

ETHICAL PROBLEM OF MORAL RESPONSIBILITY IN ARTIFICIAL INTELLIGENCE: A KANTIAN AND UTILITARIAN ANALYSIS

BY

Rev Fr Dr. John Friday Mordi

Department of Philosophy

University of Delta Agbor

john.mordi@unidel.edu.ng

Phone: 08066558754

Abstract

The rapid development and widespread application of artificial intelligence (AI) have introduced complex ethical challenges that question traditional notions of moral agency, accountability, and responsibility. As AI systems increasingly perform autonomous functions in critical areas such as healthcare, transportation, security, and governance, concerns arise regarding who should be held morally responsible when such systems cause harm or make morally significant decisions. This study examined the ethical problem of moral responsibility in artificial intelligence through the philosophical frameworks of Immanuel Kant's deontological ethics and John Stuart Mill's utilitarianism. Using a conceptual and analytical methodology, the study clarified key ethical concepts such as artificial intelligence, moral responsibility, autonomy, and ethical agency. It critically analyzed Kant's emphasis on duty, intention, rational autonomy, and respect for persons, arguing that AI systems lack the rational self-legislation required for moral agency (Kant, 1985). The work further explores the utilitarian approach, which evaluated AI actions based on consequences and overall social utility, highlighting its relevance in revealing that while policy-oriented and outcome-driven technological contexts (Mill, 1963). The study of both ethical theories offer valuable insights, neither is fully sufficient in addressing the unique moral challenges posed by AI. Kantian ethics strongly safeguards human dignity but struggles with outcome based decision-making, whereas utilitarianism accommodates technological efficiency but risks neglecting individual rights. The study concluded that moral responsibility for AI actions ultimately resides with human agents and institutions, and it recommends a hybrid ethical framework that integrates classical ethical theories with contemporary approaches to AI governance.

Keywords: *Ethics, Artificial intelligence, Moral, Responsibility, Humanity.*

Introduction

The rapid advancement of artificial intelligence (AI) has become one of the most defining features of contemporary society. AI systems now perform tasks that were once exclusively human, including medical diagnosis, financial decision-making, facial recognition, autonomous driving, and predictive policing. These developments raise profound ethical questions, particularly concerning moral responsibility, accountability, and the nature of ethical decision making in technologically mediated contexts. Traditional ethical theories were developed within human-centered moral frameworks, yet AI systems increasingly operate with a degree of autonomy that complicates conventional moral judgments.

Ethics, as a branch of philosophy, seeks to determine what is right or wrong, good or bad, and just or unjust in human actions. The emergence of AI challenges ethical philosophy to reassess its foundational assumptions about agency, intention, and responsibility. When an autonomous system causes harm, it becomes unclear whether responsibility should lie with the programmer, the user, the institution deploying the system, or the AI itself. This problem of moral responsibility forms the core concern of contemporary AI ethics.

This study examined the ethical implications of artificial intelligence through the perspectives of Immanuel Kant's deontological ethics and John Stuart Mill's utilitarianism. Kant emphasizes duty, rationality, and intention as the basis of morality, while Mill focuses on consequences and the maximization of happiness. It is by applying these classical ethical theories to modern AI systems, that this study evaluated their adequacy in addressing moral responsibility in contemporary technological society

Conceptual Clarification

Ethics

Ethics is the philosophical study of morality, concerned with principles that govern human conduct and distinguish right actions from wrong ones. According to Aristotle, ethics aims at the good life and the cultivation of virtue, while modern ethical theories emphasize rules, duties, or consequences (Aristotle, *Nicomachean Ethics*). Ethics provides a normative framework for evaluating actions, policies, and institutions.

Artificial Intelligence

Artificial Intelligence (AI) refers to computer systems designed to perform tasks that typically require human intelligence, such as learning, reasoning, problem-solving, and decision-making. Russell and Norvig (2010) defined AI as the study of agents who perceive their environment and act to maximize the achievement of goals (Russell & Norvig, 2010). Contemporary AI systems often rely on machine learning algorithms that allow them to adapt and improve performance without explicit human instruction.

Moral Responsibility

Moral responsibility refers to the condition under which an agent can be held accountable for actions and their consequences. Traditionally, moral responsibility presupposes consciousness, intention, freedom, and rational deliberation (Strawson, 1962). The attribution of moral responsibility becomes problematic when actions are performed by non-human agents such as AI systems.

Kantian Ethical Perspective on Artificial Intelligence

Immanuel Kant's ethical theory is one of the most influential deontological frameworks in moral philosophy. Kant's ethics is grounded in the belief that morality is rooted in rationality and autonomy rather than in the consequences of actions. For Kant, a morally right action is one performed out of duty and in accordance with moral law, not one motivated by inclination or the pursuit of desirable outcomes (Kant, 1785). This emphasis on duty and rational intention base significant implications for evaluating the ethical status of artificial intelligence.

Central to Kantian ethics is the concept of the categorical imperative, which serves as the supreme principle of morality. Kant formulates the categorical imperative in several ways, the most notable being: "Act only according to that maxim whereby you can at the same time will that it should become a universal law" (Kant, 1785). This principle demands universality, consistency, and rational justification of moral actions. When applied to AI, the question arises whether AI systems can formulate maxims or act from rational moral principles. Since AI operates based on algorithms, data inputs, and programmed objectives, it lacks the capacity for moral reasoning in the Kantian sense.

Another crucial Kantian principle is the formula of humanity, which states that human

beings must always be treated as ends in themselves and never merely as means (Kant, 1985). This principle is particularly relevant in contemporary debates surrounding AI applications such as facial recognition, surveillance technologies, predictive policing, and data mining. When AI systems are deployed in ways that manipulate, exploit, or dehumanize individuals such as reducing persons to data points or statistical probabilities they risk violating Kant's demand for respect for human dignity.

From a Kantian perspective, moral agency requires autonomy, understood as the capacity of rational beings to legislate moral law for themselves. AI systems, no matter how advanced, do not possess self-consciousness or the ability to act from moral duty. They do not choose moral laws; rather, they execute predefined rules or learned patterns. Consequently, AI cannot be regarded as a moral agent capable of bearing moral responsibility. Responsibility instead lies with the human agents involved in the creation, deployment, and governance of AI systems, including programmers, engineers, corporations, and policymakers.

Furthermore, Kant's emphasis on intention rather than consequences complicates the ethical evaluation of AI outcomes. In Kantian ethics, an action that leads to harmful consequences may still be morally permissible if it was performed from a good will. This creates tension when applied to AI systems whose actions are outcome-driven and lack intentional moral motivation. Nonetheless, Kantian ethics provides a strong normative framework for evaluating the duties of human actors in AI development, emphasizing accountability, transparency, and respect for human rights (Kant, 1997).

In sum, Kantian ethics rejects the idea of AI as a moral agent while strongly affirming the moral responsibility of humans. It offers a powerful critique of unethical AI practices by grounding ethical evaluation in human dignity, autonomy, and duty rather than efficiency or utility. Immanuel Kant's moral philosophy is grounded in reason, autonomy, and duty. For Kant, morality is not determined by consequences but by the intention behind an action and its conformity to moral law. The central principle of Kantian ethics is the categorical imperative which commands actions that can be universally applied without contradiction (Kant, 1985).

Kant argues that moral agents must be rational beings capable of understanding and acting according to moral law. Moral worth, therefore, arises only when actions are performed from duty rather than inclination. From a Kantian perspective, AI systems lack genuine rational autonomy because they do not possess self-consciousness or the capacity to legislate moral law for themselves. As such, AI cannot be considered moral agents in the strict Kantian sense.

However, Kantian ethics remains relevant to AI ethics by guiding human responsibility. Programmers and institutions have a duty to ensure that AI systems respect human dignity and do not treat persons merely as means to an end. For example, the use of AI in surveillance or predictive policing may violate Kant's principle of respect for persons if it undermines individual autonomy and rights (Kant, 1997).

Thus, from a Kantian standpoint, moral responsibility for AI actions ultimately rests with human agents who design, deploy, and control these systems.

Utilitarian Ethical Perspective on Artificial Intelligence

Utilitarianism, most prominently articulated by Jeremy Bentham and John Stuart Mill, offers a contrasting ethical framework grounded in consequences rather than duties. According to utilitarian ethics, the moral worth of an action is determined by its ability to produce the greatest happiness for the greatest number (Mill, 1863). Happiness, in this context, is defined as pleasure and the absence of pain, while unhappiness consists of pain and the deprivation of pleasure.

Utilitarianism is particularly appealing in discussions of artificial intelligence because AI

systems are often designed to optimize outcomes, increase efficiency, and maximize benefits. From a utilitarian standpoint, AI applications in healthcare, transportation, education, and economics may be morally justified if they lead to improved accuracy, reduced human error, lower costs, and enhanced overall well-being. For example, AI diagnostic systems that save lives or autonomous vehicles that reduce traffic fatalities can be seen as morally desirable due to their positive consequences.

John Stuart Mill further refines utilitarianism by distinguishing between higher and lower pleasures, arguing that intellectual and moral pleasures are superior to mere physical satisfaction (Mill, 1863). This distinction allows utilitarianism to address concerns about quality of outcomes, not just quantity. Applied to AI, this suggests that ethical evaluation should consider not only efficiency but also impacts on human creativity, autonomy, and moral development.

Utilitarianism also introduces the distinction between act utilitarianism and rule utilitarianism. Act utilitarianism evaluates each individual action based on its consequences, while rule utilitarianism assesses moral rules based on their general utility. In AI ethics, rule utilitarianism is often more applicable, as it supports the development of ethical guidelines and regulations that promote long-term social welfare. For instance, rules governing data privacy, algorithmic transparency, and bias reduction can be justified on utilitarian grounds because they enhance trust and reduce harm across society.

Despite its strengths, utilitarianism faces serious criticisms when applied to AI. One major concern is that utilitarian reasoning may justify morally questionable practices if they produce beneficial outcomes overall. For example, mass surveillance systems may increase public safety but at the cost of individual privacy and freedom. Critics argue that such practices risk sacrificing minority rights for majority benefit, a problem famously highlighted by opponents of utilitarianism such as Bernard Williams (Smart & Williams, 1973).

Moreover, utilitarianism struggles with the problem of predicting consequences, especially in complex AI systems whose long-term social impacts are uncertain. Bias in training data unforeseen algorithmic behavior, and unequal distribution of benefits complicate utilitarian calculations. Nonetheless, utilitarian ethics remains influential in AI policy and governance due to its practical orientation and compatibility with cost-benefit analysis. In conclusion, utilitarian ethics provides a pragmatic and outcome-focused framework for evaluating AI systems. While it lacks the strong rights-based protections found in Kantian ethics, it offers valuable tools for assessing the social impact and overall benefits of artificial intelligence. Utilitarianism, as developed by Jeremy Bentham and refined by John Stuart Mill, evaluates actions based on their consequences. According to Mill, the morally right action is that which produces the greatest happiness for the greatest number (Mill, 1863), Happiness, in this context, refers to pleasure and the absence of pain.

Utilitarianism offers a flexible framework for evaluating AI systems by focusing on outcomes rather than intentions. AI applications in healthcare, for instance, may be morally justified if they improve diagnostic accuracy, reduce costs, and save lives, even if they involve risks or errors. From a utilitarian perspective, the moral value of AI lies in its capacity to enhance overall well-being.

However, utilitarianism also faces challenges. AI systems may produce beneficial outcomes on average while still causing significant harm to minorities. Critics argue that utilitarian reasoning may justify unethical practices, such as mass surveillance, if they increase overall security and happiness (Smart & Williams, 1973).

Despite these limitations, utilitarian ethics remains influential in policy-making and AI

governance due to its emphasis on measurable outcomes and social utility.

Artificial Intelligence and Moral Responsibility

The problem of moral responsibility in artificial intelligence lies at the heart of contemporary ethical debates in philosophy of technology. As AI systems increasingly operate with a degree of autonomy in decision-making processes that have significant moral and social consequences, traditional frameworks of moral responsibility are put under pressure. Moral responsibility has historically been attributed to human agents who possess rationality, intentionality, freedom of choice, and moral awareness. The emergence of AI challenges these assumptions by introducing non-human agents capable of influencing human lives in profound ways.

Moral responsibility generally presupposes that an agent is capable of understanding moral norms and freely choosing to act in accordance with or against them (Strawson, 1962). In classical ethics, responsibility is grounded in the agent's intentions and capacity for moral reflection. Artificial intelligence systems, however, operate through algorithmic processes, statistical learning, and data-driven optimization. Although such systems may simulate intelligent behavior, they do not possess consciousness, self-awareness, or genuine moral understanding. This raises the fundamental question of whether AI can ever be considered morally responsible or whether responsibility must always remain with human actors.

One of the central issues in this debate is the distinction between moral agency and causal agency. AI systems undoubtedly function as causal agents: they initiate actions, make predictions, and influence outcomes. However, moral agency requires more than causal efficacy, it requires the capacity for moral judgment and accountability. Most philosophers argue that AI lacks moral agency because it does not act from reasons it can morally endorse, nor can it be held accountable through praise, blame, or punishment (Floridi, 2013). As a result, AI systems cannot be said to bear moral responsibility in the same way humans do.

Nevertheless, the increasing autonomy of AI systems gives rise to what has been described as the "responsibility gap." Andreas Matthias uses this term to describe situations in which harm is caused by AI systems, yet no clear human agent can be held fully responsible (Matthias, 2004). For example, in machine-learning systems, developers may not be able to predict specific outcomes due to the system's adaptive nature. Users, on the other hand, may rely on AI recommendations without fully understanding how decisions are made. This diffusion of responsibility complicates traditional moral and legal accountability structures.

Responsibility for AI-driven actions is therefore often distributed rather than centralized. It may involve programmers who design algorithms, data scientists who select training data, corporations that deploy AI systems, and governments that regulate their use. Each of these actors plays a role in shaping the behavior of AI, making moral responsibility collective rather than individual. This distributed responsibility challenges classical ethical models that focus on isolated moral agents and calls for new approaches to shared accountability.

A particularly controversial issue arises in high-stakes AI applications such as autonomous weapons systems, self-driving vehicles, and algorithmic decision-making in criminal justice. In the case of autonomous vehicles, for instance, moral dilemmas such as accident scenarios force designers to encode ethical priorities into machines. When harm occurs, it becomes difficult to determine whether responsibility lies with the manufacturer, the software developer, the owner, or the regulatory authority. Such cases reveal the inadequacy of simplistic blame attribution AI ethics.

From a Kantian perspective, moral responsibility cannot be transferred to AI systems

because they lack rational autonomy and moral intention. Kant's emphasis on human dignity implies that humans must remain accountable for technologies they create and control. Treating AI as morally responsible risks undermining human responsibility and moral agency (Kant, 1785) Kantian ethics thus insists on strict human oversight and moral accountability in AI governance.

Utilitarian ethics, by contrast, approaches moral responsibility in AI through an assessment of outcomes and social utility. From this perspective, responsibility may be assigned based on which actors are best positioned to prevent harm and maximize beneficial consequences (Mill 1863). This approach supports regulatory frameworks and institutional responsibility rather than focusing solely on individual blame. However, utilitarianism risks reducing moral responsibility to efficiency calculations, potentially neglecting justice and individual rights.

Contemporary scholars have proposed alternative frameworks to address the moral responsibility gap in AI. Luciano Floridi's information ethics emphasizes the moral value of informational entities and environments, arguing for ethical stewardship of the digital ecosystem (Floridi 2014). Others advocate for accountability by design, where ethical principles such as transparency, explainability, and fairness are embedded into AI systems from the outset. These approaches seek to ensure that responsibility remains traceable and enforceable even in complex technological systems.

In conclusion, artificial intelligence does not eliminate moral responsibility but rather transforms the way it must be understood and assigned. AI systems cannot be held morally responsible in the traditional sense, yet their growing influence demands robust ethical frameworks that clarify human accountability. Addressing moral responsibility in AI requires a combination of philosophical analysis, ethical theory, and institutional regulation. Only by maintaining human moral responsibility at the centre of AI development can society ensure that technological progress aligns with ethical values.

The question of moral responsibility in AI ethics revolves around whether AI systems can be considered moral agents or merely tools. While some scholars argue that advanced AI may eventually acquire moral agency, most contemporary philosophers maintain that AI lacks consciousness and intentionality, which are essential for moral responsibility (Floridi, 2013). Responsibility for AI actions can be distributed across multiple human factors, including developers, data scientists, users, corporations, and regulatory institutions. This distributed responsibility creates what is often referred to as the "responsibility gap" in AI ethics (Matthias, 2004).

For example, in the case of autonomous vehicles involved in accidents, it is difficult to assign blame solely to one agent. Ethical responsibility may involve shared accountability and the development of robust regulatory frameworks to ensure transparency and oversight.

Critical Evaluation

The ethical analysis of artificial intelligence through Kantian and utilitarian frameworks reveals both the enduring relevance and the significant limitations of classical ethical theories when applied to contemporary technological realities. These theories provide valuable normative guidance, they were developed within human-centered moral contexts and therefore struggle to fully capture the complexity, autonomy, and scale of modern AI systems.

From a Kantian perspective, the strength of deontological ethics lies in its unwavering commitment to human dignity, autonomy, and moral duty. Kant's insistence that persons must always be treated as ends in themselves offers a powerful moral safeguard against exploitative and dehumanizing uses of artificial intelligence, particularly in areas such as surveillance, data extraction, and algorithmic discrimination (Kant, 1785). Kantian ethics

thus functions as a moral boundary-setting framework, ensuring that technological efficiency does not override fundamental human rights.

However, Kantian ethics also faces serious limitations in the context of AI. Its strict emphasis on intention and duty makes it difficult to address situations where outcomes are morally significant but intentions are indirect or diffused. In AI systems, harmful consequences may arise without any clear malicious intent from human actors. The rigidity of Kantian universalism may also limit its ability to respond flexibly to rapidly evolving technological contexts. As such, Kantian ethics risks being normatively strong but practically constrained.

Utilitarian ethics, on the other hand, demonstrates considerable practical appeal in the evaluation of artificial intelligence. Its focus on consequences, efficiency, and overall social benefit aligns closely with the goals of technological innovation and public policy. Utilitarian reasoning is particularly useful in large-scale decision-making contexts such as healthcare, resource allocation, traffic management, and economic optimization, where AI systems promise measurable benefits (Mill, 1863). Despite its strengths, utilitarianism is vulnerable to moral criticism. Prioritizing aggregate happiness, utilitarian ethics may justify practices that infringe upon individual rights, privacy and freedom, especially when minority interests conflict with majority benefits. In AI ethics, this danger is evident in justifications of mass surveillance, predictive policing, and data profiling. Critics argue that such practices risk normalizing injustice under the guise of social utility (Smart & Williams, 1973).

Both ethical theories also struggle with the problem of moral responsibility diffusion in AI systems. Neither Kantian nor utilitarian ethics fully resolves the responsibility gap created by autonomous and self-learning technologies. Kantian ethics places responsibility squarely on human agents but offers limited guidance on how responsibility should be distributed among multiple actors. Utilitarianism, while more flexible, may dilute responsibility by focusing primarily on outcomes rather than accountability.

These limitations suggest the need for a complementary or hybrid ethical framework. Contemporary scholars argue that AI ethics must integrate elements of deontological respect for persons with utilitarian concern for social welfare, alongside modern principles such as transparency, explainability, and fairness (Floridi, 2014). Such an approach acknowledges the moral insights of classical philosophy while adapting them to contemporary technological conditions.

Both Kantian and utilitarian ethics offer valuable insights into AI ethics but also exhibit limitations. Kantian ethics provides a strong foundation for protecting human dignity and rights but struggles to address complex outcome-based scenarios common in AI applications. Utilitarianism, while pragmatic and outcome-oriented, risks sacrificing individual rights for collective benefit.

Contemporary ethicists such as Luciano Floridi advocate for an information ethics approach that integrates elements of traditional ethical theories while addressing the unique challenges of digital environments (Floridi, 2014). A hybrid ethical framework may therefore be necessary to adequately address the moral complexities of artificial intelligence.

Conclusion

This study has critically examined the ethical problem of artificial intelligence and moral responsibility through the philosophical lenses of Immanuel Kant's deontological ethics and John Stuart Mill's utilitarianism. The study has demonstrated that the rise of AI poses significant challenges to traditional ethical concepts such as moral agency, intention, and responsibility.

The analysis shows that artificial intelligence, despite its increasing autonomy, cannot be regarded as a moral agent in the strict philosophical sense. AI systems lack consciousness, rational self-legislation, and moral intentionality, which are essential conditions for moral responsibility. Consequently, moral responsibility for AI-driven actions must remain with human agents and institutions involved in the design, deployment, and regulation of these technologies.

Kantian ethics contributes a strong normative foundation for AI ethics by emphasizing human dignity, autonomy, and moral duty. It serves as a moral safeguard against the instrumentalization of persons and highlights the ethical obligations of those who create and use AI systems. Utilitarian ethics, by contrast, offers a pragmatic and outcome-oriented approach that is valuable for evaluating the social impact and benefits of AI technologies. However, both frameworks, when applied independently, prove insufficient in addressing the full complexity of AI ethics.

This paper therefore concluded that a comprehensive approach to AI ethics must move beyond exclusive reliance on any single ethical theory. A hybrid ethical framework that combines respect for human rights with concern for social welfare, supported by robust regulatory and institutional mechanisms, is necessary to address the moral challenges of artificial intelligence. Such an approach ensures that technological progress is guided by ethical responsibility and human values rather than by efficiency and innovation alone.

In conclusion, artificial intelligence does not diminish the relevance of moral philosophy, rather it reinforces the need for ethical reflection in an increasingly technological world. It is by grounding AI development and governance in sound ethical principles that society can ensure that artificial intelligence serves humanity in a morally responsible and just manner.

Artificial intelligence presents one of the most significant ethical challenges of the modern age. Through a Kantian and Utilitarian analysis, this paper has shown that classical ethical theories remain relevant but insufficient on their own to address the complexities of AI and moral responsibility. While AI systems cannot be considered moral agents, human responsibility remains central to ethical AI development and deployment.

The future of AI ethics depends on the integration of philosophical theory, technological innovation, and ethical governance. Grounding AI development in robust ethical principles, positions the society on the part to harness the benefits of artificial intelligence while minimizing its moral risks.

References

- Aristotle, (1989). *Nicomachean Ethics*. Translated by W. D. Ross. Oxford: Oxford University Press.
- Bentham, J. (1989). *An Introduction to the Principles of Morals and Legislation*. Oxford Clarendon Press.
- Floridi, L. (2013). *The Ethics of Information*. Oxford: Oxford University Press.
- Floridi, L. (2014). *The Fourth Revolution: How The Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press.
- Kant, I. (1985). *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press.
- Kant, I. (1997). *The Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press.
- Matthias, A. (2004). "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6, No. 3, 175-183.
- Mill, J. S. (1963). *Utilitarianism*. London: Parker, Son, and Bourn.
- Russell, S., Peter, N. (2010). *Artificial Intelligence: A Modern Approach*. 3rd ed. New

Jersey: Prentice Hall.

Smart, J. J. C., and Bernard W, (1973). *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Strawson, P. F. (1962). "Freedom and Resentment." *Proceedings of the British Academy* 48.