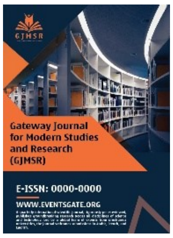
https://doi.org/10.61856/6kyazw30



Scientific Events Gate

GJMSR
Gateway Journal for Modern Studies and Research
https://gjmsr.eventsgate.org/gjmsr/

GJMSR
Gateway Journal for Modern Studies and Research

The Impact of Artificial Intelligence on Ethical Decision-Making: Challenges and Opportunities in Intelligent Systems

Rafid Mohammed Khaleefah
University of Basrah – Iraq
rafeed.rmk@uobasrah.edu.iq

Received 17/07/2025 - Accepted 8/8/2025 Available online 27/8/2025

Abstract: This study investigates the ethical implications of Artificial Intelligence (AI) in intelligent decision-making systems, focusing on sectors such as healthcare, autonomous vehicles, recruitment, and law enforcement. It proposes a multidimensional ethical risk assessment framework that integrates technological, human, and interactional factors. The research adopts a qualitative, interdisciplinary methodology drawing from philosophy, computer science, legal theory, and global policy frameworks. Through an extensive literature review and practical examples, the study explores key ethical challenges such as algorithmic bias, lack of transparency, over-reliance on AI systems, and cultural variances in ethical governance. The article addresses the absence of a unified framework for identifying and managing ethical risks across AI applications. Case studies, including self-driving cars, medical diagnosis systems, and predictive policing, demonstrate how the proposed framework can be applied in real-world contexts. The hypotheses of the study are tested through theoretical synthesis and comparative ethical reasoning rather than empirical sampling. The findings highlight the necessity of culturally adaptive and socially conscious AI governance. The study comes with advocacies for embedding ethics throughout the AI lifecycle, from design to deployment. It contributes to global discourse by providing a robust model for anticipating and mitigating ethical risks in intelligent systems, ultimately promoting fairness, accountability, and human-centred design in AI applications.

Keywords: Artificial Intelligence Ethics, Ethical Decision-Making, Intelligent Systems, Risk Assessment Framework, Human-Centred AI.

1. Introduction

Artificial Intelligence (AI) is rapidly transforming decision-making processes across various domains, including healthcare, law, finance, and security. However, as intelligent systems grow in autonomy and complexity, they introduce profound ethical challenges, particularly in contexts where decisions directly affect human lives, rights, and dignity. The integration of AI into such sensitive areas raises concerns related to bias, opacity, moral agency, and the delegation of ethical responsibility to non-human entities. To ensure that AI systems maintain alignment with societal values and ethical standards, these concerns necessitate rigorous analysis and proactive mechanisms. On this basis, the risk prevention mechanism of artificial intelligence decision-making is discussed (Guan et al., 2022).

It is widely accepted that three major sources of uncertainty shape the mechanism of artificial intelligence decision-making as data uncertainty, algorithmic opacity, and input condition variability (Rahwan et al., 2019; Floridi et al., 2018). From a technological standpoint, these uncertainties can trigger ethical risks such as privacy violations, safety threats, and opaque accountability (Calo, 2015; Mittelstadt et al., 2016). Furthermore, AI systems lack emotional intelligence and empathy, raising concerns in domains such as healthcare and law (Boddington, 2017; Moor, 2006).

Ethical risks also emerge from interactions between intelligent systems, human agents, and unpredictable environmental factors (Jobin et al., 2019; Guan et al., 2022). Addressing these risks involves ethical risk analysis, representation strategies, and mitigation mechanisms (Allen et al., 2000; van Wynsberghe, 2013). Ethical AI decision-making requires a multidimensional approach involving the technical, human, and systemic layers.

While much of the literature focuses on the functionality of decision-making algorithms, the ethical dimension has not received equivalent attention (Calo, 2015; Jobin et al., 2019). This study aims to bridge this gap by proposing a comprehensive ethical risk assessment framework covering four domains: technology, human decision-making, AI–human interaction, and ethical maturity of intelligent systems.

1.1 Research Problem and Objectives

1.1.1 Research Problem

Despite the increasing integration of Artificial Intelligence in decision-making systems, there remains a critical gap in understanding how to identify and mitigate ethical risks across diverse applications systematically. Many AI systems are deployed without sufficient ethical oversight, leading to biased decisions, opacity, and loss of public trust. This research addresses the problem of inadequate ethical frameworks that can guide the design, development, and deployment of intelligent systems in a socially responsible manner.

1.1.2 Objectives of the Study

1. To analyze the ethical implications of AI-based decision-making across domains such as healthcare, law, and autonomous systems.
2. To develop a multidimensional ethical risk assessment framework for intelligent decision-making systems.
3. To propose actionable recommendations for embedding ethics into the lifecycle of AI systems.
4. To compare international approaches to ethical AI governance.

1.2 Research Hypotheses

- **H1**: Ethical risks in AI decision-making systems can be effectively addressed through a structured and multi-disciplinary framework.
- **H2**: Cultural and legal differences significantly influence the ethical governance of AI systems.

1.3 Study Community and Sample

This study employs a theoretical and qualitative analysis rather than empirical fieldwork; therefore, it does not involve a specific community or statistical sample. Instead, it draws from a diverse corpus of academic, legal, and policy-based literature to synthesize a global ethical framework applicable across sectors.

The study does not focus on a specific geographic area or target population, it instead relies on a comparative analysis of existing academic and regulatory literature on AI ethics. This approach enables the development of a generalisable ethical framework applicable across different sectors and contexts.

1.4 Research Significance

The ethical implications of artificial intelligence (AI) have emerged as a central global concern as intelligent systems are increasingly deployed in sensitive domains such as healthcare, law, governance, and autonomous vehicles. This study focuses on ethical decision-making in AI to contribute to a pressing global debate-one that involves academic institutions, policymakers, and technology leaders.

Internationally, several research bodies and authors have emphasized the urgency of addressing ethical risks in AI. For example, scholars such as Luciano Floridi (2023), Virginia Dignum (2019), and Wendell Wallach (2008) have written extensively on algorithmic transparency and fairness. Global initiatives like the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2018) and the European Commission's High-Level Expert Group on AI (2019) have also published comprehensive ethical guidelines to promote responsible AI practices.

This research does not restrict to the identification of ethical problems but it also introduces a multidimensional ethical risk assessment framework that considers technology, human factors, and systemic interactions. Unlike other studies that focus on a single application or regional context, this paper presents a comparative, global view. It aims to contribute to building adaptable, fair, and socially responsible AI systems that align with international ethical principles while remaining culturally sensitive and practically implementable.

2. Contemporary Perspectives on Artificial Intelligence

This section explores how AI has evolved and how this evolution affects ethical considerations in intelligent systems. AI originated from symbolic logic and rule-based reasoning but has rapidly transformed through data-driven approaches, including machine learning, deep learning, and neural networks (Dent, 2020; Russell & Norvig, 2021). These models learn from vast datasets and produce probabilistic outputs, often characterized by opacity and uncertainty due to biased or incomplete training data (Guan et al., 2022).

This intrinsic uncertainty raises serious concerns when AI is applied in high-stakes domains such as healthcare, law enforcement, and autonomous vehicles. While earlier symbolic AI systems offered greater explainability, modern "black-box" algorithms frequently lack transparency, making it difficult to understand how decisions are made or to ensure accountability (Daly et al., 2019).

Furthermore, it is important to distinguish between Narrow AI, which is designed for specific tasks such as image recognition or recommendation filtering, and General AI, which aspires to simulate human-level reasoning across diverse contexts. However, from an ethical standpoint, what matters most is not the level of autonomy or generality of the AI system, but how its decisions affect human dignity, fairness, and social values (Farisco et al., 2020).

Accordingly, this study focuses on how such evolving technologies influence ethical risk, rather than simply cataloguing definitions. The goal is to investigate how these systems operate in complex social contexts and to develop an ethical risk framework that addresses technological uncertainty, human values, and system-level interactions.

2.1. History of AI Development

The history of AI development can be traced back to the emergence of the term "artificial intelligence" in 1956 at the Dartmouth Workshop. McCarthy defined AI as "the science and engineering of making intelligent machines." This new area flourished in the following years with rapid advances despite the limited computational power and data availability. AI was initially limited to rule-based symbolic methods, which caused the blind spots of first-generation expert systems to be off by 1987 due to the difficulty in constructing rules and the failure to integrate low-level perception and high-level reasoning. Machine learning emerged with the advent of the perceptron in 1957. Following successful applications by others, the competitive connectionism paradigm rekindled the interest in computational approaches to learning in the mid-1980s. Still, it quickly turned into another winter by the late 1990s.

The Internet Revival in the mid-1990s changed the landscape of AI research and applications by unlocking unprecedented amounts of data. In particular, the advent of big data and advancements in hardware such as GPUs rekindled the interest in computationally intensive but highly effective deep learning networks based on the connectionist paradigm. Under the deep learning phenomenon, AI research entered the third wave, which is widely believed to be highly promising and capable of autonomous levels of intelligence. Since GO's victory in 2016, AI has become widely accepted and no longer remains a niche concept. AI is being increasingly deployed for many applications and becoming widely pervasive at an astonishing level, greatly impacting how people work and live. AI is being adopted in more contexts that affect people's lives, it negatives consequences resulting from its failures and design flaws are becoming more visible, leading to questions about the social impact of AI.

General-purpose of AI is quickly maturing and capable of performing many higher-level thinking tasks; AI is deployed for many applications which is claimed to be an intelligent agent or system that can work on behalf of humans. They are widely believed to be capable of reasoning and making rational decisions or ethical judgements. Regardless of whether AI is intelligent or not, giving it decision-making power without responsibility is more troublesome since it may result in serious consequences, as seen in drone attacks and self-driving cars. Unintended errors or unforeseen emergencies causing severe losses to humans or society are foreseeable. Technological advancement in managing uncertainty in various aspects of AI is needed to uphold human values and ethical principles in future intelligent services. Many questions arise regarding the profound societal and ethical issues raised by the capability, responsibility, and trust in dealing with ethical or value-based disputed issues in intelligent systems.

2.2. Current Trends in AI Technology

In terms of the conception and construction of intelligent systems, current AI paradigms can be classified roughly into systems which are centred on deep learning, traditional static models, and the systems do not necessarily require frequent looks at algorithm design, such as evolutionary algorithms and automatic generation models. In general, AEOMD can be introduced into the intelligent decision systems of the first two categories. It can also be used indirectly in third-type systems. For example, in genetic algorithms, heuristic knowledge is often embedded inside fitness functions or selection operations; some grammar rules can be built inside emergence automata or L-systems can be structured in cellular automata evolution methods; improperly designed or selected genetic operators can also lead to ethical risks of the GA. After being constructed, an AEOMD will take effect via a response mechanism (Guan et al., 2022).

During system operation, the intelligent systems execute operations according to the selected model and decision algorithms. Suppose the AEOMD is implemented in such operations. In that case, the system will directly control the execution of decisions or alter the model's or algorithm's execution manner.

In case no model refinement or decision methods are required, intelligent systems output strategies based on past or present data. Though current approaches can modify decisions based on new data on safety, robustness, and quality aspects, little work has been conducted on the ethical checking of intelligent systems. Thus, a preliminary outline of an ethical decision feedback mechanism is proposed, which adjusts the behaviour and potential output space of a first-category intelligent decision system based on an ethical regular set or ethical cost function to lessen the impact of AEOMD. The details of feedback-to-rectify methods are highly related to different systematic models. Overall, a current research framework aimed at ethical decision-making in intelligent systems is constructed by investigating the recent development of AI ethics research fields (Dent, 2020).

Current technologies, scenarios, and ethical issues in AI are surveyed to enhance understanding of its ethicality.

3. Literature Review

Numerous studies have examined the ethical implications of artificial intelligence (AI) in decision-making systems across domains such as healthcare, justice, and autonomous technologies. For instance, Allen et al. (2000) introduced the concept of artificial moral agents, discussing whether machines can be moral actors. Buolamwini and Gebru (2018) explored algorithmic bias in facial recognition systems, revealing how underrepresentation leads to discriminatory outcomes. Similarly, Barocas et al. (2019) addressed fairness challenges in machine learning, emphasizing structural bias embedded in training data.

Farisco et al. (2020) proposed ethical evaluation criteria for AI, while Jobin et al. (2019) mapped global AI ethics guidelines, highlighting differences in cultural and policy approaches. Additionally, Guan et al. (2022) developed a framework for identifying ethical risks in AI decision-making, which strongly influenced this paper's risk assessment model.

Despite these valuable contributions, many studies either focus on single domains or offer fragmented views of ethics in AI. This study seeks to integrate those insights into a unified, multi-disciplinary framework for ethical risk assessment across various sectors. By synthesizing findings from multiple perspectives, this research addresses the lack of a holistic model applicable to both theory and governance.

4. Ethics in AI: Philosophical Foundations

4.1 Moral Philosophy and AI

Moral philosophy provides the foundation for evaluating right and wrong in human conduct. When applied to artificial intelligence, it challenges users to consider whether machines can and should make decisions that have ethical consequences. As intelligent systems increasingly take on roles in healthcare, autonomous vehicles, and law enforcement, the necessity to embed ethical considerations becomes unavoidable. AI technologies are not inherently moral or immoral; rather, they reflect the intentions and values encoded into them by their developers. Therefore, understanding moral reasoning becomes crucial for guiding AI design and governance.

4.2 Fundamental Ethical Theories

Several classical ethical theories provide frameworks for assessing AI behaviour. Utilitarianism, for instance, evaluates actions based on their consequences and aims for the greatest good for the greatest number. Deontological ethics, proposed by Immanuel Kant, emphasizes adherence to moral duties and rules regardless of outcomes. Virtue ethics, rooted in Aristotelian thought, focuses on the moral character of the decision-maker rather than specific rules or consequences.

Each of these theories offers unique implications for AI. A utilitarian AI might prioritize efficiency and overall benefit, while a deontological system would follow strict ethical constraints. Integrating such theories into AI development allows for more robust, transparent, and justifiable decision-making models.

4.3 Artificial Moral Agents

Artificial Moral Agents (AMAs) are AI systems designed to make or support moral decisions. The debate around AMAs focuses on whether machines can truly engage in moral reasoning or merely simulate it. Scholars like Allen et al. (2000) introduced the concept, arguing that AMAs can operate at various levels from ethical impact awareness to full moral autonomy.

Designing AMAs requires more than technical sophistication; it demands philosophical clarity. Ethical frameworks must be embedded into AI architecture and continuously audited to prevent harm, bias, or injustice. The idea of delegating ethical agency to machines raises significant questions about responsibility, accountability, and the limits of machine cognition in moral contexts.

5. Methodology

This study adopts a qualitative, comparative, and interdisciplinary approach grounded in literature analysis. Rather than relying on empirical fieldwork or data collection from a particular institution or region, the research is based on synthesizing a wide array of academic, technical, philosophical, and policy-oriented sources. The core objective is to extract recurring ethical principles, risk patterns, and governance challenges in AI-based decision-making systems.

Sources analyses include peer-reviewed journal articles, ethical frameworks from international organisations (such as UNESCO and the OECD), and regulatory proposals from major jurisdictions (e.g., the EU AI Act, U.S. sector-specific standards, and China's national guidelines). These sources were selected purposively to represent a range of global perspectives and to ensure the framework proposed in this study is culturally adaptable and ethically robust.

The methodology follows a thematic analysis structure, identifying core ethical themes such as fairness, transparency, bias, accountability, and privacy. It also involves a comparative review of how different regions interpret and apply these themes in policy and system design. This interpretive strategy supports the construction of a multidimensional ethical risk assessment model applicable to various sectors.

6. Ethics in Decision-Making

Ethical decision-making is a critical aspect of intelligent systems, especially when artificial intelligence is entrusted with tasks that significantly affect human lives, such as in healthcare, criminal justice, and autonomous vehicles. The core challenge lies in aligning algorithmic processes with moral reasoning typically applied by humans.

Ethical decision-making in AI must consider principles such as autonomy, justice, non-maleficence, and beneficence. For example, an AI system used in medical triage must weigh patients' needs objectively while avoiding discrimination based on age, gender, or ethnicity. These moral judgments, though intuitive for human practitioners, are difficult to encode algorithmically.

Moreover, ethical dilemmas often arise when values conflict, for instance, prioritising public safety versus individual privacy in surveillance systems. Intelligent systems need to be programmed with transparent and adaptable ethical frameworks to handle such conflicts.

AI systems do not "choose" ethically; rather, they follow predefined logic. Therefore, system designers must embed ethical parameters during development, ensure continuous monitoring, and establish accountability mechanisms. Failure to do so may lead to unethical or biased outcomes that damage public trust.

This section serves as a foundation for the ethical risk assessment framework proposed in the subsequent chapters, bridging theoretical ethics with practical AI governance.

6.1. Fundamental Ethical Theories

Designing ethical agents requires a clear ethical framework. Three core ethical theories are often used: utilitarianism, deontology, and virtue ethics. Utilitarianism promotes actions that maximise overall happiness and can be implemented in basic decision systems. Deontology focuses on adherence to moral rules, offering clarity for rule-based AI systems. Virtue ethics emphasises the character of the agent, complicating its implementation for AI.

While philosophical depth is challenging to encode for machines, simplified implementations can guide AI behaviour in critical scenarios. For example, a utilitarian autonomous vehicle might be programmed to minimise total harm during an unavoidable accident. However, embedding such reasoning into AI remains a technical and ethical challenge requiring further interdisciplinary collaboration.

6.2. Artificial Moral Agents

One of the most significant debates in AI ethics concerns the possibility of creating artificial moral agents – systems capable of making ethical decisions autonomously. While current AI systems are programmed to follow predefined ethical guidelines, the question remains: can they ever utterly understand or internalise moral reasoning? Unlike humans, AI lacks consciousness, emotions, and a deeper sense of moral responsibility, which raises questions about whether AI can ever be trusted to make morally sound decisions in complex scenarios.

6.3. Moral Philosophy and AI

The tendency to perceive machines as moral agents, or at least as having some degree of agency, even when they only react to external stimuli, is not new. It resembles an artificial creature that evolves along a path of persecution or a complex automaton designed to appear virtuous while hiding its inability to perform good actions. Something similar has recently been observed in the case of artificial intelligence (AI) programs that people may choose to interact with communicatively. With regard to moral responsibility more generally, one aim of the ongoing research program on agentive AI is to explore the potential and pitfalls of endowing AI agentive decision-making processes with moral agency or at least some degree of agency.

Simulation is hugely powerful in the natural sciences, but is virtually unchallenged in the field of classical AI. Straightforwardly, scholastic questions of intention and beliefs become wildly multitudinous in the context of agency. Philosophically and practically, understanding the epistemic privileges and ethical standing of AI is vitally important. However, AI systems can have some agency today; one is invited to wrestle with these questions, regardless of how daunting they may seem. Human-controlled vehicles embedded with limited AI guidance have crashed while negotiating curves too sharp for safe entry speed. Human agents controlling autonomous vehicles (AVs) can opt to countermand decisions by the AI systems. Still, unsettled ethical questions surround cases in which a vehicle is left to control itself. Then, the relevant assumptions concerning ethics in the context of human decision-making or AV behaviour can be adduced and examined. All of these efforts aim to contribute to a better understanding of both the possibilities and the pitfalls of autonomous machines making moral decisions.

Given the increasing deployment of autonomous machines and systems in fields such as education, criminal justice, and healthcare, ethical concerns surrounding machine decision-making are becoming increasingly prevalent (Seeamber & Badea, 2023). Nonetheless, these concerns are complicated by the vagueness of ethical concepts and the previously unexplored differences between machines' and humans' understandings of ethics. The moral intuitions of human beings have been widely studied within the field of machine ethics. Still, research on machine moral reasoning is limited. An experiment involving a novel legal dilemma task was designed to explore the moral decisions of LLMs, which demonstrated their ability to reason about various moral dilemmas successfully (Zhang et al., 2023). It is critical to investigate machines that are much more controllable and less capable of emitting unsolicited output than ChatGPT, but with a more limited capacity to justify their decisions.

7. AI and Ethical Decision-Making

AI-based decision-making systems are increasingly relied upon in environments characterised by rapid changes and multiple constraints. These systems identify optimal alternatives and evaluate performance based on defined criteria. However, ethical considerations have emerged as central challenges, particularly in systems that simulate moral reasoning or make decisions that directly impact human rights and safety (Seeamber & Badea, 2023). The following diagram (Figure 1) illustrates how fairness and accountability contribute to achieving transparency in intelligent systems, further clarifying the relationship between core ethical principles in AI decision-making.

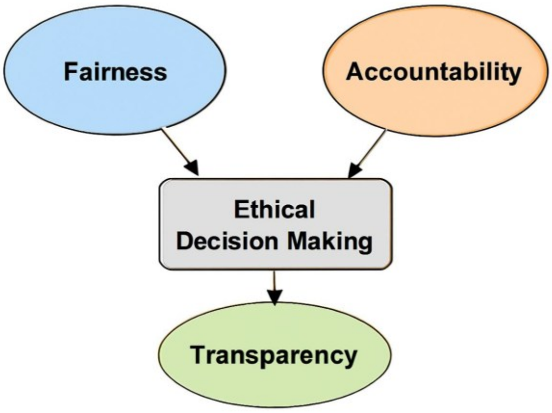


Figure 1. Core components of ethical decision-making in AI: Fairness and Accountability as foundational inputs, and Transparency as a resulting principle.

This diagram (Figure 1) illustrates the fundamental elements that underpin ethical decision-making in intelligent systems. Fairness and accountability are essential for developing ethical algorithms, ensuring that systems remain unbiased and take responsibility for their outcomes. Transparency is also a vital aspect of this process, as it facilitates understanding, review, and trust in these systems. Collectively, these three pillars form the foundation for assessing and designing ethical AI behaviours, particularly in critical situations when it is imperative to uphold human values. In light of unprecedented ethical challenges posed by AI, many scholars have begun to pay attention to the ethical issues and problems involved in AI technology. Human-machine collaborative decision-making (HMC) is a complex decision-proposition distribution problem that requires expert knowledge and team consensus. Most existing HMC processes address this problem using AI techniques and assume that decision-makers respect decision-making procedures and protocols. Studies have shown that behavioural adaptation can improve team performance during cooperative strategic tasks with minimal interaction. Under the assumption of compliance, users can reasonably derive corresponding ethical and moral principles to restrict decision options. Based on the ethical principle of respect, AI-supported systems can analyze, classify, and estimate the validity of decision options and interactions during the IDM process. In response to these ethical risks of AI, many scholars have proposed AI's ethical risk governance systems, emphasizing risk identification and assessment techniques for the intelligent decision-proposition distribution process (Guan et al., 2022).

Scholars argue over the best approach to manage ethical risks in AI: while some advocate for top-down regulation based on universal principles (Jobin et al., 2019), others highlight the limitations of imposing rigid ethical frameworks across culturally diverse contexts (Crawford & Calo, 2016; Fjeld et al., 2020). These scholars argue that ethics cannot be one-size-fits-all and must be sensitive to local norms, power structures, and political systems. Ethical governance must account for competing interests, emotional complexity, and institutional power dynamics. Simply encoding moral rules into AI without understanding their social context may lead to flawed or even harmful outcomes (Kluge Corrêa et al., 2023). In light of the fact that intelligent decision-making (IDM) processes based on AI are complicated and intelligent, it is very difficult to develop an intelligent decision-making (IDM) system that is friendly and socially acceptable for humans based on a top-down approach. In other words, machines can imitate the brain's physiological call methods, but cannot think like humans.

7.1. The Role of AI in Ethical Choices

Intelligent agents powered by artificial intelligence (AI) can make decisions autonomously, using sensors and algorithms to process data from their environment. However, this has raised many concerns about the risks associated with these decisions and the actions taken by intelligent agents. Social issues have arisen regarding how to make ethical choices, similar to humans with ethical beliefs. It remains unclear how ethics can be included in the decision-making process of intelligent systems. The entry of AI into the decision-making environment imposes considerable challenges. On the one hand, AI can identify problems, gather relevant information independently of humans, and make decisions. On the other hand, many problems arise from this technology.

A bot named ChatGPT recently released a set of programs that could correlate with human language, capable of training and creating a massive data pool for deeper learning models. This has sparked a discussion about the practicality of its application and the robustness of the corresponding ethical and legal norms. New opportunities have also arisen for data-based application promotion in various fields, and it is evident how it can enhance various domains, including technologies, education, communication, etc. However, the uncertainty remains regarding how ethical, impartial, rational, and reasonable the responses will be based solely on language comments. The smaller and newer the input formats, the less controllable the output. This feature has raised considerable public discussion. This sophisticated AI ability could also have scientific applications, but is inapplicable at present due to the risk of bias and insincerity. People consider AI fabrication software apps to enhance creativity; however, the risk of generating biassed, humiliating, or paraphrased images makes them reluctant to share their thoughts on this. There is no doubt that the release of GPT-3.5 is revolutionary for creative labour and industry.

Nevertheless, a wide breadth of academic and even non-academic industries depends on textual labour. These industries will collapse, and the price of severe job unemployment and skills depreciation will be high. Hence, this state alone hints at how sophisticated and delicate intelligent agents are.

7.2. Case Studies of AI in Ethical Decision-Making

To illustrate the practical relevance of the proposed ethical risk framework, this section presents real-world case studies in which AI decision-making raises complex ethical questions. These scenarios reflect how the theoretical model discussed in this study can be applied in diverse domains to evaluate and respond to ethical dilemmas.

7.2.1 Autonomous Vehicles (Self-Driving Cars)

One of the most widely discussed ethical dilemmas involves autonomous vehicles. For example, in an unavoidable accident scenario, an AI must choose between hitting pedestrians or sacrificing the vehicle's passengers. This real-world application echoes the classic "trolley problem" and raises questions about how utilitarian ethics should guide harm minimization. Several car manufacturers have reported working on algorithms that consider risk and harm distribution, illustrating the challenge of embedding ethical principles in technical systems.

7.2.2 Medical Diagnosis Systems

AI-based diagnostic tools, especially those trained on non-diverse datasets, have demonstrated significantly lower accuracy when diagnosing diseases in underrepresented patient groups. Such disparities have led to misdiagnosis and unequal treatment. This case highlights the ethical risk of data bias and the urgent need for fairness, transparency, and oversight in healthcare AI, which is strongly tied to the proposed ethical risk model.

7.2.3 Hiring and Recruitment Algorithms

Multiple corporations have withdrawn AI hiring tools after discovering that their systems disproportionately filtered out female candidates and ethnic minorities. These biases often stemmed from historical training data that reflected discriminatory practices. From a deontological perspective, such violations of equal opportunity rights stress the importance of rule-based ethical enforcement and algorithmic accountability.

7.2.4 Predictive Policing Systems

Law enforcement agencies have adopted predictive policing algorithms to forecast crime-prone areas or individuals. However, several studies have shown that these models reinforce systemic racial biases, disproportionately targeting marginalized communities. Applying virtue ethics here requires reconsideration of the social values encoded in these systems. It challenges the legitimacy of using AI in morally sensitive areas.

7.2.5 Maritime Ethical Dilemma Simulation

A conceptual maritime scenario challenges AI with deciding between colliding with a child in water or steering toward a rock, potentially endangering passengers. This mirrors trolley-type ethical logic under uncertainty. It tests whether the AI can apply moral reasoning that balances responsibility, harm, and survival core concerns addressed by the ethical risk framework.

These case studies demonstrate how the ethical risk framework proposed in this study can be applied to real-life dilemmas in AI deployment. By analyzing technological, human, and contextual dimensions, the framework offers a structured and interdisciplinary approach to identifying, understanding, and mitigating ethical risks.

7.3. Psychological and Social Impacts of AI Decision-Making

AI systems are increasingly incorporated into decision-making processes, their psychological and social effects cannot be overlooked.

Research has shown that humans tend to over-rely on AI decision-makers, often placing undue trust in algorithms without fully understanding their limitations (Dzindolet et al., 2003; Lee & See, 2004).

This over-reliance can lead to shifts in decision-making patterns, potentially diminishing human agency and critical thinking skills. Moreover, AI-driven decisions in sensitive areas like healthcare or law enforcement may alter societal norms, creating new ethical dilemmas regarding fairness and autonomy.

The ethical implications of AI decision-making are vast and complex. Integrating ethical frameworks in AI systems remains a challenge, but it also presents an opportunity to guide AI behaviour in ways that align with human values. As AI continues to evolve, understanding its ethical risks and applying effective governance frameworks will be crucial in mitigating potential harm.

8. Challenges in AI Ethics

The widespread adoption of artificial intelligence across critical domains, such as justice, finance, and public administration, has triggered serious ethical concerns. While AI promises enhanced efficiency, it also poses threats to fairness, accountability, and transparency. Despite growing awareness, many machine learning and deep learning systems are still designed with limited ethical foresight, failing to account for unintended consequences or societal impacts (Srivastava & Rossi, 2018; Ferrer et al., 2020).

This thought is analogous to Gödel's incompleteness theorem, which suggests that within any sufficiently complex logical system, there are true statements that cannot be proven within the system itself.

This philosophical analogy raises profound concerns about the applicability of AI and even mathematics in ethical reasoning. Current works appear as a set of exploratory analyses and cannot provide conclusive, reasonable assessments of AI ethics. A case-by-case approach to ethical concerns seems necessary (Nagel, 2008).

Addressing AI ethics requires more than just post-deployment analysis. It demands a proactive approach that integrates risk evaluation at the design stage, emphasizing explainable, bias mitigation, and the ethical alignment of algorithms. Ethical concerns must go beyond technical functionality to consider broader social, political, and cultural contexts in which these systems operate (Mennella et al., 2024).

8.1. Bias in AI Systems

Bias in AI stems from both technical and societal sources. Datasets often reflect historical prejudices, and algorithms can reinforce or even amplify these inequities during training (Barocas et al., 2019). For instance, facial recognition systems have demonstrated higher error rates for individuals with darker skin tones due to under-representation in training data. Such bias disproportionately affects marginalized groups and can lead to systemic discrimination in domains like hiring, credit scoring, and predictive policing (Buolamwini & Gebru, 2018).

8.2. Transparency and Accountability

Transparency is vital to building trust in AI systems, particularly when decisions significantly affect human lives. Explainability ensures that users, auditors, and affected individuals can understand how and why a decision was made (Doshi-Velez & Kim, 2017). However, many advanced AI models, such as deep neural networks, operate as "black boxes", making their internal logic inaccessible even to developers. This opacity poses a challenge for accountability, especially in legal or medical applications where justification is crucial.

8.3. Privacy Concerns

AI systems rely heavily on data, often gathered from users without explicit consent or awareness. This creates serious privacy risks, especially when data is sensitive, such as health records, biometric identifiers, or behavioural patterns (Zuboff, 2019). The challenge extends to data governance, where poorly defined policies can allow governments or corporations to exploit user data for control or profit. Differential privacy, data minimization, and user-centric consent models are emerging as ethical tools to mitigate these risks (Crawford & Paglen, 2021).

9. Global Perspectives on AI Ethics

Cultural and legal differences worldwide significantly shape the ethical landscape of AI. For instance, the European Union has adopted comprehensive guidelines for ethical AI that emphasize transparency and accountability. At the same time, countries like China focus more on state control and surveillance. In contrast, the United States has a more fragmented approach, with regulations varying between states. This comparative analysis highlights the challenges in creating a unified ethical framework for AI, which must account for these divergent perspectives.

Global perspectives on AI ethics reveal a dynamic and diverse ethical landscape shaped by cultural values, legal systems, economic priorities, and political ideologies. While some nations emphasize individual rights and transparency, others prioritize collective welfare and state control. For instance, the European Union leads the establishment of comprehensive regulatory frameworks like the EU AI Act, which emphasizes human oversight, transparency, risk-based categorization, and accountability in AI systems. The EU's approach is grounded in fundamental rights, seeking to ensure that AI respects human dignity, non-discrimination, and privacy.

In contrast, China adopts a more centralised model that leverages AI for national development and social stability, with a strong emphasis on surveillance technologies and government oversight. Ethical concerns in China focus on ensuring AI aligns with state-defined moral values, social harmony, and national security, which differs markedly from Western notions of liberal individualism. Meanwhile, the United States adopts a market-driven, decentralised approach, where ethical governance varies across industries and states, often guided by corporate self-regulation and sector-specific standards. This fragmented model allows for rapid innovation but has also raised concerns about inconsistencies, a lack of accountability, and insufficient protection of civil liberties.

Other regions also contribute distinct ethical perspectives. Japan and South Korea, for example, integrate traditional values, such as harmony and respect, into their AI ethics frameworks, emphasizing human-centric innovation. In the Global South, countries are increasingly participating in AI governance dialogues. However, they face challenges related to digital inequality, data sovereignty, and the ethical implications of deploying AI systems developed in foreign cultural contexts.

International organizations, such as UNESCO and the OECD, have proposed global ethical principles, including fairness, transparency, inclusiveness, and accountability, aiming to create a shared vocabulary and standards across nations. However, enforcing these principles remains complex due to geopolitical tensions and differing priorities.

This global divergence underscores the difficulty of establishing a universally accepted ethical framework for AI. Any attempt to develop global standards must navigate cultural relativism, legal pluralism, and technological asymmetries. Therefore, fostering international cooperation, mutual learning, and ethical pluralism is essential to building responsible AI systems that respect diverse human values across contexts.

10. Regulatory Frameworks

AI systems increasingly operate in high-impact environments such as healthcare, finance, and law enforcement. Due to their complexity, opacity, and rapid evolution, these systems challenge existing regulatory frameworks. As a result, ethical and legal governance has become a global concern. Governments, international organizations, and research institutions are now working to create adaptive regulatory mechanisms that align AI deployment with democratic values, privacy, and accountability (Gasser & Almeida, 2017).

10.1. Global Perspectives on AI Regulation

Several countries and international bodies have taken steps to establish regulatory frameworks for AI. The European Union's AI Act proposes a risk-based approach, categorizing AI applications by their potential to cause harm and assigning corresponding legal requirements. In contrast, the United States favours a sector-specific and innovation-driven model with lighter oversight. Meanwhile, China's approach is characterized by centralized control, algorithmic transparency requirements, and an emphasis on social harmony (Floridi et al., 2018; Metzinger, 2020). These variations reflect differing legal traditions, governance models, and cultural priorities. Understanding such differences is critical to fostering international collaboration and interoperability in AI systems.

10.2. Ethical Guidelines for AI Development

Over the past decade, numerous ethical AI guidelines have emerged from academia, industry, and international bodies. Despite differences in emphasis, common themes have converged around key principles: transparency, fairness, accountability, privacy, and human oversight. For example, the OECD Principles on AI and UNESCO's 2021 Recommendation on the Ethics of AI emphasize inclusiveness, safety, and sustainability. However, the lack of enforcement mechanisms in most of these frameworks limits their practical impact. There is a growing call to translate high-level ethical principles into operational policies and legally binding obligations (Jobin et al., 2019; Fjeld et al., 2020).

11. The Future of AI and Ethics

AI systems continue to evolve in autonomy and complexity, their ethical impact will grow significantly. Emerging developments such as artificial moral agents, generative AI, and general-purpose learning systems raise pressing questions about accountability, transparency, and moral alignment. Ensuring that future AI systems behave in ethically acceptable ways will require foundational changes in how these technologies are designed, governed, and evaluated.

One important research direction involves formalising ethical reasoning in AI through computational models that simulate normative theories, such as utilitarianism, deontology, or virtue ethics. These models aim to embed moral principles directly into decision-making systems operating in high-stakes environments, such as healthcare, autonomous vehicles, or military applications (Allen et al., 2005). However, significant challenges remain regarding explainability, contextual adaptability, cultural diversity in ethical values, and the technical feasibility of embedding such models in real-world systems.

AI systems become deeply integrated into societal infrastructures, their ethical deployment must be guided by inclusive, transparent, and participatory processes. Guidelines should prioritize fairness, privacy, and accountability while ensuring that AI does not perpetuate systematic bias or harm marginalized populations. Ethical integration also requires sustained collaboration among AI developers, ethicists, civil society organisations, and policymakers to ensure that technologies align with social values, legal norms, and international human rights standards across diverse cultural contexts (Floridi, 2019).

Looking forward, AI ethics must become anticipatory and adaptive, proactively addressing ethical risks before they manifest and evolving along with technological advancement. Rather than treating ethics as an external audit or a final checkpoint, it should be embedded in every phase of the AI system lifecycle – from design and data collection to deployment, monitoring, and decommissioning.

12. The Public's Perception of AI Ethics

Understanding public perception of AI ethics is key to creating socially acceptable technologies. AI systems become more pervasive in everyday life rom healthcare diagnostics to autonomous vehicles and algorithmic content moderation . Public trust plays a vital role in determining how widely and safely they are adopted. Research shows that concerns around bias, loss of control, job displacement, and privacy violations significantly influence the public's attitude toward AI deployment (Zhang & Dafoe, 2019).

12.1. Public Opinion and Trust

Empirical studies indicate that while people are generally optimistic about the potential of AI, they also express deep concern about its ethical consequences. For instance, surveys conducted in the EU and the U.S. show that individuals prefer human oversight in AI decisions, especially in critical sectors like criminal justice, hiring, and healthcare. The degree of public trust often correlates with the transparency and explainability of AI systems, as well as the presence of accountability mechanisms (Miller et al., 2020). Public acceptance is not merely a function of technical performance. Still, it is shaped by values such as fairness, inclusiveness, and democratic oversight.

12.2. The Role of Media in Shaping Perceptions

Media and digital platforms significantly shape how the public perceives AI and its ethical implications. Popular narratives often oscillate between utopian portrayals of AI as a revolutionary force for good and dystopian fears of surveillance, loss of autonomy, or existential threats. Such framings influence how individuals assess the role of AI in society and their level of support for regulation. Sensationalist coverage may exaggerate risks or ignore ethical nuances, whereas responsible reporting can improve public understanding and foster informed debate (Cave et al., 2018). Therefore, public education and media literacy are essential tools for aligning public perception with realistic and evidence-based expectations about AI systems.

13. Interdisciplinary Approaches to AI Ethics

The ethical design and deployment of artificial intelligence systems require interdisciplinary collaboration across multiple domains. Computer scientists alone cannot foresee the wide-ranging societal, cultural, and legal implications of AI decisions. Ethical AI must incorporate perspectives from philosophy, sociology, psychology, law, and policymaking to ensure it aligns with human values and social contexts (Floridi et al., 2018).

An ethics-by-design approach integrates ethical reflection into every stage of AI development – from data collection to algorithm selection and system deployment. This process includes participatory design, where diverse stakeholders contribute to shaping ethical goals and constraints. Interdisciplinary teams help ensure that AI systems remain transparent, accountable, and responsive to the needs of vulnerable populations (Vakkuri et al., 2021).

Furthermore, ethical reasoning in AI must be culturally sensitive. Values such as fairness or autonomy may differ across societies. Without interdisciplinary input, AI risks enforcing biased, Western-centric moral frameworks on global users. A pluralistic, inclusive design process enhances both the legitimacy and social acceptability of intelligent systems (Jobin et al., 2019).

14. Discussion

This study has explored the ethical complexities associated with AI-based decision-making in intelligent systems, focusing on risk factors such as data uncertainty, algorithmic opacity, and limitations in emotional intelligence. These factors raise significant concerns regarding fairness, accountability, and public trust. The analysis has shown that ethical risks are not merely technical issues but are deeply rooted in sociocultural, legal, and philosophical dimensions.

The research findings highlight the importance of incorporating ethical design principles throughout the AI system lifecycle – from data collection and algorithm development to deployment and long-term monitoring. Furthermore, no single discipline can adequately address these ethical challenges. Instead, interdisciplinary cooperation between engineers, philosophers, policymakers, legal experts, and social scientists is essential.

These findings align with the conclusions drawn by several foundational studies. For instance, the emphasis on fairness, transparency, and accountability as core ethical pillars echoes the principles outlined in Jobin et al. (2019) and the OECD guidelines. The proposed risk assessment framework builds upon prior models such as Guan et al. (2022), who extended their approach by integrating a cultural and interdisciplinary dimension. Moreover, the study addresses critiques found in Buolamwini and Gebru (2018) and Barocas et al. (2019) regarding algorithmic bias by incorporating it as a key component of ethical risk. Thus, the discussion here not only reflects existing literature but also offers a broader, more integrative perspective.

15. Recommendations

Based on the analysis conducted in this study, the following targeted recommendations are proposed to address ethical risks in AI-based decision-making systems:

1. Integrate Ethical Risk Assessment in Design Stages: Developers and engineers should embed ethical risk analysis during the design and prototyping phases, ensuring systems to consider bias, fairness, and explainability before deployment.
2. Establish Interdisciplinary AI Ethics Committees: Institutions deploying intelligent systems should form ethics review panels including legal experts, philosophers, AI engineers, and end-users to evaluate the ethical implications of AI decisions.
3. Mandate Transparency in Automated Decision-Making: Policymakers should introduce regulations that require AI systems to disclose decision logic and allow the affected individuals to contest outcomes, especially in high-risk sectors like healthcare and justice.
4. Develop Context-Specific Ethical Frameworks: Ethics cannot be universally imposed; it is crucial to adapt frameworks to local cultural, legal, and social values. Comparative research should continue to refine context-aware models.
5. Monitor AI Decisions Post-Deployment: Ongoing auditing mechanisms should be implemented to track real-world ethical consequences of AI decisions and allow for timely corrections.

These recommendations do not aim to promote general ethical awareness but they also attempt to directly shape how ethical principles influence concrete decision-making processes in AI systems.

16. Future Work

Future research should continue exploring how ethical principles can be formalized, operationalized, and evaluated in intelligent systems. A key challenge lies in translating high-level moral theories into machine-executable frameworks that can handle real-world ambiguity, cultural diversity, and conflicting values. Collaboration between AI engineers and moral philosophers will be critical in designing systems capable of contextual moral reasoning.

Moreover, further investigation is needed for emergent AI models that exhibit self-reflective or adaptive ethical behaviour. These include reinforcement learning agents with embedded ethical constraints and multi-agent systems designed to negotiate ethical trade-offs. Research should also address gaps in current regulatory mechanisms, particularly the lack of tools for assessing long-term ethical consequences prior to deployment.

Finally, AI ethical social impacts should be prioritized. This involves examining how users, institutions, and public opinion affect the ethical outcomes of AI and how to foresee and manage ethical concerns. To ensure AI technologies benefit society, empirical investigations and stakeholder-engaged design are essential.

Future research may complement the present qualitative analysis by incorporating statistical methods to empirically measure the impact of AI on ethical decision-making across various domains. Quantitative studies using surveys, behavioural experiments, or data-driven impact assessments could validate and operationalize the ethical risk framework proposed in this study.

17. Conclusion

In conclusion, aligning AI systems with human values requires more than mere reactive regulation. It calls for proactive, transparent, and inclusive approaches that embed ethics into the core of intelligent system development. As AI continues to evolve, future research must focus on operationalizing ethical principles and developing tools that help evaluate and enforce responsible AI behaviours.

References

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261. https://doi.org/10.1080/09528130050111428

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press. ISBN: 978 0262048613. https://fairmlbook.org/

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics*, 5, 3265–3279. https://doi.org/10.1007/s43681-024-00653-w

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer. https://doi.org/10.1007/978-3-319-60648-4

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency* (pp.77–91). https://proceedings.mlr.press/v81/buolamwini18a.html

Calo, R. (2015). Robotics and the lessons of cyberlaw. *California Law Review*, *103*(3), 513–563. https://doi.org/10.2139/ssrn.2402972

Cave, S., Dihal, K., & Dillon, S. (2018). Portrayals and perceptions of AI and why they matter. *Nature Machine Intelligence*, *1*(2), 1–3. https://www.lcfi.ac.uk/resources/portrayals-and-perceptions-ai-and-why-they-matter

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311–313. https://doi.org/10.1038/538311a

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & Society*, 36, 1–12. https://doi.org/10.1007/s00146-021-01162-8

Daly, A., Hagendorff, T., Hui, L., Mann, M., Marda, V., Wagner, B., Wang, W., & Witteborn, S. (2019). Artificial Intelligence, Governance and Ethics: Global Perspectives. SSRN. https://doi.org/10.2139/SSRN.3414805

Dent, K. (2020). Ethical considerations for AI researchers. https://doi.org/10.48550/arXiv.2006.07558

Dignum, V. (2019). *Responsible Artificial Intelligence: How to develop and use AI responsibly.* Springer. https://doi.org/10.1007/978-3-030-30371-6

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

European Commission – High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy artificial intelligence. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52019DC0168

Farisco, M., Evers, K., & Salles, A. (2020). Towards establishing criteria for the ethical analysis of artificial intelligence. *Science and Engineering Ethics*, 26, 2001–2026. https://doi.org/10.1007/s11948-020-00238-w

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2020). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *39*(3), 72–80. https://doi.org/10.1109/MTS.2021.3056293

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Centre Research Publication*, (2020-1). https://doi.org/10.2139/ssrn.3518482

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-X

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15–38. https://doi.org/10.1007/s13347-023-00643-6

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Gasser, U., & Almeida, V. A. F. (2017). A layered model for AI governance. IEEE Internet Computing, 21(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835.

Guan, H., Dong, L., & Zhao, A. (2022). Ethical Risk Factors and Mechanisms in Artificial Intelligence Decision Making. *Behavioral Sciences*, *12*(9), 343. https://doi.org/10.3390/bs12090343

Herrera-Poyatos, A., Del Ser, J., López de Prado, M., Wang, F.-Y., Herrera-Viedma, E., & Herrera, F. (2025). Responsible artificial intelligence systems: A roadmap to society's trust through trustworthy AI, auditability, accountability, and governance. *arXiv preprint*. https://doi.org/10.48550/arXiv.2503.04739

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2018). Ethically Aligned Design: A vision for prioritizing human well-being with autonomous and intelligent systems (Version 2). IEEE Standards Association. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kluge Corrêa, N., Galvão, C., Santos, J. W., Del Pino, C., Pontes Pinto, E., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & de Oliveira, N. (2022). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, *4*(10), Article 100857. https://doi.org/10.1016/j.patter.2023.100857

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, *10*(4), Article e26297. https://doi.org/10.1016/j.heliyon.2024.e26297

Metzinger, T. (2020). Europe's approach to regulating AI. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang. (Eds.). *The Oxford Handbook of AI Governance*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.001.0001

Miller, T., Howe, P., & Sonenberg, L. (2020). Explainable AI: Understanding, trust, and acceptance. *Artificial Intelligence*, 287, Article 103385. https://doi.org/10.1016/j.artint.2020.103385

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

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

Nagel, E. (2008). *Gödel's Proof* (Rev. ed.). New York University Press. ISBN: 9780814758373.

Radanliev, P., Santos, O., Brandon-Jones, A., & Joinson, A. (2024). Ethics and responsible AI deployment. *Frontiers in Artificial Intelligence*, 7, Article 1377011. https://doi.org/10.3389/frai.2024.1377011

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., ... & Lazer, D. (2019). *Machine behaviour. Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education. ISBN: 9780134610993

Schrage, M., & Kiron, D. (2025, January). The great power shift: How intelligent choice architectures rewrite decision rights. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/the-great-power-shift-how-intelligent-choice-architectures-rewrite-decision-rights/

Seeamber, R., & Badea, C. (2023). If we aim to build morality into an artificial agent, how might we begin to go about doing so?. *IEEE Intelligent Systems*, *38*(6), 35–41. https://doi.org/10.1109/MIS.2023.3320875

Srivastava, B., & Rossi, F. (2018). Towards a composable bias rating of AI services. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. (pp. 284–289). https://doi.org/10.1145/3278721.3278744

Tigard, D. W. (2021). Responsible AI and moral responsibility: A common appreciation. *AI and Ethics*, *1*(2), 113–117. https://doi.org/10.1007/s43681-020-00009-0

Tractenberg, R. E. (2023). What is ethical AI? Leading or participating in an ethical team and/or working in statistics, data science, and artificial intelligence. *SocArXiv*. https://doi.org/10.31235/osf.io/8e6pv

Vakkuri, V., Jantunen, M., Halme, E., Kemell, K. K., Nguyen-Duc, A., Mikkonen, T., & Abrahamsson, P. (2021). The time for the AI (Ethics) maturity model is now. In *Proceedings of the 53rd Hawaii International Conference on System Sciences. arXiv preprint arXiv:2101.12701*. https://doi.org/10.48550/arXiv.2101.12701

van Wynsberghe, A. (2013). Designing robots for care: Care-centred value-sensitive design. *Science and Engineering Ethics*, *19*(2), 407–433. https://doi.org/10.1007/s11948-011-9343-6

Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press. https://academic.oup.com/book/10768

Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends.* Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://doi.org/10.7910/DVN/SGFRYA

Zhang, Y., Wu, J., Yu, F., & Xu, L. (2023). Moral judgments of human vs. AI agents in moral dilemmas. *Behavioral Sciences*, *13*(2), 181. https://doi.org/10.3390/bs13020181

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs. ISBN-13: 978 1610395694.