COMP4920 Essay 1

Adrian Balbalosa, z5397730

September 2024

1 Introduction

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.

One of the strongest parts 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

In this section I argue that Kantian Ethics is not an appropriate framework for automated ethics, specifically in developing artificial moral agents (AMAs) that can make ethical judgements. The reason being that creating these AMAs goes against the ethos of Kantian ethics itself.

Artificial Moral Agents that are built on top of Kantian Ethics are not considered to have any autonomy. To be considered a moral agent, one must have freedom of choice (Manna and Nanth 2021, p. 141). Artificial Moral Agents are not considered to be free because they are programmed to act in a certain way (Tonkens 2009, p. 429). They are not free enough to operate outside of the boundaries of which they are programmed, meaning they are not able to exhibit any form freedom that is characteristic of moral agents of Kantianism (Manna and Nanth 2021, p. 149). As AMAs are not able to exhibit any sort of free will, they cannot be considered free enough to justify moral actions.

Furthermore, the creation of Kantian AMAs goes against the categorical imperative. The categorical imperative requires actions are performed as universally necessary, but AMAs only follow a hypothetical imperative (Manna and Nanth 2021, p. 149). This is because they are programmed to work in a particular way in order to accomplish some goal, contrasting the nature of the categorical imperative. If we were to establish an Artificial Moral Agent as a rational being, we would be treating them as a means to an end, which violates the notion of the categorical imperative (Tonkens 2009, p. 432). Thus, the creation of these AMAs ends up contradicting the notions that were established by Kantian ethics.

While the development of these AMAs seem ethically dubious, Kantian AMAs have shown potential in getting us a step closer to creating machines which can reason about ethical dilemmas. Kantian ethics has proved itself ot be a useful ethical framework as a foundation for building an AMA (Singh 2022, p. 3). Kantian ethics has shown to be easier to embed in AMAs as it requires little information about the world compared to other ethical frameworks (Singh 2022, p. 17).

4 Conclusion

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.