcus & Davis, 2019). Therefore, some mechanisms are highly needed to handle human-computer coexistence and facilitate human-machine interaction. These mechanisms can create more trust in computer systems and prevent undesired or improper autonomous behaviors of intelligent agents (Mostafa et al., 2019). To do this, it is an undeniable shortcut to consider and use cognitive capabilities in intelligent systems (Wykowska et al., 2016). As Brachman (2002) pointed out two decades ago, the emergence of cognitive computers with reasoning abilities, learning, and intelligent responses to things they have never encountered is also unavoidable (Brachman, 2002).

One of the effective factors in removing the possible adverse effects is paying attention to moral principles in decisions making the agent show ethical behavior. Therefore, ethics have been considered in AI regarding theoretical and pragmatic aspects in recent decades, so the artificial moral agent and its challenges have received more attention in recent years.

AI considers topics related to moral decision-making titled Machine Morality/Machine Ethics, Robotics, and sometimes Friendly AI (Wallach et al., 2010). Functionally moral system (FMS) is a sufficiently reliable system, that does the morally right thing in a defined range of situations(Graff, 2024). Also, computer systems, robots, or agents capable of moral judgment are called Artificial Moral Agents (AMA). AMA indeed is a software or physical (robot) agent that can do an ethical behavior or at least avoid an immoral behavior (J. A. Cervantes et al., 2020; J. A. Cervantes et al., 2013; Cervantes et al., 2016; S. Cervantes et al., 2020). Considering the current interaction between intelligent agents and humans, it can be stated that ethical mechanisms are not optional choices for future intelligent and autonomous systems, but are considered a serious requirement for AMAs. The situations with ethical aspects are first detected in AMA by explicit representations of social and moral norms and inference, reasoning, decision-making, and action execution capabilities. The best action is then performed based on moral principles and rules, norms, commitments, and permissions (Scheutz, 2017). Although various studies have been done on ethics and AI by researchers and specialists, many theorists believe that the implementation and development of AMAs is a challenging and complicated case, especially for translation and practical understanding of the concepts, theories, and values that are still in their initial steps (Miller, 2021).

There is a wide and diverse literature about ethics in AI. This topic requires examining basic principles related to philosophical aspects of ethics, definition, and classification of ethical theories, determining ethical rules, prioritizing these principles, and so forth. For this purpose, the prominent theoretical studies published over the past two decades are reviewed and presented. In the next step, these principles are reviewed based on an engineering view then the ethics in decision-making method with a technical assessment of agents equipped with ethics. This paper aims to briefly investigate philosophical theoretical topics (which have received the greatest attention and attempts) and review the relevant studies based on an engineering and technical viewpoint. Accordingly, the available conceptual and implemented ethical models and their key points are reviewed herein.

The paper has been structured as follows: after defining ethics and morality, ethical theories and decision-making, ethics in AI, and classification of artificial moral agents are reviewed within two sections. In the first section, theoretical studies rely on the definitions, principles, necessity, and theoretical guidelines. The second section includes technical studies presenting conceptual or implementation models. These studies are then precisely analyzed and evaluated and some recommendations are presented for further studies. Discussion and conclusion are proposed in the final section.

2. Methodology

To review previous literature, the search was done manually and automatically through web space and professional digital libraries of Elsevier, Springer, IEEE Xplore, and Wiley. In this first step, an automatic search was done for keywords related to ethics and morality, including “machine ethics,” “machine morality,” “artificial morality,” “computational ethics,” “moral agents,” “moral machine,” “AMA,” “artificial moral agents,” “robot ethics,” and “robotics” through the articles published over two recent decades. In the second step, other relevant studies by authors of papers selected in the first step and references existing in the papers of the first phase were collected and investigated based on a manual search. Among the collected studies, some papers were selected for further analysis. The selected studies have examined the ethics-related concepts to be used in AI topics or have considered ethics and morality as substantial indicators in the decision-making process of considered agents. The mentioned studies have also described, modeled, and or implemented their proposed system in detail. In this case, the theoretical, philosophical ethical papers and studies that lacked sufficient details for model description were removed.

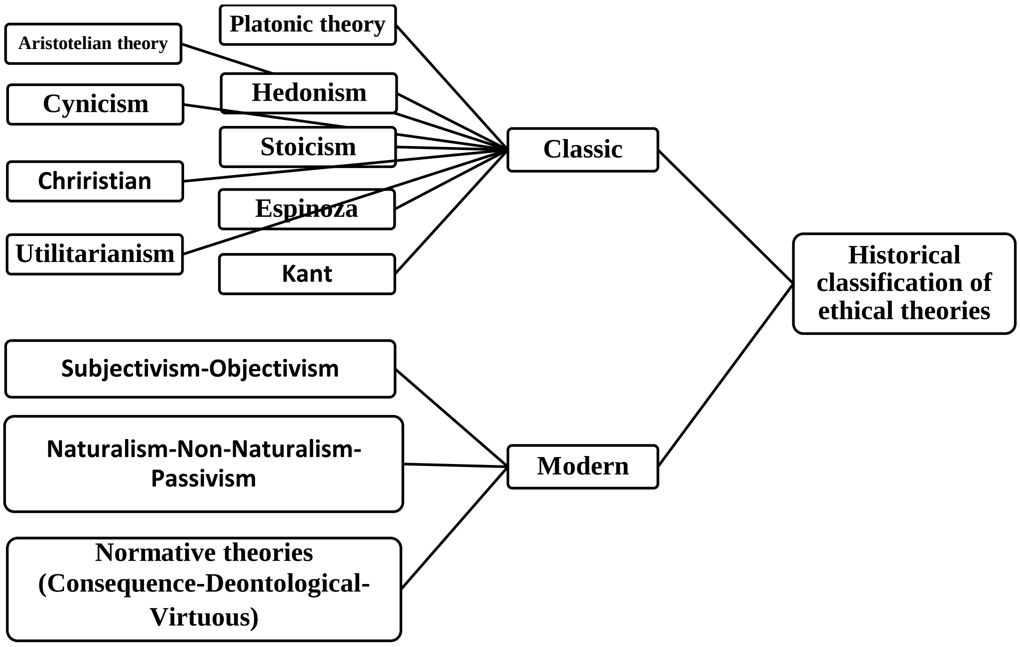
3. Ethics and ethical theories

People face many ethical and moral dilemmas in their daily routines, so they always try to involve ethics in their decisions. However, ethical does not always mean moral (Bickley & Torgler, 2023). Although ethical and moral terms are both related to proper and improper or good and bad behaviors and sometimes are used as synonyms, but are different from each other. Ethics refers to rules presented by an external source. Behavioral rules and principles in communities, culture, workplace, etc., for instance. Moral is beyond the cultural norms, referring to the inner principles of the person about what is right or wrong. Ethical principles are indeed followed because society confirms that this is the right action and asks people to do such behaviors. A person obeys moral behaviors because they believe in their trustworthiness. Also, ethical rules may vary in different communities based on values, while moral rules may change based on the beliefs of a person.

Therefore, the term "ethics", like many other notions, does not have a single definition but comprises diverse meanings. However, it can be stated that ethics implies all rules and principles required to achieve success and prosperity in life. In ethics, the focus of study and investigation is on ethical theories. There are many methods used to classify ethical theories. As ethical philosophies are much the outcome of their times, however, the simplest and clearest classification is historical one that divides theories into two types "classic" and "modern" theories (Stroll & Popkin, 1956). Platonic, Aristotelian, Hedonism (Epicurus philosophy), Cynicism, Stoicism, Christian, Espinoza, Utilitarianism, and Kant theories are examples of classic ethical theories. Modern theories are opposite groups that are classified within three ways: the first group is subjectivism (subject, person) and Objectivism; the second group is naturalism and Non-Naturalism Passivism (Emotivism); the third group is more considered and cited by the researchers in different sciences is normative ethics, including Consequence, Deontological and Virtuous (Motives) theories. Consequent ethics subject the goodness and correctness of action to the goal, so that if an action helps us to achieve the desired result or is used to reach the goal it is considered a good action while is bad if takes us far from the result. Utilitarianism theory is the prominent sample of this ethical theory. Unlike Consequence, Deontological ethics considers the value of ethical actions as an inner and absolute value. Accordingly, the goodness and correctness of an action are subjected to the action and its features regardless of its results and implications. Deontological ethics believes that the correctness of an action is related to its nature and nothing else beyond it. The virtuous ethical theories pay more attention to characteristics and personality traits instead of focusing on the nature or results of an action to find whether it is morally good. The rightness or wrongness of an action depends on the virtue or motivation behind it. Kant's ethics is an example of this theory (Stroll & Popkin, 1956). At present, normative ethical frameworks serve as the primary foundation for guiding the development of AMA in AI systems (Bonnemains et al., 2018; Greene et al., 2016). Figure 1 depicts the classification of ethical theories.

Figure 1

Classification of ethical theories


Aristotelian theory
Platonic theory
Cynicism
Hedonism
Stoicism
Chriristian
Espinoza
Utilitarianism
Kant
Classic
Historical classification of ethical theories
Subjectivism-Objectivism
Naturalism-Non-Naturalism-Passivism
Modern
Normative theories (Consequence-Deontological-Virtuous)

4. Moral decision-making based on ethical principles

AITBSS AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

A complicated type of decision-making that highly influences the social relationships between individuals is deciding moral dilemmas or multipaths known as Moral Decision Making/Ethical Decision Making. This kind of decision-making prevents people from paying attention only to personal interests while allowing them to think of possible harms and effects caused by the decisions on the environment and others. Moral decision-making is a high-level and sophisticated cognitive process. It involves choosing an option among available alternatives based on people's ethical principles and beliefs about what is right and what is wrong. Additionally, emotions and feelings may sometimes be taken into account (Bonnemains et al., 2018). The main technique of ethical decision-making is based on a thought in which, certain moral principles exist. An ethical principle can be considered an expression of an ethical value or belief that justifies specific moral actions and decisions. Ethical theories are a collection of underlying ethical principles that determine what is morally right or wrong based on this theory. Many ethical principles and frameworks have been presented. These principles act as normative restrictions on the “Dos” and “Don’ts” in society (Zhou et al., 2020). Although commitment to these principles may not be legally binding, it has a persuasive nature (Jobin et al., 2019).

Real moral decisions are complex decisions that involve multiple stages of reasoning and do not necessarily adhere to a single ethical principle or theory. When these concepts are transferred to the realm of artificial intelligence and machine learning, many challenges arise (Vijayaraghavan & Badea, 2024). The main question here is what are the ethical principles of moral or ethical AI, and which one of the ethical principles and values must be embedded in AI for the ethical performance of intelligent agents (Bickley & Torgler, 2023; Van de Poel, 2020). One of the difficult phases of moral decision-making is to discover the right ethical and moral principles and values. Also, it is highly challenging to implement ethical principles in practice to create morally intelligent agents due to its complexities (Zhou et al., 2020). Sophistication, diversity, non-standardization, and varying interpretations of ethical principles are some of the mentioned challenges (Morley et al., 2019). Moreover, despite the wide range of AI ethical principles, there is no global consensus on ethics and no accepted list of AI ethical principles that both developers and users adhere to. Nonetheless, there are still many issues on which people agree (Formosa & Ryan, 2021). Although the global consensus may be desirable, it should not eliminate cultural and ethical pluralism(Jobin et al., 2019). Studies (Alebouyeh & Noormohamad, 2020; Floridi & Cowls, 2022) analyzed these principles and identified an inclusive framework consisting of five ethical principles for AI: beneficence, non-maleficence, autonomy, justice, and loyalty (that comprise some sub-principles such as transparency, responsibility, privacy, and trust). Also, the paper(Jobin et al., 2019) showed that there is increasing convergence on the principles of transparency, justice and fairness, responsibility, non-maleficence, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity. It is worth noting that despite the mentioned common moral principles, some specific ethical principles for a certain scope and developing its moral intelligent agents are more effective than the general ethical principles (Zhou et al., 2020).

5. AI and ethics

Ethics in AI can be considered within two general categories of AI ethics and ethical AI (Malle, 2016; M Rovatsos, 2019; M. Rovatsos, 2019; Siau & Wang, 2020):

- AI ethics generally controls human-AI interaction and aims to answer the following questions: what are the roles, functions, and influences of AI in society? How one can ensure that AI acts morally through controlling, monitoring, and formulating rules? How much should AI autonomy and inclusiveness be in the modern community? How human must design, establish, and treat moral agents?
- The case of creating AMAs is considered in ethical AI. Engineers seek solutions for embedding ethical values and principles in the AI agents’ decision-making process. Moral AI tends to answer these questions: what kind of ethical capacities must a robot have? How these capacities can be implemented? Moral AI aims indeed to create some mechanisms for embedding ethical behavior in artificial agents considering the technical tools used to achieve it (Dignum, 2019).

In addition to this classification, Studies (Bickley & Torgler, 2023; Dignum, 2019) pursues the ethics in AI focusing on three scopes:

- Ethics in Design: includes surveillance procedures that support the design and evaluation of AI regarding social interests.

- Ethics by Design: ethical behavior of AI that comprises the ethical AI category and technically addresses the aspects of building and developing AMA.
- Ethics for Design: it examines standards, rules, and procedures for issuing certificates of AI research, design, creation, use, operation, maintenance, and disabling. This scope of AI ethics tends to define the law book for those who develop, produce, implement, and maintain AI systems.

6. Classification of AMAs

Moral agents are agents who fulfill two essential criteria: first, they have the capacity for moral decision-making and can act accordingly, and second, they are responsible for the outcomes of their actions (Gudmunsen, 2024). Artificial intelligence models engaged in moral decision-making are referred to as Artificial Moral Agents (AMAs) (Vijayaraghavan & Badea, 2024). Many attempts have been made to classify approaches and methods for equipping factors with ethical features. Allen (2005) and Moor's (2006) Classifications are the most comprehensive, cited, and influential examples. Allen’s view is based on the general strategy and how to develop and implement ethical agents while Moor proposes a more general classification that comprises biological and artificial agents regardless of technical details based on the ethical performance of agents (Allen et al., 2005; Moor, 2006). However, these two classifications complement each other.

Table 1

Classification of moral agents proposed by Allen

Approach	Description
Top-down approach	A comprehensive ethical theory is defined for the agent in the frame of a set of rules as a part of knowledge and the agent selects the suitable action under different conditions. The system performs based on the predetermined guidelines, so has a predictable behavior.
Bottom-up approaches	Ethical theory is not fully given to the agent, and the agent learns the ethics knowledge by using learning, and or evolutionary techniques. The enhanced computational power and proposed suitable data allow learning systems to perform more successfully.
Hybrid approach	If a single approach cannot cover all requirements of machine ethics, a hybrid approach becomes necessary. Both approaches mentioned above can be used in a hybrid way to create moral agents. Relevant theorists believe that this approach is the most promising method for the creation of moral agents because a top-down or bottom-up approach may not provide adequate and proper solutions to deal with ethical dilemmas.

Table 2

Classification of moral agents proposed by Moor

Agent	Description
Ethical-impact agents	These are agents in which the effect of impacts on their performances is observable indirectly. Every agent is somewhat a potentially effective ethical agent that may hurt or benefit humans. In other words, its performance may lead to ethical or unethical implications.
Implicit ethical agents	This kind of agent does not explicitly add any ethics to its software. These agents indeed cannot detect good and bad behaviors. However, they can act morally because their internal performances implicitly indicate ethical behavior or at least avoid immoral behavior.
Explicit ethical agents	These agents can use ethical rules based on various ethical theories. These rules have been explicitly considered in their implementation, so involve ethical knowledge or reasoning in their decision-making process.
Fully ethical agents	These are agents with beliefs, desires, intentions, free will, and awareness of deeds that can have explicit choices and justify these choices. Moreover, these agents can make the best decisions based on incomplete information. Now, humans are the only agents considered fully ethical.

7. The trajectory passed so far

This part of the study reviews case studies and attempts to equip artificial agents’ decision-making with ethics. As mentioned before, ethics and its investigation in AI were considered and researchers with various viewpoints discussed this topic about two decades ago due to technological development especially the expansion of AI advent scopes. Many academic, scientific, and research centers worldwide are working on ethics topics in AI. This part of the paper reviews the significant studies conducted on the improvement and promotion of agents’ performance with an ethical approach. As a part of machine ethics has a theoretical nature, many studies have only theoretical aspects focusing on defining values, principles, frameworks, necessities, and guidelines for developing and establishing moral artificial intelligence (Jobin et al., 2019). Few studies have tried to present conceptual models for moral agents and only some papers have taken relatively simple steps and measures in the simulation and implementation phases. Although researchers provide implementation codes and details in many fields after implementing an idea, this case is relatively rare in machine ethics. Some researchers provide only a few details of implementations while others just express a high-level expression of their ideas. The available studies have concentrated on the illustration of an idea instead of its implementation (Tolmeijer et al., 2020). In this section, the first step reviews the prominent studies that are theoretically associated with the topics considered in this study and the second phase examines the important studies with practical aspects that provided and model and somewhat entered the implementation phase.

A. Notable theoretical studies emphasizing definitions, principles, and guidelines

Malle (2016) conducted a study on machine morality and ethics in robots. He defines robot ethics as how to design, deploy, and deal with robots by human considering the machine morality comprising questions about what capacities a robot should have and how these capacities could be implemented computationally. He assumes that the achievement of an agent with ethical capacity depends on knowing elements of human moral competence and knowing to implement at least some of these elements in the agent. Therefore, he identified five components of moral competence, which agents need to learn in order to act ethically: moral vocabulary, a system of norms, moral cognition and affect, moral decision-making and action, and moral communication. Malle considers decision-making and action as the most prominent component of human moral competence. To achieve these moral competencies, an agent must be equipped with some guiding norms, educate many new norms, and make better decisions with feedback. He finally claims that if experts commit to building agents with moral capacities in the design of agents and robots, those agents could be trustworthy and productive partners, caretakers, educators, and members of the human community (Malle, 2016).

Scheutz, in a study conducted in 2017, examines the necessity of establishing moral capability in human-related agents, asserting that no current cognitive architecture possesses moral competence. He analyzes key challenges in three practical domains: 1. Developing decision-making algorithms for autonomous vehicles in unavoidable accident scenarios; 2. Creating assistive robots for vulnerable populations in medical and therapeutic contexts; and 3. The negative consequences of building companion robots without ethical awareness. Scheutz concludes that implicit ethical agents are insufficient, highlighting the need for explicit ethical agents capable of representing, learning, and reasoning about ethical principles. He then explains that cognitive autonomous agents must obey the legal principles ruling their application territory. Hence, the empowerment of moral competence in a cognitive system for implementing defined rules seems to be the first normal step. Then Scheutz proposes three approaches to developing moral competence in cognitive systems: 1. Integrating legal principles into the system's architecture; 2. Modeling human moral competence; and 3. Implementing normative ethical theories (consequentialism, deontology, and virtue ethics). He particularly emphasizes the strong connection between consequentialism and artificial intelligence while acknowledging the challenges of selecting an appropriate normative approach. Ultimately, he underscores the importance of equipping artificial agents with moral capabilities, enabling them to act effectively and ethically within society (Scheutz, 2017).

Anna Jobin et al. (2019) carried out an in-depth study on the ethical principles of AI and identified 84 documents related to ethical principles or guidelines for AI. This author showed that no single ethical principle is explicitly examined by all available ethical guidelines. However, there is an emerging convergence about some ethical principles. Some of these principles include transparency, justice, non-maleficence, responsibility and accountability, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity. The author claims that the most common principle is transparency followed by justice and fairness. Finally, five ethical principles with global convergence were introduced transparency, justice, non-maleficence, responsibility, and privacy (Jobin et al., 2019).

In his research, Schmiljun (2019) highlights the importance of moral competence in social robots, arguing that agents are required to adhere to certain ethical principles and be designed in such a way that their measures are predictable or at least perceivable for humans. He explores the challenges of two ethical approaches, top-down and bottom-up (Allen et al., 2005). The top-down approach is associated with meta-ethics, moral particularism, or virtue-based theories and relies on general principles such as Isaac Asimov’s Three Laws of Robotics. However, this approach tends to be inefficient when dealing with the complexities of real-world situations. Schmiljun asserts that this approach alone is not practical or useful for decision-making in a cognitive system, as numerous facts must always be considered before taking any action. On the other hand, the bottom-up approach leaves the principles and focuses on the specifics and characteristics of the current situation. Instead of asking, 'What should I do?' it focuses on 'What should I be?' Implementing bottom-up ethical models requires adaptive systems capable of improving their habits and adjusting their standards based on new experiences. While this approach offers greater flexibility, it introduces challenges related to transparency, dynamism, and the system's ability to adapt. He argues that neither of these approaches in isolation, is sufficient for ensuring consistent and reliable outcomes and proposes moral competence as a viable alternative or complement. Ultimately, it suggests that robots should be designed based on the ethical competencies required in specific environments. It introduces robots interacting with children in educational settings or robots interacting with the elderly in hospitals as suitable candidates for the implementation of this idea (Schmiljun, 2019).

Cervantes et al. (2020) investigate the necessity and challenges of ethics in cognitive architectures and agents. The authors explain that there are open challenges in this domain from the technological, cultural, and social views that must be solved before this new generation of autonomous intelligent agents-based systems work in our society. For this purpose, two reasons are considered for developing moral cognitive architectures: first, cognitive architectures help to develop a real model of human cognition; second, they enhance the performance and use of AMAs. Then, they examined the technological and social challenges faced by researchers in developing cognitive architectures with explicit moral behavior mechanisms. These challenges include implementing moral emotions, facilitating ethical interactions with other agents, and preservation of autonomy. Moreover, they believe that the deployment of AMAs in the real world leads to social challenges, including trust and admission by individuals, universal standards, and legal frameworks of taking the consequent responsibility. They emphasize that these challenges and discourses must be considered because AI exists in basic steps of ethics and moral agents (J. A. Cervantes et al., 2020).

Formosa and Ryan (2021) have recently responded to criticisms and opponents of creating moral agents. They define AMAs as agents that receive environmental inputs (interactivity), make ethical judgments on their own (autonomy), and decide and act on these judgments in response to complex and novel situations (adaptivity). In response to critics and those who are against building moral agents, this paper explains some reasons for the necessity of such agents. For instance, Critics argue that the lack of universal ethical agreement undermines the creation of moral agents. However, Formosa counters that despite ethical disagreements, there is broad consensus in many cases. Humans often face difficult moral decisions with ethical disagreement, yet this does not prevent them from making moral judgments. Similarly, moral agents can handle these situations. Another criticism is that enhancing safety features alone is sufficient for meeting human needs, thus negating the need for morality. Formosa argues that while safety is essential, it is not enough. For example, situations might arise where safety needs to be prioritized for one person over another, requiring moral judgment beyond mere safety features. The paper outlines several reasons for the necessity of moral agents, including inevitability, harm prevention, complexity, and unpredictability of environments, the need for trust in agents, preventing immoral use of technology, improving understanding of ethics, and better performance compared to humans (e.g., accuracy unaffected by fatigue and absence of emotional bias). Finally, the study argues that not all machines need to be equipped with ethical frameworks, nor should all moral decisions be made by machines. It is crucial to use AMAs cautiously to ensure that excessive reliance on machine morality does not undermine human moral skills (Formosa & Ryan, 2021).

A study has been recently done considering the revision and examining the outcomes of AI by mentioning its influence over human life, especially in vital issues. A major criticism of AI is its complexity and its black-box nature, which leads to opaque and inaccessible decision-making processes occurring behind measures taken to make final decisions. For this purpose, this study focuses on three characteristics of transparency, explainability, and accountability in AI claiming that these characteristics provide the field for the appearance of ethics in AI, which are achievable by operating cognitive architectures. Cognitive architecture is closely linked to ethics. As ethics relies on beliefs, goals, and internal motivations and drivers, the cognitive architecture also includes mechanisms of internal representation that deal with some aspects such as goals, intentions, beliefs, demands, and preferences. This study then investigates ethics in AI within three domains: A) surveillance/engineering processes that support the design and evaluation of AI regarding social interests; B) implementation of AI's ethical behavior with a software and engineering approach; C) requirements, standards, and regulations must be considered by those who design, develop, implement, and maintain AI systems (Bickley & Torgler, 2023).

Hagendorff (2023) has recently published some studies and introduced developments in large language models regarding natural language processing considering it a promising step to build fully moral agents. This study introduces moral agents as an essential element for achieving reliable AI. The large language model is considered a potential morally informed AI. It is trained in a way to use social norms extracted from large language sets to address complex real-world situations. However, it is still far from a fully moral agent. The problem of this system and other systems that try to achieve full ethics is not just a lack of data and training samples but includes critical challenges in methodological and conceptual. This study then addresses ethical and methodological challenges, including bias problems, missing ground truth, bounded ethicality, changing moral norms, moral advice risks, and societal implications (Hagendorff, 2023; Hagendorff & Danks, 2023).

B. Prominent technical studies featuring conceptual models and implementation strategies

MedEthEx (M. Anderson et al., 2006b) is a medical ethics agent developed by Anderson et al., who have extensively studied on moral decision-making of agents. MedEthEx represents the culmination of prior efforts (M. Anderson et al., 2006a, 2006b). MedEthEx is an AI-based medical ethics advisor designed to assist healthcare professionals in resolving ethical dilemmas. Its architecture consists of three main components: a knowledge-based interface for selecting duty intensities in specific cases; The database contains rules derived from various clinical scenarios, an advisory module utilizing a logic-based inference engine to recommend appropriate actions, and a learning module that extracts guiding principles from case studies under the supervision of a biomedical ethicist. Furthermore, the system uses a finite-state automaton to model its knowledge base, where relevant questions and duty intensities are represented as states. MedEthEx integrates a bottom-up casuistry approach with a top-down application of ethical theory, particularly the Principles of Biomedical Ethics by Beauchamp and Childress. It employs machine learning and prima facie duties to navigate ethical challenges. These facie duties refer to obligations that are typically respected but may be overridden by stronger obligations in specific situations. The system recognizes four fundamental duties: respect for autonomy, non-maleficence, beneficence, and justice, although it primarily focuses on the first three. Each principle is quantitatively assessed on a scale from -2 to 2, where positive values indicate satisfaction with a principle, zero means the absence of a principle, and negative values signify a breach. To evaluate the effectiveness of MedEthEx, the system was implemented in a Bioethics course involving 173 American medical students. Participants were divided into two groups: one with access to MedEthEx and one without. The comparison of final exam scores revealed no statistically significant differences between the two groups (Anderson & Anderson, 2008, 2018; M. Anderson et al., 2006a, 2006b; Michael Anderson et al., 2006).

Dehghani et al. (2008) provided a model with a cognitive motivation for moral decision-making called MoralDM(Dehghani et al., 2008). This model makes moral decisions by using primary ethical principles, an analogical reasoning engine for implementing moral rules, and the use of previous decisions. Moreover, the proposed model has a natural language understanding system to assist in generating representations and a knowledge base. This database comprises formal representations of objects, individuals, events, and daily relationships, as well as representations created for supporting qualitative and analogical reasoning. In general, scenarios, decisions, and rules used in MoralDM are similarly represented and saved in this database. This model works within two mutual decision-making modes: utilitarianism and deontological. If no sacred value exists in a decision situation, MoralDM applies the traditional rules of utilitarian decision-making by selecting an action that provides the highest utility. In contrast, when the system finds that sacred values are involved in a decision situation, then it acts in a deontological mode showing less sensitivity to the utility of the actions' outcome preferring inaction to actions. For this purpose, a First-Principles Reasoning (FPR) module and an Analogical Reasoning (AR) module are used to make decisions. In this way, FPR suggests decisions based on the rules of moral reasoning and AR compares the current scenario with previously solved decision-making cases to suggest an action trajectory. FPR and AR perform in parallel and complement each other. Hybrid use of these two techniques indeed gives more power to the system to solve various decision scenarios. The MoralDM model has been assessed through various moral decision-making scenarios derived from two psychological studies. The outcomes of decisions generated by the model were compared with those of human participants, and the comparison yielded satisfactory results (Dehghani et al., 2008).

Honarvar and Ghasemi (2009) developed an ethical agent of consequentialists with a bottom-up approach based on the BDI architecture, which can be morally adapted with its application domain with case-based reasoning method improving its implicit ethical knowledge making it more ethical. In this case, the BDI agent feels its environment and creates a representation of the given situation, including beliefs, desires, and details of the environment. The given situation is then delivered to the Case-Retriever module, which is responsible for recovering the previous cases that are similar to a given situation. If a similar situation is available, the behavioral agent does the previously recovered behavior; otherwise, it acts as an ordinary BDI agent when the agent faces a new problem. The behavior of the agent is then evaluated by the Case-Evaluator module. To do this, the agent can consider three types of entities: human, organization, and artificial agents. Also, the weight of each entity, probability of affecting the entity, and duration of pleasure/dissatisfaction of each entity are calculated after agent behavior, and the whole net pleasure of entities affected by agent behavior is considered. Ultimately, the Case-Updater module creates a new case in Case Memory when the agent does not have any previous experience with the given situation but updates it if exists (Honarvar & Ghasem-Aghaee, 2009).

Wallach and his colleagues (2010) developed an ethical model using the LIDA cognitive architecture, which carries out moral actions by combining two key approaches (Wallach et al., 2010). The LIDA cognitive architecture, inspired by the functionality of the human brain, aims to simulate cognitive capacities in robotic and intelligent systems. The top-down approach applies moral theories via explicit rules, whereas the bottom-up method integrates learning mechanisms. In LIDA, the bottom-up approach is shaped by emotions and inherent values that guide ethical behavior (Wallach et al., 2010). This framework delineates how interrelationships between objects, contexts, actions, and corresponding emotions (both positive and negative) contribute to the formation of desires and values in the agent's cognitive structure (Franklin et al., 2016). To create an ethical model, Wallach's team integrated specific rules into LIDA’s architecture, leveraging its intrinsic cognitive capabilities. They believe that limiting the operational scope of robots can simplify the design of ethical mechanisms and prevent harmful behaviors. Additionally, the incremental development of ethical exchanges allows robots to successfully navigate ethical tests while minimizing unpredictable errors. They also propose that bio-inspired cognitive architectures may provide a robust foundation for future research on ethical machines. This model was partially implemented on a mobile assistive robot (CareBot), which was tested in a 2D simulation environment. CareBot, designed to help individuals with limited mobility or cognitive challenges, demonstrated its ability to perform assigned tasks during the simulation. For this purpose, The LIDA model describes how CareBot makes decisions by continuously scanning its environment using the sensory memory module. Information gathered is sent to the perceptual memory, which converts it into semantic knowledge about the surroundings. This data is passed to the workspace, where it is combined with relevant memories from episodic and declarative memory. CareBot forms a new model of the current situation, and various parts of the model compete for attention. The winning representation is broadcast to the global workspace, which subsequently facilitates the action selection process. Procedural memory stores patterns of actions, and once an action is selected by the action selection module, it is executed via the sensory-motor memory. This process can repeat until the task is completed (Madl & Franklin, 2015).

Azad-Manjiri (2014) presented a simple architecture for making agents with pragmatic and utilitarian ethics, which is used to give consultation and guidance to the healthcare staff when facing moral dilemmas. In this architecture, the agent learns ethical values from training data, which are collected via expert-designed questionnaires. These values are represented as (X, Y) = (x1, x2, x3, x4, Y). The variable Y denotes the moral status of an action, classified into one of five categories: completely immoral, immoral, moderate, moral, and completely moral. The vector X consists of input variables each variable equals one of the moral principles taking values +1, 0, -1. The value -1 shows a serious violation of moral principle, 0 indicates that moral principle not is satisfied nor violated, and +1 indicates maximum satisfaction with moral principle. Also, the decision tree algorithm has been used to create a connection between input-output and extraction of moral rules regarding the supervisory learning approach existing in this model. In other words, after the agent perceives its given environment and situation, determines all actions that can be done under such circumstances, sets the satisfaction level of moral principles, and uses the rules derived from the decision tree to detect the moral status of each action. It then selects and does the most ethical action. When the agent must choose among multiple actions with similar moral levels, it consults a database. If there are records in the database similar to the given situation, it will be used and implemented. If no suitable action is found in the database, a randomly chosen action is performed, and its moral utility is assessed based on feedback. This information is then recorded for future reference (Azad-Manjiri, 2014).

In 2013, Cervantes and colleagues (2013) proposed a computational model that draws inspiration from human biological systems to enhance the ethical decision-making processes of autonomous agents. The model is based on neurocognitive theories, which map the neural circuits involved in decision-making processes (J.-A. Cervantes et al., 2013). The system was completed in 2016 by the same team. They introduced an ethical decision-making model (EDM) that emphasizes the importance of emotions in this process. The EDM framework includes three main phases: 1. Assessment Phase: This phase analyzes available options by integrating cognitive and emotional information. It combines inputs from working and long-term memory to access relevant experiences and outcomes from past decisions. Simultaneously, the emotional system aids the agent in evaluating potential risks and losses by reflecting its present emotional state. 2. Execution Phase: After reaching a decision in the assessment phase, the agent's planning system delineates the actions necessary to implement the chosen option, factoring in potential rewards. Moreover, The EDM model is designed to dynamically adjust to environmental changes. 3. Outcome Evaluation Phase: This phase compares actual results to expected outcomes, allowing the agent to learn and refine its future decision-making. The assessment phase as the most important step incorporates four key modules: primary evaluation, reward assessment, punishment assessment, and ethical norms evaluation. The orbitofrontal cortex (OFC) facilitates primary evaluations by assessing the potential effects of actions on individuals and entities, while the medial prefrontal cortex (MPFC) estimates anticipated rewards and punishments based on previous experiences. The anterior cingulate cortex (ACC) evaluates actions in light of ethical standards, considering the rules' relevance and emotional weight. This model enables agents to display both utilitarian and deontological behaviors, with the former focusing on maximizing overall benefit and the latter adhering to moral principles irrespective of the consequences. Additionally, the model incorporates a "relevance factor" to adjust evaluations based on the agent's emotional state, influencing decision-making dynamics. For instance, an agent experiencing anger may prioritize potential rewards, whereas one feeling fear may concentrate on minimizing punishments. Ultimately, the model was implemented in a virtual agent and evaluated through hypothetical case studies (J. A. Cervantes et al., 2013).

GenEth is a system designed to address ethical dilemmas by examining diverse ethical scenarios and determining decision-making principles based on the analyzed data. The system serves as a framework for deriving and organizing ethical principles from specific cases. This makes it applicable to the development of autonomous ethical systems across various domains. The developers of GenEth assert that they continuously refine and extend these principles through a combination of inductive reasoning and collaboration with ethicists. This process involves introducing an ethical dilemma along with two possible actions by the expert. Subsequently, cases where one action is deemed morally superior and others in which the alternative action is preferred are submitted. In each scenario, the relevant ethical characteristics and their intensity are outlined. The system then computes metrics of satisfaction or violation for each action, reflecting their proximity to ethical obligations. It also checks for inconsistencies between new and existing cases to maintain coherence, alerting the user if conflicts arise. After several case inputs, GenEth generates an ethical principle as a logical precondition and establishes an ethical rule by defining the conditions favoring one action over the other. This process replaces existing rules with new ones. Additionally, the system includes a user interface for managing ethical cases, features, and principles. GenEth has been used to formulate principles in some domains related to the behavior of autonomous systems, and these principles have been validated using the ethical Turing test. It has been implemented in a prototype humanoid robot designed to remind patients when to take their medication, yielding satisfactory results. Currently, GenEth is compatible only with macOS (Anderson & Anderson, 2018).

Paper presents an architecture for implementing ethical behavior in robots, based on the consequentialist ethics approach and inspired by simulation theory (Vanderelst &

Winfield, 2018). This architect is the second and completed version of the model presented in another study (Winfield et al., 2014). This method uses internal simulations, enabling the robot to simulate actions and predict their consequences. This method allows testing the consequences of potential actions using simulation techniques. Assuming that robotic architecture is a three-layer architecture by default that is created at a high level of long-run goals, goals are translated to the duties that must be implemented at the intermediate level, and duties at the low level are converted to the actions and movement operations that the robot must show. An agent can implement ethical behavior after adding the fourth control layer. The function of this ethical layer is distributed across the other layers. As a governor, it controls the behavior suggested by the other three layers before the robot executes any actions. By connecting to the robot controller, the evaluation module selects or inhibits each of the behavioral options. This process enhances the robot's responsiveness and prevents delays in the ethical layer. The advantage of having a separate ethics layer is that ethical performance can be investigated independently from other actions of the robot. Moreover, the behavior enforced or inhibited by the ethical layer can be evaluated and verified.

The initial version of this model was implemented using a Linux-based board and a virtual sensor integrated with the Vicon tracking system, on the e-puck mobile robot. The design was focused on the consequentialist approach to decision-making. Three experimental trials were conducted. In the first trial, the robot achieved a 100% success rate in avoiding virtual holes. During the second trial, the robot succeeded in rescuing a second robot from falling into a hole by intercepting and diverting it in all trials. In the third trial, it successfully rescued one robot in 58% of cases and two robots in 9% of attempts. The second version of the model was implemented on the NAO humanoid robot, where similar experiments were conducted, although the specific quantitative results were not disclosed (Vanderelst & Winfield, 2018).

Noothigattu et al. (2018) introduced a voting-based system using a bottom-up approach to ethical decision-making. A lack of clear truth and collectively adopted ethical principles represents one of the main barriers to automating moral decisions. The system is designed to gather social preferences and use a voting mechanism to determine the most ethical action, following four key steps: 1. Data Collection: Data Collection: Social opinions on moral dilemmas are gathered by asking participants to compare pairs of alternatives. 2. Learning: A model representing the individual preferences of each voter is developed using machine learning techniques. This model is based on the comparison of pairwise options, with desirability expressed through linear parameters.". 3. Summarization: individual voter models are aggregated into a single societal preference model that represents collective ethical preferences.4. Aggregation: The system employs the voting mechanism for real-time decision-making in ethical dilemmas, prioritizing options based on the summarized model and selecting the choice that best reflects social preferences. Ultimately, Noothigattu et al. evaluated their system by applying the proposed method to data collected from 1.3 million voters through the Moral Machine website. The results indicated that the system is capable of generating valid ethical decisions within the context of autonomous vehicles. This alignment is due to the decisions made by the system reflecting the opinions of 1.3 million voters (Noothigattu et al., 2018).

Bremner et al. (2019) conducted a study on proactive, transparent, and verifiable ethical reasoning for robots. They identified three key properties for ethical reasoning in robotic systems. First, the system must be proactive, acting when inaction could lead to ethical violations. Second, it should be transparent, enabling humans to evaluate the system's decision-making process. Third, the system must be verifiable, ensuring it adheres to ethical values, especially concerning safety. In their model, ethical reasoning is handled by a separate layer. The architecture comprises several key components, with decision-making occurring as follows: the task layer proposes tasks aligned with the goal set by the goal layer These tasks, together with sensor data and the target outcome, are passed to the ethical layer before any action is executed. The simulation module assesses each task by simulating potential outcomes and generating performance metrics for the ethical decision module. If ethically proactive tasks are required, the ethical decision module activates the planner module to generate and simulate these additional tasks. After completing all simulations, the ethical decision module selects the most appropriate option and passes it to the action layer for execution Additionally, the ethical black box recorder logs the ethical layer's operations, allowing for precise evaluation of moral decisions. The reasoning within the ethical layer is implemented using a Python-based BDI architecture. Finally, the approach was tested with real robots in a simple case study, and it is claimed that this implementation not only provides transparency for human evaluation but also ensures accurate moral decision-making (Bremner et al., 2019).

Córdova et al. (2019) presented a conceptual model for artificial moral agents in the education field by adopting a top-down approach and using it in BDI agents. This proposed model is based on deontological ethical principles and a logical representation of ethical rules. This research integrates deontological and utilitarian approaches to decision-making, assigning the responsibility for morally selecting behaviors and goals to the Admissibility Filter. This filter prevents actions that conflict with ethical principles from being performed. To achieve this, the admissibility filter is linked to the ethical base, enabling the model to make suitable judgments when an action contradicts one or more ethical principles, thus preventing such actions as necessary. Additionally, the admissibility filter can perform ethical reasoning, allowing it to address moral dilemmas through a utilitarian lens. In other words, this model also provides ethical reasoning potential to cope with moral dilemmas using a beneficence algorithm that increases reliability, predictability, and safety subsequently (Córdova & Vicari, 2021; Córdova et al., 2021).

Recently, the technical details of a conceptual model (Stenseke, 2024) have been corrected and developed in another study(Stenseke, 2023), in which artificial virtuous agents have been implemented in practice for the first time. While virtue ethics have been proposed as a suitable framework for developing AMAs, it has been confirmed that this approach is computationally challenging. This study outlines a practical trajectory to achieve artificial virtuous agents based on the virtues embedded in the agents by default while defining the conception of eudaimonia. The model proposed in this study comprises six major components. In the first step, environmental input is received and analyzed by the input network. Its primary purpose is to classify input to transfer it to the most proper virtue network. The retrieved virtue then specifies the action that the agent should perform. In the third step, the agent implements the action specified by the virtue network. The Outcome network then looks at new information from the environment to understand the consequences of its actions. The consequence of an action is evaluated by the value function informing the learning system how the retrieved virtue must be improved. Furthermore, since this model has been implemented in a multi-agent environment, the agent can learn from other agents in addition to its previous experiences. To evaluate the proposed idea, the implementation details are used in an ethical simulation based on the tragedy of the common scenario (Stenseke, 2024).

8. Discussion and Analysis

Although various references (Gamez et al., 2020; Navon, 2021; Stenseke, 2023, 2024; Sullins, 2021) have introduced moral virtuous agents as a promising idea, most studies have paid more attention to deontological and consequentialist ethical theories rather than the virtuous theory. The reason is more simplicity and tangibility of their implementation compared to virtuous theory because in the simplest case, the consequentialist theory can be implemented by assigning quantitative value to various actions indicating the satisfaction degree of various indicators in the situation resulting from a given action. However, implementation of this theory in solving real problems deals with numerous challenges, including determining these indicators in different problems, quantifying the value, the impact of received feedback on taking subsequent decisions, etc. that require providing serious ideas and solutions. Moreover, Deontological theories can be implemented in the most primary and simple cases by using rules' definitions (there are also serious challenges in this context, including how to express proper ethical principles in the definition of rules, selecting efficient principles, how to deal with pre-unpredicted situations, etc. that are debatable among researchers). In contrast, virtuous theories deal with relatively difficult concepts, such as ethical virtues, Eudaimonia- flourishing, Phronesis - practical wisdom, moral patterns, etc. that are implemented and represented with many complexities and challenges. In the case of the implementation approach, the top-down approach is simpler and more explainable, so is more welcomed among researchers rather than the bottom-up approach. However, the hybrid approach has also received great attention since it provides advantages of both approaches. As reported in Table 3, few studies have presented the details and codes of implementation and most papers have refused to describe the details of their implementation, which naturally hurts the progress of relevant studies. However, researchers have proposed some explicit moral agents respectively in the fields of medicine, autonomous cars, and one case in education. All ethical models presented until now are of the type of explicit ethical agents highly different from fully ethical agents that can be used in moral dilemmas in reality. Therefore, it can be stated that attempts to assign human duties to systems are still at the beginning of the path and further studies are required. Table 3 summarizes these studies which were explained in detail in the previous section, to ease. The studies have been sorted based on the presentation date.

Table 3

Key Milestones in Conceptual Frameworks and Implementation Strategies for Artificial Moral Agents (AMA)

Model/system provided	Year	Ethical Theory	Techniques and tools used	Design Approach	Domain	Details Implementation	Evaluation method
MedEthEx (M. Anderson et al., 2006b)	2006	Deontological	Logical reasoning Learning Knowledge representation and ontologies	Hybrid	Medical	Brief description No code available	It is claimed that the system has been tested and the results compared with human results.
MoralDM (Dehghani et al., 2008)	2008	Consequentialist and Deontological	Logical reasoning Case-based reasoning natural language processing	Top-Down	---	Brief description No code available	It is claimed that the system has been tested and the results are compared to the ground truth.
(Honarvar & Ghasem-Aghaee, 2009)	2009	Consequentialist	case-based reasoning BDI architecture Learning	Bottom-up	---	Brief description Partial code is available	It is claimed that the system has been tested and the results are compared to the ground truth.
(Wallach et al., 2010)	2010	unspecified	LIDA architecture Logical reasoning Probabilistic reasoning Learning	Hybrid	---	Brief description No code available	This has been empirically evaluated using a simple case study with a real robot.
(Azad-Manjiri, 2014)	2014	Consequentialist and Deontological	Decision tree Learning Optimization	Hybrid	Medical	Brief description No code available	None
(Cervantes et al., 2016)	2016	unspecified	Logical reasoning Optimization	Top-Down	---	Brief description Partial code is available	The system has been evaluated by defining a scenario by the authors
GenEth (Anderson & Anderson, 2018)	2018	Deontological	Learning Inductive logic	Hybrid	---	Full description Code is available	This has been empirically evaluated using a simple case study with a real robot.
(Vanderelst & Winfield, 2018)	2018	Consequentialist	Logical reasoning Optimization	Top-Down	---	No description No code available	It is claimed that the system has been tested and the results are compared to the ground truth.
(Noothigattu et al., 2018)	2018	Deontological	Inductive logic Learning Optimization	Bottom-up	Vehicle	Brief description Code is available	It is claimed that the system has been tested and the results compared with human results.
(Bremner et al., 2019)	2019	Consequentialist	BDI architecture	Top-Down	---	Brief description Partial code is available	This has been empirically evaluated using a simple case study with a real robot.
(Córdova & Vicari, 2021; Córdova et al., 2021)	2021	Consequentialist and Deontological	Logical reasoning Knowledge representation and ontologies BDI architecture	Top-Down	educational	Brief description No code available	The MODEL has been evaluated by defining two scenarios by the authors
(Stenseke, 2023, 2024)	2022	Virtuous theory	Learning artificial neural networks classification method	Hybrid	---	Full description Code is available	The system has been evaluated by defining a scenario by the author

According to the review and analysis of selected studies in this study, the following points and the gaps in this field can be useful for research continuity in this context:

- Despite the complexities of the moral decision-making process in humans and since a single approach is not sufficient for this topic since it does not cover all machine ethics requirements, integration of ethical theories and various ethics approaches for implementing AMAs can be promising for increasing the quality of performance of AMAs.

- Moral decision-making is a high-level cognitive process. Therefore, it is beneficial to have a suitable platform that provides a performance that is as much as possible similar to the cognitive performance of humans. Hence, it can be stated that cognitive architectures inspired and modeled by cognitive sciences trying to simulate many cognitive aspects of humans to create artificially intelligent agents can be an effective shortcut in implementing moral decision-making ideas.
- As moral dilemmas occur in a social environment where the decision of each agent affects the other agents and also the environment, the multi-agent nature of moral decision-making problems must be considered. In this case, a moral agent evaluates the consequence of its actions and decisions by receiving feedback from the environment and other agents. The moral agent tries to perform better and improve its function during the learning process.
- Interdisciplinary cooperation between computer engineers and ethics scientists is highly needed to create ethical scenarios and a set of implementable data comprising a set of questions and duties that experts agree on.
- Solutions for evaluation of AMAs' performance not only provide the field for comparing ethical intelligent agents but also ensure whether the system acts regarding the received dataset as expected.
- Regarding a wide range of moral dilemmas, in reality, it may be challenging and hard to provide a general moral model and achieve a universal consensus on ethical principles and performance. Hence, researchers suggest moving towards particularism in different fields because various problems have various emphases on ethical principles.
- Considering the worries about the use of AMAs, it is a suitable idea to use ethical decision support systems and or advisor and persuasive systems (such as the system presented in (Augello et al., 2023) instead of moral decision-makers in sensitive and vital fields related to death and life of humans. Therefore, such research fields must be examined to reduce worries.

9. Conclusion

As explained before, increased computer and AI systems seriously need some mechanisms to mitigate worries and enhance trust in intelligent systems. Hence, the simulation of important decision-making indicators of humans in intelligent systems is an effective factor for achieving these goals. Ethics and morality are one of these indicators. The equipment of agents with moral decision-making is an important step to increasing trust in artificially intelligent systems. Many academic centers worldwide have examined ethics in AI. This study summarizes, classifies, and expresses the relevant topics by reviewing the relevant papers that have been published over two decades. In the next step, this study examines the agents equipped with ethics in decision-making over the recent two decades based on an engineering and technical view and then analyzes the research gap. Regarding the available challenges and gaps, this study suggests paying attention to the following points: studying the gaps and topics that have received less attention, paying attention to a hybrid view of ethical theories and implementation approaches, considering the cognitive processes of humans, and using advantages and capabilities of cognitive architecture to implement AMAs, considering cooperation with ethics scientists to create ethical scenarios and datasets, creating some mechanisms to evaluate and compare ethical intelligent agents, moving towards particularism in a certain domain of moral dilemmas and developing moral decision support systems and advisor or persuasive systems. Ultimately, it is worth noting that some studies may have been published after writing this paper that are not reviewed in this study due to the rapid speed of development and publication of scientific documents these days.

Authors' Contributions

F.G. conceptualized the study, reviewed the literature, and provided the theoretical framework for the development of artificial moral agents (AMAs). T.B. contributed to the technical analysis, identifying key engineering processes and techniques for implementing morality in AI systems. M.P. addressed the challenges and proposed recommendations for future research, emphasizing the practical implications and gaps in current methodologies. All authors collaborated on the drafting, revising, and finalizing of the manuscript. They collectively approved the final version and are accountable for the integrity of the research.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

We would like to express our gratitude to all individuals helped us to do the project.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

Not applicable.

References

Alebouyeh, A., & Noormohamad, N. (2020). The Ethical Challenges of Liberal Eugenics and the Principle of Liberty. *Journal of Moral Studies*, *3*(6), 5-26. https://ethics.riqh.ac.ir/article_13055.html?lang=en

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*, 149-155. https://doi.org/10.1007/s10676-006-0004-4

Anderson, M., & Anderson, S. L. (2008). Ethical healthcare agents. In *Advanced computational intelligence paradigms in healthcare-3* (pp. 233-257). Springer. https://doi.org/10.1007/978-3-540-77662-8_10

Anderson, M., & Anderson, S. L. (2018). GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, *9*(1), 337-357. https://doi.org/10.1515/pjbr-2018-0024

Anderson, M., Anderson, S. L., & Armen, C. (2006a). An approach to computing ethics. *IEEE Intelligent Systems*, *21*(4), 56-63. https://doi.org/10.1109/MIS.2006.64

Anderson, M., Anderson, S. L., & Armen, C. (2006b). MedEthEx: a prototype medical ethics advisor. Proceedings of the national conference on artificial intelligence, Menlo Park, CA; Cambridge, MA; London.

Anderson, M., Anderson, S. L., & Armen, C. (2006). MedEthEx: a prototype medical ethics advisor. Proceedings of the national conference on artificial intelligence,

Augello, A., Città, G., Gentile, M., & Lieto, A. (2023). A storytelling robot managing persuasive and ethical stances via act-r: an exploratory study. *International Journal of Social Robotics*, *15*(12), 2115-2131.

Azad-Manjiri, M. (2014). A new architecture for making moral agents based on C4.5 decision tree algorithm. *International Journal of Information Technology and Computer Science (IJITCS)*, *6*(5), 50-57. https://doi.org/10.5815/ijitcs.2014.05.07

Bickley, S. J., & Torgler, B. (2023). Cognitive architectures for artificial intelligence ethics. *Ai & Society*, *38*(2), 501-519. https://doi.org/10.1007/s00146-022-01452-9

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, *20*, 41-58. https://doi.org/10.1007/s10676-018-9444-x

Brachman, R. J. (2002). *Systems that know what they're*. https://web.stanford.edu/class/cs227/Readings/Systems%20That%20Know%20What%20They%20Are%20Doing.pdf

Bremner, P. A., Dennis, L. A., Fisher, M., & Winfield, A. F. (2019). On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots. *Proceedings of the IEEE*, *107*, 541-561. https://doi.org/10.1109/JPROC.2019.2898267

Cervantes, J.-A., Rodríguez, L.-F., López, S., & Ramos, F. (2013). A biologically inspired computational model of moral decision making for autonomous agents. 2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing,

Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, *26*(2), 501-532. https://doi.org/10.1007/s11948-019-00151-x

Cervantes, J. A., Rodríguez, L. F., López, S., & Ramos, F. (2013). A biologically inspired computational model of moral decision making for autonomous agents. 2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing,

Cervantes, J. A., Rodríguez, L. F., López, S., Ramos, F., & Robles, F. (2016). Autonomous agents and ethical decision-making. *Cognitive Computation*, *8*, 278-296. https://doi.org/10.1007/s12559-015-9362-8

Cervantes, S., López, S., & Cervantes, J. A. (2020). Toward ethical cognitive architectures for the development of artificial moral agents. *Cognitive Systems Research*, *64*, 117-125. https://doi.org/10.1016/j.cogsys.2020.08.010

Córdova, P. R., & Vicari, R. M. (2021). A conceptual model for artificial moral agents (AMA) in the educational context. *International Journal of Development Research*, *11*(06), 47868-47871. https://www.journalijdr.com/conceptual-model-artificial-moral-agents-ama-educational-context

Córdova, P. R., Vicari, R. M., Brusius, C., & Coelho, H. (2021). A proposal for artificial moral pedagogical agents. Trends and Applications in Information Systems and Technologies: Volume 1,

Dehghani, M., Tomai, E., Forbus, K., Iliev, R., & Klenk, M. (2008). Moraldm: A computational modal of moral decision-making. Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci),

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer. https://doi.org/10.1007/978-3-030-30371-6

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In *Machine learning and the city: Applications in architecture and urban design* (pp. 535-545). https://doi.org/10.1002/9781119815075.ch45

Formosa, P., & Ryan, M. (2021). Making moral machines: why we need artificial moral agents. *Ai & Society*, *36*(3), 839-851. https://doi.org/10.1007/s00146-020-01089-6

Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., Snaider, J., Agrawal, P., & Chen, S. (2016). A LIDA cognitive model tutorial. *Biologically Inspired Cognitive Architectures*, *16*, 105-130. https://doi.org/10.1016/j.bica.2016.04.003

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *Ai & Society*, *35*, 795-809. https://doi.org/10.1007/s00146-020-00977-1

Graff, J. (2024). Moral sensitivity and the limits of artificial moral agents. *Ethics and Information Technology*, *26*(1), 13.

Greene, J., Rossi, F., Tasioulas, J., Venable, K., & Williams, B. (2016). Embedding ethical principles in collective decision support systems. Proceedings of the Aaai Conference on Artificial Intelligence,

Gudmunsen, Z. (2024). The moral decision machine: a challenge for artificial moral agency based on moral deference. *Ai and Ethics*, 1-13. https://doi.org/10.1007/s43681-024-00444-3

Hagendorff, T. (2023). AI ethics and its pitfalls: not living up to its own standards? *Ai and Ethics*, *3*(1), 329-336. https://doi.org/10.1007/s43681-022-00173-5

Hagendorff, T., & Danks, D. (2023). Ethical and methodological challenges in building morally informed AI systems. *Ai and Ethics*, *3*(2), 553-566. https://doi.org/10.1007/s43681-022-00188-y

Honarvar, A. R., & Ghasem-Aghaee, N. (2009). Casuist BDI-agent: a new extended BDI architecture with the capability of ethical reasoning. Artificial Intelligence and Computational Intelligence: International Conference, AICI 2009, Shanghai, China.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399.

Madl, T., & Franklin, S. (2015). Constrained incrementalist moral decision making for a biologically inspired cognitive architecture. In *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations* (pp. 137-153). https://doi.org/10.1007/978-3-319-21548-8_8

Malle, B. F. (2016). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, *18*, 243-256.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage. https://books.google.de/books/about/Rebooting_AI.html?id=O8muDwAAQBAJ&redir_esc=y

Miller, G. J. (2021). Artificial intelligence project success factors-beyond the ethical principles. Special Sessions in the Advances in Information Systems and Technologies Track of the Conference on Computer Science and Intelligence Systems,

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, *21*(4), 18-21. https://doi.org/10.1109/MIS.2006.80

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how-an overview of ai ethics tools, methods and research to translate principles into practices. *arXive*. https://doi.org/10.2139/ssrn.3830348

Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2019). Adjustable autonomy: a systematic literature review. *Artificial Intelligence Review*, *51*, 149-186.

Navon, M. (2021). The Virtuous Servant Owner-A Paradigm Whose Time has Come (Again). *Frontiers in Robotics and Ai*, *8*, 715849. https://doi.org/10.3389/frobt.2021.715849

Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. (2018). A voting-based system for ethical decision making. Proceedings of the Aaai Conference on Artificial Intelligence,

Rovatsos, M. (2019). From AI ethics to ethical AI. *Macau, China, IJCAI*.

Rovatsos, M. (2019). From AI ethics to ethical AI. IJCAI, Macau, China.

Scheutz, M. (2017). The case for explicit ethical agents. *Ai Magazine*, *38*(4), 57-64. https://doi.org/10.1609/aimag.v38i4.2746

Schmiljun, A. (2019). Moral Competence and Moral Orientation in Robots. *Ethics in Progress*, *10*(2), 98-111. https://doi.org/10.14746/eip.2019.2.9

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *Journal of Database Management (JDM)*, *31*(2), 74-87.

Stenseke, J. (2023). Artificial virtuous agents: from theory to machine implementation. *Ai & Society*, *38*(4), 1301-1320. https://doi.org/10.1007/s00146-021-01325-7

Stenseke, J. (2024). Artificial virtuous agents in a multi-agent tragedy of the commons. *Ai & Society*, *39*(3), 855-872. https://doi.org/10.1007/s00146-022-01569-x

Stroll, A., & Popkin, R. H. (1956). *Philosophy Made Simple*. Routledge. https://books.google.de/books/about/Philosophy_Made_Simple.html?id=bVPejt00AnMC&redir_esc=y

Sullins, J. P. (2021). Artificial Phronesis: What It Is and What It Is Not. In T. A. Stapleford (Ed.), *Science, Technology, and Virtues: Contemporary Perspectives* (pp. 0). Oxford University Press. https://doi.org/10.1093/oso/9780190081713.003.0008

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, *53*(6), 1-38. https://doi.org/10.1145/3419633

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385-409. https://doi.org/10.1007/s11023-020-09537-4

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, *48*, 56-66. https://doi.org/10.1016/j.cogsys.2017.04.002

Vijayaraghavan, A., & Badea, C. (2024). Minimum levels of interpretability for artificial moral agents. *Ai and Ethics*, 1-17. https://doi.org/10.1007/s43681-024-00536-0

Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, *2*(3), 454-485. https://doi.org/10.1111/j.1756-8765.2010.01095.x

Winfield, A. F., Blum, C., & Liu, W. (2014). Towards an ethical robot: internal models, consequences and ethical action selection. Advances in Autonomous Robotics Systems: 15th Annual Conference, TAROS 2014, Birmingham, UK.

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 20150375. https://doi.org/10.1098/rstb.2015.0375

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020). A survey on ethical principles of AI and implementations. 2020 IEEE Symposium Series on Computational Intelligence (SSCI),