1 Introduction

In this section, I justify my choice to automate Kantian ethics, as opposed to virtue ethics or consequentialism. Kantian ethics is easy to formalize for three reasons. First, evaluating a maxim requires little data about the world. Second, a maxim is relatively easy to represent to a computer. Third, while any ethical theory has debates that automated ethics would need to take a stance on, these debates are less frequent and less controversial for Kantian ethics.

The argument is not that no form of consequentialism or virtue ethics is tractable to automate, but rather that Kantian ethics is easier. I do not present a comprehensive comparision of ethical theories, but instead sketch the basic principles that most consequentialist and virtue ethical theories share.

2 Kantian Ethics

2.1 Kant Crash Course

First, I explain the concepts of practical reason, the will, and maxims. I then present Kant's argument that because the will is autonomous, only the will has authority over itself [Korsgaard and O'Neill, 1996]. Finally, I argue that a will is definitionally only bound by those imperatives that are implied by practical reason itself [Velleman, 2005]. From there, I present the three formulations of the categorical imperative, focusing on the Formula of Universal Law (FUL) [Kant, 1785]. This is not a full defense of Kantian ethics but is instead a quick sketch of one argument for the FUL.

2.2 Kant is Easy to Formalize

The FUL is easy to formalize because it is a purely formal principle that is only evaluates the form of the maxim that an agent acts on. No other information is relevant to the test, so moral judgement can proceed with a relatively small amount of data.

2.3 Common Debates

I acknowledge the debates[1] in Kantian ethics that my project takes a stance on: the correct way to interpret the FUL and the definition of a maxim. The stances I take are generally accepted by most Kantians, though some still disagree.

3 Consequentialism

3.1 Consequentialism Crash Course

I define a consequentialist theory as one that evaluates the consequences of an action, acknowledging that this definition itself is controversial [Sinnott-Armstrong, 2021]. I then present the debates over theories of the good, which consequences to evaluate, and the aggregation of consequences.

3.2 Consequentialism is Hard to Formalize

3.2.1 Requires Lots of Data About States of Affairs

Making a consequentialist moral judgement requires data about the entire state of affairs following an action, posing many challenges. First, collecting this data is difficult. Second, in order to trust the

[1] I have a separate piece of writing that provides a literature review of these debates and justifies my stances. Would that fit here or somewhere else?

system's judgements, we have to trust the ethical theory, its theory of the good, and the many judgements required to assign each state of affairs a goodness measurement [Driver, 2014].

3.2.2 Tradeoff Between Aggregation vs Wholistic Evaluation

Consequentialism also faces the further problem of aggregating goodness across people. Consequentialists who abandon aggregation must instead find some wholistic evaluation function for a state of affairs. There is a tradeoff between the difficulty of aggregation and the complexity of making judgements about an entire state of affairs, as opposed to about a single person. A reasoner will need large, complicated datasets to settle these questions.

4 Virtue Ethics

4.1 Virtue Ethics Crash Course

I understand the concept of virtue as those traits that are good for the posessor [Hursthouse and Pettigrove, 2018]. I briefly explain Aristotle's eudaimonistic conception of virtue and present some examples of virtues (courage, temperance, equanimity).

4.2 Virtue Ethics is Hard to Formalize

4.2.1 What is Virtue?

Automated virtue ethics will need to plant a flag in the messy, controversial debate over the exact list of virtues. While most Kantians agree on one interpretation of the FUL, most virtue ethicists have their own interpretations of what the virtues are.

4.2.2 Representing Moral Character is Difficult

Automated virtue ethics has to evaluate moral character, which is much more challenging than evaluating a maxim. Moral character is a complex concept to precisely represent to a computer.

4.2.3 Machine Learning and Virtue Ethics

There is a connection between the ideas of cultivating habit and mimicking virtuous action and machine learning, which mimics patterns in datasets[2]. There's been some work using machine learning to learn moral behavior from a dataset of actions tagged as ethical [Jiang et al., 2021]. One drawback of this approach is the fact that machine learning algorithms have trouble explaining why they made the judgements they made and often pick up on patterns that human beings would not see as indicative of causation. In contrast, my system can explain exactly which axioms and principles resulted in a maxim being obligated or prohibited and can even present human-reconstructable proofs of its results.

References

[Driver, 2014] Driver, J. (2014). The History of Utilitarianism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2014 edition.

[2] Would love feedback on whether this should go here or in a Related Work section.

[Hursthouse and Pettigrove, 2018] Hursthouse, R. and Pettigrove, G. (2018). Virtue Ethics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2018 edition.

[Jiang et al., 2021] Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., and Choi, Y. (2021). Delphi: Towards machine ethics and norms.

[Kant, 1785] Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Cambridge University Press, Cambridge.

[Korsgaard and O'Neill, 1996] Korsgaard, C. M. and O'Neill, O. (1996). *The Sources of Normativity*. Cambridge University Press.

[Sinnott-Armstrong, 2021] Sinnott-Armstrong, W. (2021). Consequentialism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition.

[Velleman, 2005] Velleman, J. D. (2005). *A Brief Introduction to Kantian Ethics*, page 16–44. Cambridge University Press.