1 Introduction

In this section, I hope to explain why I choose to automate Kantian ethics, as opposed to another ethical theory. My general approach is to argue that Kantian ethics is easier and more natural to formalize and automate than other theories. Given that my thesis is an early-stage, proof-of-concept for computational ethics, formalizing the most formal ethical theory makes sense. I don't have time or space to evaluate every ethical theory, but I hope to sketch the spirit of consequentialist and virtue ethical theories and argue why they would be more difficult to formalize than Kantian ethics.

2 Kantian Ethics

2.1 Kant Crash Course

I will present a brief introduction to Kantian ethics. First, I will explain the concepts of practical reason, the will, and maxims. I will then present Kant's argument that the only the will has authority over itself. Finally, I will argue that a will is, definitionally, only bound by those imperatives that are implied by practical reason itself. From there, I will present the three formulations of the categorical imperative, focusing on the Formula of Universal Law. My goal is not to present a full defense of Kantian ethics, but instead to quickly sketch the argument in a way that conveys the concept of a maxim and of practical reason.

2.2 Kant is Easier to Formalize

Armed with this understanding of Kantian ethics, I will argue that the FUL is easy to formalize because it is a purely formal principle that is only concerned with the form of a maxim. Essentially, evaluating action under the FUL merely requires formulating the maxim of that action. No other information beyond the form of the maxim is relevant to the test, so moral judgement can proceed with a very small amount of data about the action.

2.3 Flags Planted

Here, I will acknowledge that Kantian ethics has many debates and my project necessarily takes a stance on some of them. Specifically, I take a stance on the correct way to interpret the FUL and the definition of a maxim. I will note that the stances I take are generally accepted by most Kantians, and thus do not open my project to huge philosophical critcism, though some will still disagree with my choices.

3 Consequentialism

3.1 Consequentialism Crash Course

I will present a brief overview of different consequentialist theories. I will crudely define a consequentialist theory as one that evaluates the consequences of an action, acknowledging that this definition itself is controversial. I will then present debates over what qualifies as a good consequence, which consequences to evaluates, and the aggregation of consequences across people.

3.2 Consequentialism is Hard to Formalize

3.2.1 Data About States of Affairs

Using debates about consequentialist theories of the good as a backdrop, I will argue that making a moral judgement in the consequentialist context requires data about the entire state of affairs following an action. Kantian ethics, on the other hand, merely requires the form of the maxim itself. This poses many challenges. First, collecting this data is difficult. Second, in order to trust the system's judgements, we have to trust its theory of the good, but this is a point of contention. Third, a critic could not only question the system's theory of the good, they could also question the huge number of judgements that will go into assigning each state of affairs a goodness measurement.

Here I will note that Kantian ethics also needs to take stances on debates about interpretations of the theory (and will point to one such debate about the meaning of the FUL).The larger point is that every ethical theory needs to "plant such flags," but Kantian ethics plants fewer and less controversial flags.

3.2.2 Aggregation vs Wholistic Evaluation

Consequentialism also faces the further problem of aggregating goodness across people. On the other hand, consequentialists who abandon aggregation must instead find some wholistic evaluation function for a state of affairs. Each approach poses challenges, with 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. Again, a reasoner will need to plant flags in these debates and will need large, complicated datasets to settle these questions.

4 Virtue Ethics

4.1 Virtue Ethics Crash Course

I will focus my exposition of virtue ethics on the concept of virtue as those traits that are good for the posessor. I will briefly explain Aristotle's eudaimonistic con-

ception of virtue and present some examples of virtues (courage, temperance, equanimity).

4.1.1 What is Virtue?

I will present a common debate in virtue ethics over the exact list of virtues. I will argue that automated virtue ethics will need to plant a flag in this messy, controversial debate. While most Kantians agree on one interpretation of the FUL, most virtue ethicists have their own unique interpretation of what the virtues are.

4.1.2 How Do We Represent Moral Character?

Lastly, I will argue that automated virtue ethics has to evaluate moral character, which is much more challenging than evaluating maxims as in Kantian ethics. Moral character is a complex concept that human beings don't really understand how to represent to ourselves, let alone to make precise to a computer.

4.1.3 Machine Learning and Virtue Ethics?

Would love feedback on whether this fits here or in a separate, related work section. There's a proposed connection between the ideas of cultivating habit, moral education, and mimicking virtuous action and the operation of machine learning, which learns and mimics patterns in datasets. There's been some work using machine learning to learn moral behavior from a dataset of actions tagged as ethical or not. I want to present this as a valid alternative to my approach, with its own set of pros and cons. One con I want to focus on is explainability, or the idea 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 significant or indicative of causation or any meaningful property of a dataset. 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.

end