AI and Ethics: Tackling AI Ethics Through Utilitarianism and Virtue Ethics

Dr. Sk Nijamatulla
Assistant Professor (Guest Faculty)
School of Philosophy
Gangadhar Meher University
snijamatulla@gmail.com

Abstract

Artificial Intelligence (AI) has become one of the most influential technologies of the modern world. It is widely used in areas such as healthcare, finance, education, transportation, and communication. While AI systems can improve efficiency and solve complex problems, they also raise important ethical concerns. Issues such as algorithmic bias, lack of transparency, privacy risks, and the potential replacement of human labour have prompted scholars to question how AI should be developed and used responsibly.

One way to examine the ethical challenges of AI is through classical moral philosophy. Although these philosophical theories were developed long before modern technologies, they still provide valuable tools for evaluating contemporary ethical issues. Two particularly relevant approaches are utilitarianism and virtue ethics. Utilitarianism evaluates actions based on their consequences and seeks to promote the greatest good for the greatest number of people. Virtue ethics, by contrast, focuses on the character and moral responsibilities of individuals.

This article explores how these two philosophical perspectives can help address AI ethics. It first discusses major ethical concerns associated with AI. It then examines the key ideas of utilitarianism and virtue ethics and their relevance to AI development. The article argues that utilitarianism helps evaluate the social consequences of AI technologies, while virtue ethics guides the ethical behaviour of developers and institutions. Together, these frameworks offer a balanced approach to responsible AI governance.

Keywords: Artificial Intelligence, AI ethics, utilitarianism, virtue ethics, future of AI

1. Introduction

Artificial Intelligence (AI) has rapidly developed over the past few decades and has become a central part of modern technological progress. Today, AI systems are used in many different sectors, including healthcare, finance, education, transportation, and communication. These systems are capable of processing large amounts of data, identifying patterns, and making decisions that once required human intelligence. Because of these capabilities, AI has the potential to improve efficiency, increase productivity, and solve complex problems that were previously difficult for humans to manage (Russell & Norvig, 2021).

However, the growing influence of AI has also raised serious ethical concerns. As AI systems become more involved in decision-making processes, they can significantly affect people's lives, opportunities, and rights. For example, algorithms are now used in hiring processes, credit scoring systems, medical diagnosis, and law enforcement. When these systems make mistakes or reflect biases present in their training data, they can lead to unfair or harmful outcomes. In some cases, individuals may not even understand how these systems reach their decisions, which creates further concerns about transparency and accountability.

One well-known example of this problem is algorithmic bias. AI systems learn from data that often reflects existing social inequalities. If the data used to train an AI system contains biased patterns, the system may reproduce those patterns in its decisions. Research has shown that some facial recognition technologies have higher error rates for people with darker skin tones, which can lead to discriminatory outcomes in areas such

as security or law enforcement (Buolamwini & Gebru, 2018). Such examples highlight the importance of considering ethical issues when developing and using AI technologies.

Privacy is another major concern related to AI. Many AI systems rely on large datasets that include personal information about individuals. This information may involve online behaviour, health records, financial data, or location tracking. If these datasets are misused or poorly protected, individuals may lose control over their personal information. As a result, questions about data protection, surveillance, and digital rights have become central to discussions about AI ethics.

In addition to bias and privacy concerns, AI technologies also raise questions about accountability and responsibility. When an AI system causes harm or makes a harmful decision, it is often unclear who should be held responsible. Responsibility may lie with the developers who created the algorithm, the organization that deployed it, or the institutions that regulate its use. This complexity makes it difficult to apply traditional ethical and legal frameworks to modern AI systems.

Because of these concerns, scholars and policymakers have increasingly emphasized the need for ethical guidelines in AI development. AI ethics refers to the set of principles and moral frameworks that aim to guide the design, development, and use of artificial intelligence in ways that respect human values and promote social well-being (Floridi et al., 2018). Many organizations and international institutions have proposed ethical principles for AI, such as fairness, transparency, accountability, and respect for human rights (Jobin, Ienca, & Vayena, 2019). While these principles provide important guidance, they often require deeper philosophical foundations to explain why these values are important and how they should be applied in practice.

This is where classical moral philosophy becomes useful. Philosophical traditions that were developed centuries ago continue to provide valuable tools for thinking about modern ethical problems. Ethical theories help us understand how to evaluate actions, responsibilities, and social consequences. By applying these theories to new technological contexts, scholars can develop clearer frameworks for ethical decision-making in AI.

Two philosophical approaches that are particularly relevant to AI ethics are utilitarianism and virtue ethics. Utilitarianism is a consequentialist theory associated with philosophers such as Jeremy Bentham and John Stuart Mill. According to utilitarianism, the moral value of an action depends on its consequences, and the ethically correct action is the one that produces the greatest happiness or well-being for the greatest number of people (Mill, 1863/2001). When applied to AI, utilitarianism encourages policymakers and developers to evaluate technologies based on their overall social benefits and harms.

Virtue ethics, on the other hand, has its roots in the philosophy of Aristotle. Instead of focusing primarily on consequences, virtue ethics emphasizes the importance of moral character and the cultivation of virtues such as honesty, responsibility, fairness, and wisdom (Aristotle, trans. 2009). In the context of AI, virtue ethics shifts attention to the individuals and organizations that design and implement technological systems. It encourages developers and policymakers to act responsibly and consider the broader impact of their work on society.

Both of these ethical traditions offer valuable insights for addressing the challenges created by AI. Utilitarianism provides a framework for evaluating the overall consequences of AI technologies, helping society determine whether their benefits outweigh their potential harms. Virtue ethics, meanwhile, emphasizes the moral responsibilities of those who develop and deploy AI systems, encouraging ethical behaviour and responsible innovation.

This article examines how these two philosophical frameworks can help address the ethical challenges of artificial intelligence. It begins by exploring the major ethical concerns related to AI. It then discusses the principles of utilitarianism and virtue ethics and examines how these theories can be applied to modern technological contexts. Finally, the article argues that combining these two approaches can provide a more comprehensive and balanced framework for AI ethics.

2. Understanding AI Ethics

Artificial Intelligence (AI) is no longer limited to research laboratories or experimental technologies. It is now widely used in everyday life and plays an important role in many social, economic, and political systems. AI technologies are used in search engines, social media platforms, healthcare systems, financial markets, and even in public administration. Because these systems influence important decisions, questions about their ethical use have become increasingly significant. The study of AI ethics focuses on understanding and addressing the moral issues that arise from the development and use of artificial intelligence.

AI ethics can be defined as the field that examines how AI technologies should be designed, developed, and used in ways that respect human values and promote social well-being. It seeks to ensure that AI systems operate in ways that are fair, transparent, accountable, and beneficial to society. As AI becomes more powerful and widely adopted, it becomes necessary to think carefully about the ethical principles that should guide its development. Without ethical guidance, AI technologies could create serious social problems or reinforce existing inequalities (Floridi et al., 2018).

One of the most widely discussed issues in AI ethics is algorithmic bias. AI systems learn from large datasets, and these datasets often reflect historical social patterns and inequalities. If the data used to train an AI system contains biased information, the system may produce biased outcomes. For example, AI tools used in hiring processes might unintentionally favour certain groups over others if the training data reflects past hiring patterns that were discriminatory. Similarly, facial recognition systems have been shown to perform less accurately for individuals with darker skin tones compared to those with lighter skin tones (Buolamwini & Gebru, 2018). Such biases can lead to unfair treatment and reinforce existing social inequalities.

Another important ethical concern is transparency. Many AI systems operate as complex algorithms that are difficult for humans to understand. These systems are sometimes described as “black boxes” because it is not always clear how they arrive at particular decisions. When AI systems are used to make decisions that affect people’s lives, such as loan approvals, medical diagnoses, or legal judgments, it becomes important for individuals to understand how those decisions were made. A lack of transparency can reduce trust in AI systems and make it difficult to challenge or correct mistakes.

Closely related to transparency is the issue of accountability. When AI systems cause harm or make incorrect decisions, it can be difficult to determine who should be held responsible. Traditional systems of responsibility usually assume that decisions are made by human agents. However, AI systems can operate autonomously or semi-autonomously, which complicates questions of responsibility. For instance, if an autonomous vehicle causes an accident, it may not be immediately clear whether the responsibility lies with the software developers, the vehicle manufacturer, the user, or the regulatory authorities. Addressing this issue requires clear ethical guidelines and legal frameworks that define responsibility in the context of AI technologies.

Privacy is another major concern in AI ethics. Many AI systems rely on large amounts of data to function effectively. This data often includes personal information about individuals, such as their browsing habits, location data, medical records, or financial history. If this data is collected or used without proper safeguards, it may lead to violations of personal privacy. Furthermore, the increasing use of surveillance technologies powered by AI has raised concerns about the potential misuse of data by governments or corporations. Protecting individuals’ privacy while still allowing technological innovation is therefore a major challenge in AI ethics.

Another ethical issue relates to automation and its impact on employment. AI systems have the ability to perform tasks that were previously carried out by humans. In some cases, automation can increase efficiency and reduce costs, which benefits businesses and consumers. However, it can also lead to job displacement in certain industries. Workers in manufacturing, transportation, and administrative roles may face significant changes as AI systems become more capable. This raises broader ethical questions about social justice, economic inequality, and the responsibility of governments and organizations to support workers during technological transitions.

In addition to these concerns, AI technologies may also influence human autonomy and decision-making. Recommendation systems used by social media platforms and online marketplaces can shape what information people see and what choices they make. While these systems can improve user experience by providing personalized suggestions, they can also influence behaviour in subtle ways. Some scholars worry that excessive reliance on AI systems could reduce human autonomy or encourage passive decision-making.

Because of these complex challenges, many governments, international organizations, and research institutions have begun to develop ethical frameworks for AI. These frameworks often emphasize core principles such as fairness, transparency, accountability, privacy, and human oversight. For example, global surveys of AI guidelines show that these principles appear consistently in many international discussions about responsible AI development (Jobin et al., 2019).

However, while these principles are widely accepted, applying them in real-world situations can be difficult. Ethical principles often need deeper philosophical foundations to guide decision-making when conflicts arise. For instance, there may be situations where improving efficiency conflicts with protecting privacy, or where increasing safety might require collecting more personal data. In such cases, ethical theories can help clarify how different values should be balanced.

This is why classical philosophical traditions remain important in discussions about AI ethics. Ethical theories provide structured ways of thinking about moral problems and evaluating different courses of action. By applying these theories to modern technological contexts, scholars and policymakers can develop more consistent and thoughtful approaches to ethical decision-making.

Two philosophical approaches that are particularly relevant to AI ethics are utilitarianism and virtue ethics. Utilitarianism focuses on evaluating actions based on their consequences for overall well-being, while virtue ethics emphasizes the moral character and responsibilities of individuals involved in decision-making. Together, these perspectives provide useful tools for examining the ethical challenges created by artificial intelligence.

3. Utilitarianism and AI Ethics

Utilitarianism is one of the most influential ethical theories in modern philosophy. It evaluates actions based on their consequences and seeks to maximize happiness or well-being for the greatest number of people. The theory was developed primarily by Jeremy Bentham and John Stuart Mill (Bentham, 1789/1996; Mill, 1863/2001).

Bentham argued that human beings naturally seek pleasure and avoid pain. Therefore, moral decisions should aim to maximize pleasure and minimize suffering. Mill expanded this idea by emphasizing that some forms of happiness, such as intellectual and moral pleasures, are more valuable than others.

Utilitarianism is particularly relevant to technological ethics because it focuses on the overall social consequences of actions. When evaluating technologies that affect large populations, utilitarian reasoning asks whether the benefits outweigh the harms.

In the context of AI, utilitarian thinking encourages policymakers and developers to consider whether AI technologies improve overall well-being. For example, AI systems used in healthcare can assist doctors in diagnosing diseases more accurately. Early detection of diseases such as cancer can significantly improve patient outcomes and save lives.

Similarly, autonomous vehicles may reduce road accidents caused by human error. If these technologies improve public safety and reduce fatalities, they may produce significant benefits for society.

However, utilitarian reasoning must also consider potential harms. Predictive policing systems, for example, may improve efficiency in law enforcement but may also reinforce existing biases in policing data. If such systems disproportionately target certain communities, the social harm may outweigh their benefits.

Automation also raises utilitarian concerns. AI technologies may increase productivity and economic growth but can also lead to job displacement. Policymakers must therefore balance technological benefits with measures such as retraining programs and social support for affected workers.

Although utilitarianism provides valuable insights, it also faces limitations. Focusing solely on overall happiness may overlook issues of fairness or individual rights. A decision that benefits the majority might still harm a minority group. Therefore, utilitarian reasoning is often complemented by other ethical frameworks.

4. Virtue Ethics and AI Development

Virtue ethics offers a different perspective from utilitarianism. Rather than focusing primarily on consequences, virtue ethics emphasizes the moral character of individuals. The theory originates in Aristotle's philosophy, particularly in his work *Nicomachean Ethics* (Aristotle, trans. 2009).

According to Aristotle, the goal of human life is eudaimonia, often translated as flourishing or living well. This goal is achieved through the cultivation of virtues such as honesty, courage, justice, and wisdom.

Virtues represent balanced character traits that guide individuals toward ethical behaviour. Aristotle described this balance through the concept of the golden mean, where virtue lies between extremes of excess and deficiency.

Another important concept in virtue ethics is practical wisdom (*phronesis*). Practical wisdom enables individuals to make sound moral judgments in complex situations where rules alone may not provide clear guidance.

In the context of AI, virtue ethics highlights the importance of the individuals and institutions that develop these technologies. AI systems do not operate independently of human decisions; they are created and managed by people whose values shape technological outcomes.

Developers and engineers therefore have ethical responsibilities beyond technical performance. Virtues such as fairness, responsibility, honesty, and transparency should guide AI development. Ethical developers must actively address potential harms, such as algorithmic bias or privacy violations.

Organizational culture also plays an important role. Technology companies and research institutions influence the ethical behaviour of their employees. Institutions that encourage ethical reflection and accountability are more likely to develop technologies that respect human values (Jobin et al., 2019).

Education is another key factor. Many universities now include ethics courses in computer science and engineering programs to help students understand the social impact of technology.

By emphasizing character and responsibility, virtue ethics highlights the human dimension of technological innovation. Ethical AI development requires not only technical expertise but also moral awareness and responsible decision-making.

5. Integrating Utilitarianism and Virtue Ethics

Utilitarianism and virtue ethics provide complementary perspectives for addressing AI ethics. Utilitarianism focuses on the consequences of AI technologies, while virtue ethics focuses on the character and responsibilities of those who develop them.

Utilitarian reasoning helps policymakers evaluate whether AI systems produce overall social benefits. For example, AI technologies used in healthcare or environmental monitoring may improve public welfare.

However, utilitarianism alone may overlook concerns about fairness or individual rights. Virtue ethics addresses this limitation by emphasizing moral responsibility and ethical character (Hursthouse & Pettigrove, 2018).

Combining these approaches allows ethical evaluation to occur at two levels. Utilitarianism assesses the social outcomes of AI systems, while virtue ethics guides the behaviour of developers and institutions.

For instance, when implementing AI in healthcare diagnostics, utilitarian analysis would examine whether the technology improves patient outcomes. At the same time, virtue ethics would require developers to ensure transparency, protect patient data, and acknowledge the system's limitations.

This integrated approach is reflected in many contemporary AI ethics guidelines, which emphasize both social impact assessments and responsible professional conduct (Floridi et al., 2018; Jobin et al., 2019).

6. Challenges and Future Directions

Despite the usefulness of ethical frameworks, applying them to AI technologies remains challenging. One difficulty is measuring well-being and social consequences. Utilitarian evaluation requires comparing benefits and harms, but these outcomes may be difficult to quantify (Bostrom & Yudkowsky, 2014).

Another challenge involves predicting long-term consequences. AI technologies evolve rapidly, and their broader social effects may not be immediately apparent.

Global differences in cultural values also complicate the development of universal AI ethics guidelines. Different societies may prioritize values such as privacy, security, or economic development differently (Jobin et al., 2019).

Furthermore, many advanced AI systems are developed by large technology companies with significant influence. Ensuring that these companies prioritize ethical responsibilities requires effective governance and regulation.

Addressing these challenges requires interdisciplinary collaboration among philosophers, engineers, policymakers, and social scientists. Ethical education, regulatory frameworks, and public engagement will also play crucial roles in shaping responsible AI development.

7. Conclusion

Artificial Intelligence is transforming modern society, offering both significant benefits and serious ethical challenges. Issues such as bias, privacy, accountability, and automation demonstrate that technological innovation must be accompanied by ethical reflection.

This article has shown how classical moral philosophy can contribute to discussions about AI ethics. Utilitarianism helps evaluate the social consequences of AI technologies and encourages decisions that promote collective well-being (Mill, 2001). Virtue ethics emphasizes the moral responsibilities of individuals and institutions involved in technological development (Aristotle, trans. 2009).

By integrating these two perspectives, a more comprehensive framework for AI ethics can be developed. Utilitarianism evaluates the outcomes of AI systems, while virtue ethics ensures responsible behaviour among those who create them.

As AI technologies continue to evolve, ethical reflection will remain essential. Through interdisciplinary collaboration, responsible governance, and ethical education, societies can ensure that artificial intelligence develops in ways that respect human dignity and promote the well-being of humanity.

References

Aristotle. (2009). *Nicomachean ethics* (W. D. Ross, Trans.). Oxford University Press. (Original work published ca. 350 BCE)

Bentham, J. (1996). *An introduction to the principles of morals and legislation*. Oxford University Press. (Original work published 1789)

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316–334). Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1–15.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (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

Hursthouse, R., & Pettigrove, G. (2018). Virtue ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2018 ed.). Stanford University. https://plato.stanford.edu/entries/ethics-virtue/

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

Mill, J. S. (2001). *Utilitarianism* (2nd ed.). Hackett Publishing. (Original work published 1863)

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