# COMP4920 Essay 1

Adrian Balbalosa, z5397730

September 2024

## 1 Introduction

The notion of Kantian ethics to treat all rational beings as equal is a strength of this ethical framework. However, the flaws of Kantainism begin to show when we begin to justify more complex ethical dilemmas. The use of Kantian ethics to design an automated ethics, particularly the creation of Artificial Moral Agents (AMAs) has provided interesting implications. Kantian ethics use as a foundation for an Artificial Moral Agent can help to serve as a guide for human decision-making, since the laws which govern this theory are shown to be computationally tractable. However, there are still risks to consider when using Kantian ethics to design an AMA. The first being that AMAs do not have genuine autonomy nor a consciousness, which erodes the genuineness of moral decisions. The second is that an over-reliance on AMAs will undermine the development of moral character and accountability for humans.

## 2 An Assessment of Kantian Ethics

Kantian Ethics is a deontological ethical theory which places emphasis on duty and moral principles over consequences. Central to this framework is the notion of the Categorical Imperative, which states that one should act according to the maxims which can be universally applied. Another formulation of this is that individuals should treat others as an ends, not just a means to an end. Kantian Ethics prioritises rationality and personal freedoms, and argues that ethical actions should arise from a sense of duty and adhering to moral law, rather than from emotional or situational considerations.

A notable strength of Kantian Ethics is that it places emphasis on respect for the individual. A fundamental aspect of Kantian Ethics is that humans should be respected because we are rational agent because we have the capacity for rational behaviour, and can be free from our impulses (Bennet 2015, p. 77).

This underpins the notion that humans ought to never be treated as means to our own devices, because we are rational beings (Bennet 2015, p. 77). By ensuring that people are treated as an ends, it upholds human dignity and rights.

However, there is a fundamental flaw of Kantian Ethics, in that when duties conflict, it is not clear what action we are to take and how to resolve those dilemmas. A classical example used by critics of Kantianism is the murderer at the door scenario, where the correct response is to respect the autonomy of the murderer and tell the truth (Bennet 2015, p. 81). Lying could potentially save the life of someone, but we cannot lie as we would be disregarding the autonomy of the murderer, which is paradoxical in nature. As a result of this, we are left to deliberate with difficult ethical decisions in a complex situation like this.

# 3 The Applicability of Kantian Ethics to Automated Ethics

An opportunity which Kantian ethics presents is that it has been shown to be computationally tractable. 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). It is noted that there are still limitations to this implementation because the inputs and outputs require a specific structure (Singh 2022, p. 10). Despite this limitation, we can observe that the use of Kantian ethics makes the implementation of AMAs feasible.

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. This could potentially lead to the dehumanisation of the moral decision-making process, since AMAs lack the capacity for free and rational choice, a fundamental Kantian value.

Another risk that is present is that an over-reliance on artificial moral agents would have the potential to diminish human engagement with moral responsi-

bility. 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 generated by the programmer's command. The implications of this are even when these machines are programmed to follow ethical guidelines, they perform them without actually understanding the moral reasoning behind them. Thus, the actions of AMAs are considered to lack moral worth and are considered amoral. Through relying on AI to make our moral decisions, we risk reducing moral actions to a mechanical process which lack the moral depth of deliberation. Manna and Nanth (2021, p. 148) also state that an AI agent "works according to hypothetical rules" rather than follow the categorical imperative, since their actions are not performed out of a sense of duty, which is a central part of the categorical imperative. Hence, these machines follows rules that are conditional rather than universal. The implications of this are that AI machines are unable to understand the universal nature of the categorical imperative. Thus, with over-reliance on these systems, we may become accustomed to following the rules without engaging with the deeper moral reasons that are essential for moral action.

# 4 Conclusion

While Katian ethics has shown to be an ethical theory which upholds human dignity and equality, it fails in more complex situations where maxims may clash against each other. The application of Kantian ethics to automated ethics presents some interesting opportunities. One of them being that Kantian ethics is generally simpler to automate, as the categorical imperative provides an algorithmic process for making ethical judgements. We must however consider that AMAs lack the genuineness of a rational being as a result of a lack of autonomy, and that we must be careful to not over-rely on them for guidance.

# 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).