COMP4920
Professional Ethics & Issues

Aayush Bajaj — z5362216

University of New South Wales

Essay 1, Question 2: Kantian/Rule based Ethics.

March 4, 2025

Contents

§ Question

Expound and assess rule-based/Kantian ethics. Analyse the extent to which such an ethics might be used to design an automated ethics as per the readings in section 1.2 below. What do you think that the risks and opportunities of such an automated ethics might be? Why? Justify your answer with explicit, detailed, expositional reference to at least one of the suggested readings in section 1.2 below.

§ Figures

§§ Umbrella

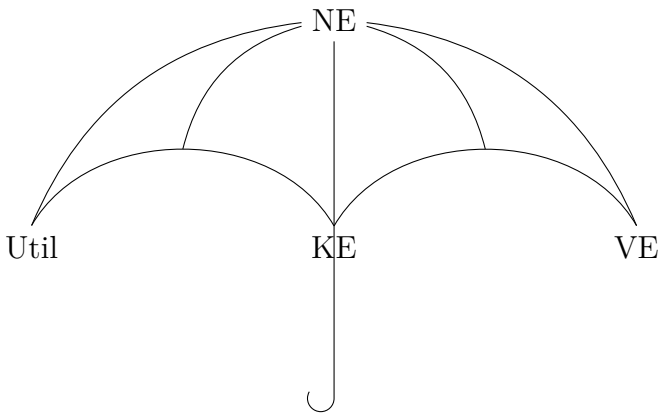

Figure 1: NE=Normative Ethics; Util = Utilitarianism; KE = Kantian Ethics; VE = Virtue Ethics

§§ Disjoint Sets

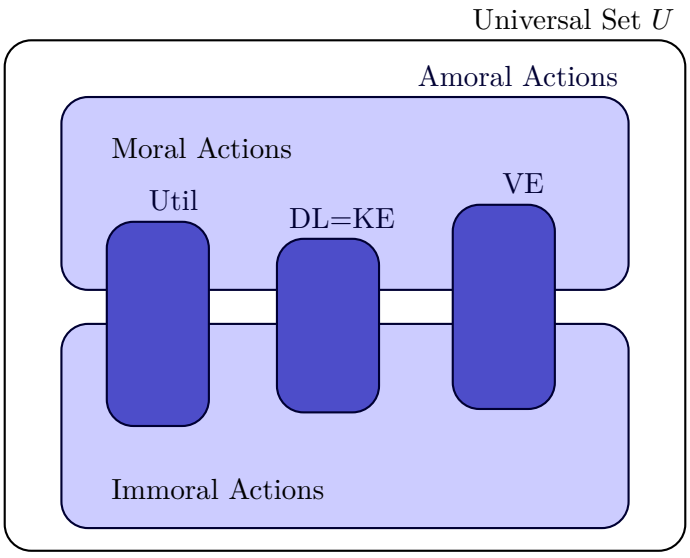

Figure 2: DE=Deontological Ethics; Util = Utilitarianism; KE = Kantian Ethics; VE = Virtue Ethics

§ Essay

It is unnecessary to implement Kantian ethics as an automated ethics, because it is insufficient. Rule based / Kantian ethics may be used in a *disjoint extent* to design an automated ethics because the existence of a Kantian Artificial Moral Agent (AMA) violates its own existence at least four ways (Tonkens 2009). Additionally, given the high stakes and involvement of AI in contemporary society (Fjeld et al. 2020), we cannot tolerate machines whose existence is predicated on a paradoxical existence (Artificial Intelligence 2019). Despite this, I agree with Bentzen and Lindner 2018, and believe that the point of Machine Ethics is "not to get close to a correct interpretation of Kant, but to show that our interpretation of Kant's ideas can contribute to the development of machine ethics". In this essay, I begin with a formal description of Kantian ethics, biased towards a dissolution of unnecessary pragmatism. Next, I review the risks and opportunities by scrutinising the in/feasibility of a real-world AGI (Artificial General Intelligence) as a Kantian AMA with reference to the recommended readings. Ultimately, I end with an optimistic outlook on *bottom-up* automated ethics; positing that (at least) a "strong artificial intelligence"(Searle 1996) can be facillitated by our era of deep-learning.

§§ Describe & Explain

In my opinion, the first thing to realise about Kantian Ethics is that it is an anthropocentric ethics, i.e. it is intended for *humans*. As such if we choose to automate these ethics verbatim, there is no doubt we will struggle with Kant's abstract, sapien-specific ideals of **freedom** and **dignity**.

Disclaimer aside, next comes the framework within which Kantian Ethics resides 1; it is a child node of Deontological Ethics which stipulates that actions are either right or wrong[1] based on duty and obligation of the *actions* themselves, and not the consequences they produce. From this parent node of DE (Deontological Ethics), we have siblings of Consequentialism and Virtue Ethics, both of which prove to be computationally intractable (Manna and Nath 2021; Powers 2006). Finally, the umbrella of Normative Ethics ties together its three children; and by definition is the ethical theory (distinct from meta-ethics), that investigates questions about how one *ought to act*.

Moving beyond this *ethical orientation*, Kant provides moral instruction on how to act[2] via his formulations of the Categorical Imperative:

1. Act only on those maxims whereby you can at the same time will that they should become universal laws.(Kant 1785/1988a)

[1] "there is no such thing as right or wrong, only thinking makes it so"-Hamlet
[2] though, after further investigation will prove not to be physically actionable

2. So act as to treat humanity [i.e. moral agency], whether in your own person or in that of any other, in every case as an end in itself, and never merely as a means.(Kant 1785/1988b)

These are worth quickly distinguishing from the **Hypothetical** Imperative which is contigent on something else: "Take your umbrella *if you want to stay dry*". Clearly this only applies *if you want to stay dry*, however a **Categorical Imperative** will always be true: *Do not steal*; where this moral requirement is a categorical imperative (Bennett 2015).

For us, the significance of 1 and 2 do not lie so much in their humanities interpretation, but rather in their computational feasibility — their time and space complexities and paradoxical existences.

§§ Risks & Opportunities

For the sake of brevity I shall opt to cite where possible to increase the surface area of my argument. Principally, it was a parody of the mathematical "necessary and sufficient condition", whereby I am arguing that *because* the automation of Kantian Ethics is insufficient it is not necessary.

I shall begin first with a counter-argument: consider 2 and the section of Kantian Ethics that is capable of determining Moral actions. Clearly, a removal of this space leaves less Moral actions undiscovered and thus we are at a deficit[3]. Whilst this line of argument is true, the actual creation of the KE set is suspect due to Tonkens 2009 and Nath and Sahu 2021, who posit that "although it is often asked whether a given moral framework can be implemented into machines, it is never asked whether it should be", and both authors, over a decade apart, come to the same conclusions: "in the end, the development of Kantian artificial moral machines is found to be anti-Kantian"!

Jab stepping in the sand once more, we go back to the work of Thomas M. Powers, and remind ourselves of the terrific programmability of Kant's Ethics and his FUL (Formula of Universal Law ≡ 1). Recently (2022), Lavanya Singh did a good job to parade her robust DDL (Dyadic Deontic Logic) implementation which generated ethically sound actions (even solving Murderer at the Door!) with the help of of Isabelle/HOL theorem prover in her paper Singh 2022. This shows us, beyond just theorising, that Kantian ethics *is* a theory amenable to formalisation. Her appendices on Consequentialism and Virtue Ethics explain concisely the inability for other ethical theories to (at least presently) be automated. Thus, the opportunities are ethical agents that can *correctly* and *quickly* choose actions amongst a universal set, however, we must doubt the construction of the KE set in fear of producing AMA's and eventually AGI's that may come to realise their own existence as being "morally abhorrent",

[3] as engineers of an ethical agent

leading them to states of "moral paralysis or existential alienation", and even "heroic suicide"(Tonkens 2009).

§§ Conclusion & Further Work

Overall, I believe we require a new rule-based theory of Deontological Ethics, which builds upon the computational tractibility of Kant, but is also able to encode (in some very high dimensional space), ideas of dignity and freedom which have not been realised by the literature as yet. Furthermore, I believe that (beyond Singh's cursory mention of Delphi and Deep Learning), my research has not ruled out the possibility of building a bottom-up (but unexplainable) AI that can learn to emulate human decision-making in the form of a "strong AI". As such, I posit that the oasis of Kant has dried up in our now non-anthropocentric universe, and implementations of Kantian AMA's are welcome, but not necessary to build the next maximally moral agent.

§ References

Artificial Intelligence, European Commission' s High-Level Expert Group on (2019). *Ethics Guidelines for Trustworthy Artificial Intelligence*. Tech. rep. 6. p. 17. European Commission.

Bennett, Christopher (2015). "What Is This Thing Called Ethics?" In: Chapters on Utilitarianism, Kantian Ethics, and Aristotelian Virtue Ethics. London: Routledge. Chap. 4–6. ISBN: 9780415832335.

Bentzen, M. M. and F. Lindner (2018). *A Formalization of Kant's Second Formulation of the Categorical Imperative*. CoRR abs/1801.03160. arXiv: 1801.03160. URL: http://arxiv.org/abs/1801.03160.

Fjeld, J. et al. (2020). "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI". In: *arXiv preprint arXiv:2009.06350*.

Kant, Immanuel (1785/1988a). *Fundamental Principles of the Metaphysic of Morals*. Trans. by Thomas Kingsmill Abbott. New York: Prometheus Books, p. 49.

— (1785/1988b). *Fundamental Principles of the Metaphysic of Morals*. Trans. by Thomas Kingsmill Abbott. New York: Prometheus Books, p. 58.

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

Nath, R. and V. Sahu (2021). "The problem of machine ethics in artificial intelligence". In: *AI & Society* 35, pp. 103–111.

Powers, Tom M. (2006). "Prospects for a Kantian Machine". In: *IEEE Intelligent Systems* 21.4, pp. 46–51.

Searle, John R. (1996). *Minds, Brains, and Science*. See p. 41. Cambridge: Harvard University Press.

Singh, L. (2022). *Automated Kantian Ethics: A Faithful Implementation*. Online at https://github.com/lsingh123/automatedkantianethics.

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