COMP4920 Essay 1 (Question 2)

Adrian Balbalosa, z5397730

September 2024

1 Introduction

Kantian ethics, with its principle of treating all rational beings as inherently equal offers a conceptually appealing framework for moral decision-making. Its emphasis on consistency and fairness makes it an appealing candidate for designing an automated ethics, particularly in the development of Artificial Moral Agents (AMAs). However, the limitations of Kantian ethics become evident when addressing complex ethical dilemmas. This essay argues that while Kantian ethics presents an interesting foundation for automated ethics due to its rule-based nature, it faces challenges, including issues of a lack of empathy and ambiguous accountability, which must not be overlooked.

2 An Assessment of Kantian Ethics

Kantian Ethics is a deontological ethical theory focused on the principle that actions are morally right if they are done in accordance with universal moral laws (Tonkens 2009, p. 426). The central idea of this theory is the categorical imperative, which states that "one should only act according to the maxims which that can at the same time be willed as a universal law" (Tonkens 2009, p. 427). Meaning that the rules governing one's action should be applicable to everyone, without contradiction. Kantian ethics places emphasis on the importance of intention, asserting that the moral worth of an action lies not in its consequences, but from its motivation by a sense of duty (Tonkens 2009, p. 428). Another key principle is that rational beings must be treated as an ends, rather than a means to an end (Tonkens 2009, p. 427), recognising the intrinsic dignity of each individual. This ethical framework values reason, autonomy and consistency, promoting actions guided by a sense of duty rather than emotions or consequences.

A notable strength of Kantian Ethics is that it places emphasis on respect for

the individual, grounded in the ideal that humans as rational agents possess intrinsic worth. Kant argues that because we have the capacity for rational behaviour and can act independently beyond our impulses, we must always be treated as an ends in ourselves, never merely as a means (Bennet 2015, p. 77). This ensures the protection of human dignity and rights, such that individuals are valued for their capacity for rational thought.

However, Kantian Ethics faces significant challenge when duties conflict, as it provides no clear guidance for how to resolve those dilemmas. A classical example used by critics of Kantianism is the "murderer at the door" scenario, where the morally correct response is to tell the truth, even though lying would save the life of someone (Bennet 2015, p. 81). In this case, respecting the autonomy of the murderer by telling the truth appears contradictory in nature, as it disregards the potential harm to another individual. This highlights a limitation of Kant's ethical system, as it struggles to navigate complex moral situations where duties may clash.

3 The Applicability of Kantian Ethics to Automated Ethics

An opportunity of the application of Kantian ethics to automated ethics is its potential for ethical consistency and fairness. Singh (2022, p. 16) argues that "Kantian ethics is more natural to formalise" compared to other ethical theories, as "the Formula of Universal Law evaluates the form and structure of an agent's maxim" and requires less knowledge about the "state of affairs" or "moral character". This argument is further solidified through their implementation of an AMA that can successfully evaluate certain ethical scenarios, like the nature of joking and lying (Singh 2022, pp. 6–7). Since Kantian ethics is a rule-based system which focuses on upholding clear and universal rules, it has been shown to be compatable with such automated systems by being computationally tractable. As a result of this, such automated systems are able to remain consistent when following a set of predefined rules perscribed by Kantian ethics.

A risk of the application of Kantian ethics to automated ethics is that artificial moral agents lack genuine autonomy or consciousness. Kant states that transcendental freedom is "fundamental requirement of morality" (Manna and Nanth 2021, p. 142). That is in order to be considered a moral agent, a rational being should have the capability of acting autonomously rather than being controlled from external influences. Manna and Nanth (2021, p. 149) argue that "AI Systems are deterministic models of agency that do not exceed its initial programming". Since AMAs are only able to act within the bounds of their programming, AMAs do not possess any consciousness or autonomy. The actions of AMAs are mechanically driven, rather than driven by a sense of rationality.

The implications of this are that Kantian machines will be devoid of any moral intuition or empathy, which is crucial to respecting the dignity of a rational being.

Another risk that is present is that an over-reliance on artificial moral agents would have the potential to diminish human engagement with moral responsibility. Manna and Nanth (2021, p. 146) claim that "AI's moral deeds are not generated from the 'freedom of will' and the sense of 'duty' itself", but are simply a result of the programmer's instructions. This raises concerns about who is truly responsible for the outcomes of ethical decisions that are made by AMAs. Despite being programmed to follow ethical guidelines, the actions behind them lack moral reasoning, as they are incapable of understanding the principles behind the decisions they make. Thus, the nature of who takes responsibility when an ethical failure occurs becomes ambiguous.

Manna and Nanth (2021, p. 148) further argue that an AI agent "works according to hypothetical rules" rather than the categorical imperative. This means they do not act out of a sense of duty, but instead follow a set of conditional rules established by humans. This raises the issue that AMAs are incapable of comprehending the universal principles which govern human responsibility. Thus, an over-reliance on these systems will risk diminishing human engagement with moral accountability. This could lead individuals to trust these automated agents to make such moral decisions, which absolves an individual of their responsibility to reflect on the ethical implications of their own actions.

4 Conclusion

Kantian ethics provides a powerful framework for moral reasoning, placing emphasis on duty, autonomy and adhering to universal principles, especially in its application to automated ethics. However, its limitations become apparent in complex moral dilemmas and in the context of automated ethics, where artifical moral agents lack true autonomy and moral intuition. Despite these challenges, Kantian ethics remains a valuable tool for ensuring ethical consistency, but care must be taken to preserve human responsibility in moral decision-making.

References

Bennet, C. (2015). *What is this thing called ethics*. second edition. Routledge.

Manna, R. and Nanth, R. (2021). "Kantian Moral Agency and the Ethics of Artificial Intelligence". *Problemos* 100, pp. 139–151.

Singh, L (2022). *Automated Kantian Ethics: A Faithful Implementation*. URL: `https://github.com/lsingh123/automatedkantianethics`. (accessed: 15.09.2024).

Tonkens, R. (2009). “A Challenge for Machine Ethics”. *Minds & Machines* 19, pp. 421–438.