Philosophical Writing

Lavanya Singh

October 28, 2021

Contents

1 Choice to Formalize the FUL

In *Groundwork of the Metaphysics of Morals*, Kant presents three formulations, or versions, of what he calls the "supreme law of morality." I will focus on the first of these three formulations, and below I explain the formulations and defend my choice.

Kant argues that if morality exists, it must take the form of a categorical imperative or a law that holds unconditionally. Categorical imperatives are contrasted with hypothetical imperatives, which take the form of conditionals as in, "If I want to get good grades, I must study hard." Hypothetical imperatives only have force so long as the antecedent holds, but the categorical imperative is unconditionally binding [Kant, 1785, 28]. In the first half of *Groundwork*, Kant examines what the categorical imperative, if such a thing exists and has force, must be. He concludes that there are three "formulations" of the categorical imperative, or three ways of articulating the supreme law of morality.

The first formulation of the categorical imperative is the formula of universal law (FUL), which reads, "act only according to that maxim through which you can at the same time will that it become a universal law." [Kant, 1785, 34] This formulation generates the universalizability test, which "tests" the moral value of a maxim by imagining a world in which it becomes a universal law and attempting to will the maxim in that world. The second formulation of the categorical imperative is the formula of humanity (FUH): "So act that you use humanity, in your own person, as well as in the person of any other, always at the same time as an end, never merely as a means." [Kant, 1785, 41]. This formulation is often understood as requiring us to acknowledge and respect the dignity of every other person. The third formulation of the categorical imperative is the formula of autonomy (FOA), which Korsgaard summarizes in her introduction to the Groundwork as, "we should so act

that we may think of ourselves as legislating universal laws through our maxims." [Korsgaard, 2012, 28] While closely related to the FUL, the FOA presents morality as the activity of perfectly rational agents in an ideal "kingdom of ends," guided by what Kant calls the "laws of freedom."

I choose to focus on formalizations of Kant's first formulation of the categorical imperative, the formula of universal law (FUL), because it is the most formal and thus the easiest to formalize and implement. Onora O'Neill explains that the formalism of the FUL allows for greater precision in philosophical arguments analyzing its implications and power [O'Neill, 2013, 33]. This precision is particularly useful in a computational context because any formalism necessarily makes its content precise. The FUL's existing precision reduces ambiguity, allowing me to remain faithful to Kant's writing and philosophical interpretations of it. Precision reduces the need to make choices to resolve debates and ambiguities. Some of these choices may be well-studied and grounded in literature, but some may be unique to formalizing the FUL and thus understudied. Minimizing these choices minimizes arbitrariness in my formalization and puts it on solid philosophical footing. Given that this thesis is a proof-of-concept, the formalism of the FUL is attractive because it reduces both the computational and philosophical complexity of my work.

While some criticize the FUL for its formalism and percieved "sterility" [O'Neill, 2013, 33], Kantian constructivists embrace it [Ebels-Duggan, 2012, 173]. My project is not committed to Kantian constructivism. I believe that computational ethics is likely a valuable tool for any ethicist, and I make the case for Kantian ethics specifically. Nonetheless, Kantian constructivists may find the focus on the FUL particularly appealing.

Though Kantians study all formulations of the categorical imperative, Kant argues in Groundwork that the three formulations of the categorical imperative are equivalent [Kant, 1785]. While this argument is disputed [Johnson and Cureton, 2021],

for those who believe it, the stakes for my choice of the FUL are greatly reduced. If all formulations are equivalent, then a formalization of the FUL lends the exact same power as a formalization of the second or third formulation of the categorical imperative. In fact, future work could formalize the other formulas and try to prove that they are identical. Kant believes that his argument for the equality of the formulas is analytical, and if he is correct, it should be possible to recreate the argument in logic.

2 Definition of a Maxim

The central unit of evaluation for the universalizability test is a "maxim," which Kant defines in a footnote in *Groundworkd* as "the subjective principle of willing," or the principle that the agent acts on [Kant, 1785, 16]. Modern Kantians differ in their interpretations of this definition. The naive view is that a maxim is an act, but Korsgaard adopts the more sophisticated view that a maxim is composed of an act and the agent's purpose for acting [Korsgaard, 2005]. She also compares a maxim to Aristotle's logos, which includes these components and information about the circumstances and methods of the act. O'Neill concludes that Kant's examples imply that a maxim must also include circumstances [O'Neill, 2013], and Kitcher [Kitcher, 2003] uses textual evidence from the Groundwork to argue for the inclusion of a maxim's purpose or motivation. In order to formalize the notion of a maxim, I must adopt a specific definition and defend my choice.

I define a maxim as a circumstance, act, goal tuple (C, A, G), read as "In circumstances C, act A for goal G." Isabelle's strict typing rules mean that the choice of the type of each member of this tuple is significant. A circumstance is represented as a set of worlds t where that circumstance holds. A goal is also a term because it can be true or false at a world if it is realized or not. An act is an open sentence

because an act itself is not the kind of thing that can be true or false (as in, an act is not truth-apt), but the combination of a subject performing an act can be true or false at a world depending on whether or not the act is indeed performed by that subject. For example, "running" is not truth-apt, but "Sara runs" is truth-apt.

My definition of a maxim is inspired by O'Neill's work on maxims. I will defend my representation below and consider an additional component that Kitcher argues for.

2.1 O'Niell's Original Schematic and The Role of Practical Judgement

O'Neill [O'Neill, 2013, 37] presents what Kitcher [Kitcher, 2003] calls the widely accepted view that a maxim is a circumstance, act, goal tuple. A maxim is an action-guiding rule and thus naturally includes an act and the circumstances under which it should be performed, which are often referred to as "morally relevant circumstances."

She also includes a purpose, end, or goal in the maxim because Kant includes this in many of his example maxims and because Kant argues that human activity, because it is guided by a rational will, is inherently purposive [Kant, 1785, $4:428$]. A rational will does not act randomly (else it would not be rational), but instead in the pursuit of ends which it deems valuable. This inclusion is also essential for the version of the universalizability test that I will implement, explained in Section ??.

O'Neill's inclusion of circumstances is potentially controversial because it leaves open the question of what qualifies as a relevant circumstance for a particular maxim. This is gives rise to "the tailoring objection" [Kitcher, 2003, 217][1], under which maxims are arbitrarily specified to pass the FUL. For example, the maxim "When my name is Lavanya Singh, I will lie to get some easy money," is universal-

[1] Kitcher cites [Wood, 1999] as offering an example of a false positive due to this objection.

izable, but is clearly a false positive. One solution to this problem is to argue that the circumstance "When my name is Lavanya Singh" is not morally relevant to the act and goal. This solution requires some discussion of what qualifies as a relevant circumstance.

O'Neill seems to acknowledge the difficulty of determining relevant circumstances when she concedes that a maxim cannot include all of the infinitely many circumstances in which the agent may perform the action[O'Neill, 2013, 4 : 428]. She argues that this is an artifact of the fact that maxims are rules of practical reason, the kind of reason that helps us decide what to do and how to do it [Bok, 1998]. Like any practical rule, maxims require the exercise of practical judgement to determine in which circumstances they should be applied. This judgement, applied in both choosing when to exercise the maxim and in the formulation of the maxim itself, is what determines the "morally relevant circumstances."

The upshot for computational ethics is that the computer cannot perform all ethical activity alone. Human judgement and the exercise of practical reason are essential to both formulate maxims and determine when the actual conditions of life coincide with the circumstances in which the maxim is relevant. Choosing when to exercise a maxim is less relevant to my project because analyzing a formal representation of the FUL requires making the circumstances in a given scenario precise, but will be important for applications of computational ethics to guiding AI agents. The difficulty in formulating a maxim, on the other hand, demonstrates the important fact that ethics, as presented here, is not a solely computational activity. A human being must create a representation for the dilemma they wish to test, effectively translating a complex, real situation into a flat logical structure. This parallels the challenge that programmers face when translating the complexity of reality to a programming langauge or computational representation. Not only will some of the situation's complexity inevitably be lost, the outcome of the universalizability

test will depend on how the human formulates the maxim and whether or not this formulation does indeed include morally relevant circumstances. If the human puts garbage into the test, the test will return garbage out.

While this may appear to be a weakness of my system, I believe that it actually allows my system to retain some of the human complexity that many philosophers agree cannot be automated away.[2] Ethics is a fundamentally human activity. Kant argues that the categorical imperative is a statement about the properties of rational wills. In fact, Korsgaard argues that morality derives its authority over us, or normativity, only because is it a property of a rational will, and we, as human beings, are rational wills. If ethics is meant to guide human behavior, the role of the computer becomes clear as not a replacement for our will, but instead as a tool to help guide our wills and reason more efficiently and more effectively. Just as calculators don't render mathematicians obsolete, computational ethics does not render human judgement or philosophy obsolete. Chapter 4 Section ?? will be devoted to a more complete discussion of this issue.

2.2 Exclusion of Motive

Kitcher begins with O'Niell's circumstance, act, goal view and expands it to include the motive behind performing the maxim [Kitcher, 2003]. This additional component is read as "In circumstance C, I will do A in order to G because of M," where M may be "duty" or "self-love." Kitcher argues that the inclusion of motive is necessary for the fullest, most general form of a maxim in order to capture Kant's idea that an action derives its moral worth from being done for the sake of duty itself. Under this view, the FUL would obligate maxims of the form "In circumstance C, I will do A in order to G because I can will that I and everyone else

[2] Powers presents the determination of morally relevant circumstances as an obstacle to the automation of Kantian ethics [Powers, 2006].

simultaneously will do A in order to G in circumstance C." In other words, if Kant is correct in arguing that moral actions must be done from the motive of duty, the affirmative result of the FUL becomes the motive for a moral action.

While Kitcher's conception of a maxim captures Kant's idea of acting for duty's own sake, I will not implement it because it is not necessary for putting maxims through the FUL. Indeed, Kitcher acknowledges that O'Neill's formulation suffices for the universalizability test, but is not the general notion of a maxim. In order to pass the maxim through the FUL, it suffices to know the circumstance, act, and goal. The FUL derives the motive that Kitcher bundles into the maxim, so automating the FUL does not require including a motive. The "input" to the FUL is the circumstance, act, goal tuple. My project takes this input and returns the motivation that the dutiful, moral agent would adopt. Additionally, doing justice to the rich notion of motive requires modelling the operation of practical reason itself, which is outside the scope of this project. My work focuses on the universalizability test, but future work that models the process of practical reason may use my implementation of the FUL as a "library." Combined with a logic of practical reason, an implementation of the FUL can move from evaluating a maxim to evaluating an agent's behavior, since that's when "acting from duty" starts to matter.

3 Practical Contradiction Interpretation

Kantians debate the correct interpretation of the formula of universal law because Kant appears to interpret the universalizability test in different ways. My project uses Korsgaard's practical contradiction interpretation, broadly accepted as correct within the philosophical community [Ebels-Duggan, 2012, 177]. Below, I briefly reconstruct Korsgaard's argument for the practical contradiction interpretation. While

she believes that the text partially supports this interpretation, her argument is philosophical and derives its strength from the plausibility of the practical contradiction interpretation.

Recall that the formula of universal law is "act only in accordance with that maxim through which you can at the same time will that it become a universal law" [Kant, 1785, $4:421$]. To determine if a maxim can be willed as a universal law, one must use the "universalizability test," which requires imagining a world in which everyone for all of time has willed the maxim. If willing the maxim in such a world generates a contradiction, then the action is prohibited. There are three interpretations of what sort of contradiction is necessary: (1) the teleological view, prohibiting actions that conflict with some assumed teleological end when universalized, (2) the logical contradiction view, prohibiting maxims that are logically impossible when universalized, and (3) the practical contradiction view, prohibiting maxims that are self-defeating when universalized.

Under the logical contradiction interpretation, falsely promising to repay a loan to get some quick cash fails the universalizability test because, in such a world, the practice of promising would die out so making a false promise would be impossible. Korsgaard appeals to Dietrichson [Dietrichson, 1964] to construct the example of a mother killing her children that tend to cry more than average so that she can get some sleep at night. Universalizing this maxim does not generate a logical contradiction, but it is clearly morally wrong. The problem here is that killing is a natural action, which Korsgaard distinguishes from a practice, like promising. Natural actions will never be logically impossible, so the logical contradiction view fails to prohibit them.

Under the teleological contradiction interpretation, a maxim is prohibited if it undercuts some natural or assigned purpose for some practice, act, or object. For example, the purpose of promising is to create a system of mutual trust and false

promising undercuts this purpose and is thus prohibited. The problem with this view is that it assumes that the agent is committed, either because of their own goals or because of some property of a rational will, to some teleological system. Acton formulates Hegel's argument that [Ewing, 1972], an agent doesn't have to be committed to promising as a system of mutual trust. Korsgaard concludes that assigning teleological purposes to actions is difficult because "such purposes may have nothing to do with what the agent wants or ought rationally to want, or even with what any human being wants." If the agent is not committed to the purpose, then will not see a contradiction in willing an act that violates this purpose.

This difficulty with the teleological contradiction interpretation drives Korsgaard to look for purposes that an agent must necessarily be committed to, and she concludes that this must be the purpose of the maxim itself. By willing a maxim, an agent commits themselves to the goal of the maxim, and thus cannot rationally will a system in which this goal is undercut. This system satisfactorily handles natural actions like those of the sleep-deprived mother: in willing the end of sleeping through the night, she is implicitly willing that she be alive in order to secure and enjoy her sleep. If any mother is allowed to kill any loud child, then she cannot be secure in the possession of her life, because her own mother may have grown frustrated with her crying. Her willing this maxim thwarts the end that she sought to secure.

The practical contradiction interpretation not only addresses the problems with the first two interpretations, it also offers a much more satisfying explanation of why certain maxims are immoral. The problem is not the existence of a contradiction itself, but instead the fact that these maxims involve parasitic behavior on social conditions that the agent seeks to benefit from. The false promiser simultaneously wants to abuse the system of promising and benefit from it, and is thus making an exception of themselves. It is this kind of free-riding that the universalizability test

seeks to draw out. The test raises the same kinds of objections that the question "What if everyone did that?" seeks to draw out.

4 Philosophical Contributions

I argue that computational ethics should be useful for and interesting to philosophers for two reasons. First, it could serve as the basis for AI agents with the capacity for philosophically sophisticated ethical reasoning. For example, my project contributes an implementation of the Formula of Universal Law that an AI agent could use to reason about the world using the categorical imperative. Second, computational ethics helps philosophers think about ethics in the same way that theorem provers help mathematicians think about math. I am not arguing that the computer can replace human reasoning or prove things that humans theoretically couldn't do. Instead, I argue that the computer bolsters human reasoning by forcing precision due to the rigid syntax of a computer program. Below, I explore these contributions in greater detail.

4.1 AI Agents

As artifical intelligence becomes more powerful, science-fiction predictions about "evil AI" and calls from regulators are intensifying the need for "ethical AI". My project contributes a "top down" approach automating a particular ethical theory. My work on automating the categorical imperative could serve as one component of a partially or fully artificial ethical reasoner. Specifically, my project could be repurposed into a "categorical imperative library" that takes as input the logical representation of a maxim and determines its moral status (if it is obligatory, prohibited, or permissible).

As it stands, my project can evaluate the moral status of maxims represented in my logic and potentially serves as one component of an "ethics engine" that an AI agent could use to make ethical decisions. For example, my system could be combined with an input parser to translate moral dilemmas as represented to the

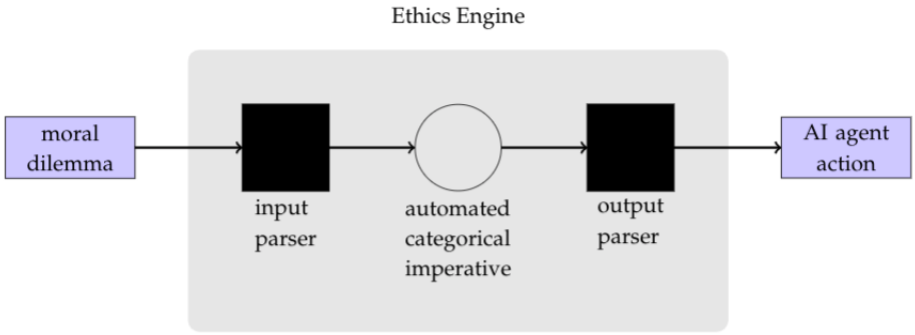

Figure 1: An example of an ethics engine for an artificial agent. I contribute the automated categorical imperative component.

AI agent into maxims in my logic. The ouput of my system could be fed into an output parser to translate this output into a prescription for the action the AI agent should take. Figure 1 depicts the workflow of this example ethics engine.

In this workflow, an AI agent is faced with a moral dilemma in some internal representation. This internal representation would need to be translated by an input parser into an appropriate logical representation, i.e. a circumstance, act, goal tuple. This input parser is the most technically and ethically challenging component of the system. It is this input parser that determines which circumstances are "morally relevant" for a maxim, a judgement that requires commonsense reasoning and knowledge about moral relevance. Translating everyday situations into appropriate maxims is the bulk of the work that a Kantian human being does when making decisions. Common misconceptions about Kantian ethics[3] often result from incorrectly formulated maxims, and the entire field of applied Kantian ethics

[3]For example, some wonder why the FUL doesn't prohibit gay sex, as the maxim "marry someone of the same sex" appears to result in human extinction when universalized. The solution to this dilemma is that the maxim a gay person acts on is usually something of the form, "marry the person you love because you love them," which is perfectly reasonable to universalize.

is devoted to generating the right kinds of maxims to test. For more discussion of challenges involved with defining morally relevant circumstances, see Section UNWRITTEN.

This representational question will be one of the biggest hurdles to actually using my categorical imperative library in an AI ethics engine. Currently, it may be reasonable for a human being to perform the role of the input parser. Once an AI agent stumbles onto an ethical dilemma, a human being could take over, formulate the right question, and feed it into the categorical imperative library to see what action the categorical imperative would prescribe. This may actually be a feature, not a bug. Proponents of the "human-in-the-loop" model argue that fully automated decision-making is doomed to ethical failure, and that the inclusion of a human being injects common-sense sanity into otherwise dangerous decisions[4].

It is likely that, regardless of the strengths of the human-in-the-loop model, fully automated AI agents will exist. Even if developing this kind of AI is irresponsible, such developments are likely and will require ethics engines, or risk no consideration of ethics at all. Even if fully automated AI is scary, such AI with automated ethics is better than such AI without. In such a world, the input parser in my ethics engine would have to be automated. This would require that the parser translate the AI agent's internal representation to the appropriate logical representation. The input parser would need enough common sense reasoning to determine what circumstances are morally relevant to a maxim. This is a question that, like all of ethics, philosophers debate robustly [5]. It is likely that, just as different implementations of automated ethics choose a particular ethical theory and implement it, different implementations of such an input parser would need to adopt different interpretations of commonsense reasoning and morally relevant circumstances. Automating

[4]For more discussion of different models of AI ethics, see Section UNWRITTEN

[5]Powers [Powers, 2006] identifies this as a challenge for automating Kantian ethics and briefly sketches solutions from O'Neill [O'Neill, 1990], Silber [Silber, 1974], and Rawls [Rawls, 1980].

this level of commonsense reasoning would represent signifiant technical progress in computational ethics.

Once the input has been parsed, either by a human or a machine, into a sentence in my logic, my project can evaluate its moral status using my implementation of the FUL. Concretely, my project would return a value indicating if the maxim is obligatory, permissible, or prohibited. The maxim would be prohibited if it fails the universalizability test, permissible if it passes, and obligatory if its negation fails the universalizability test. All three of these properties amount to testing if a certain theorem holds or not in my logic, a calculation that I demonstrate in my tests.

This output could then be converted into some actionable, useful response with another output parser, and then passed back to the AI agent. For example, if the AI agent is equipped to evaluate natural language prescriptions, the status of the maxim could be parsed into a natural language sentence. This output will be passed back to the AI agent, which will use it to make a decision. The input parser, categorical imperative library, and output parser together constitute an "ethics engine" that AI agents could use as a black box implementation of an ethical theory.

The ethics engine depicted above is a high-level example of one way to use my project to guide an artifical agent. The upshot is that an automated version of the categorical imperative could become part of the ethical engine for an AI agent, with much work to parse the input and the output. Effectively, the kind of automated ethics I implement could be a library that AI developers use to give AI agents the capacity for sophisticated ethical reasoning faithful to philosophical literature. This represents an improvement over existing ethics engines, which rarely attempt to capture the complexity of any ethical theory that philosophers plausibly defend. Moreover, a logic programming approach is potentially more explainable than a black-box deep learning approach, as theorem provers like Isabelle can explicitly

list the axioms used to generate moral prescriptions. For more on how my project is situated among other work in automated ethics, see Section Related Work.

4.2 Computational Philosophy

Above I explained how my system offers a mechanism for humans to build ethical AI agents. I also argue that computational ethics is a mechanism for computers to help humans think differently about philosophy. Just as theorem provers make mathematics more efficient and push mathematicians to think precisely about the phenomena they are modelling, computational ethics can help philosophers think more precisely about philosophy. Below I share a personal example of the kind of philosophical insight that computational ethics can prompt and analyze the value that this tool offers to philosophers.

4.2.1 Example of a Philosophical Insight

As I implemented a formalization of the categorical imperative using an interactive theorem prover, I discovered logical insights. These logical insights led to a philosophical insight that was novel to me and is potentially novel to the field. While this insight could have been reached without the help of the computer, my system's logical results provoked an interesting philosophical conversation. Effectively, the system spat out a logical principle, and I examined the philosophical plausibility of this principle for the ethical theory I am formalizing. In this section, I will first present the logical insight, then the philosophical insights, and then its implications for a debate about self-doubt. I will then generalize my personal experience and argue that computational ethics is a new, useful methodology for philosophers.

I arrived at the logical insight while testing my formalization of the FUL. I realized that my formalization was inconsistent unless I specified that the FUL only held for "well-formed maxims," such that neither the act nor goal were already achieved

in the given circumstances. Precisely, a circumstance, act, goal tuple (c, a, g) is well-formed if $(\neg(c \longrightarrow a)) \wedge (\neg(c \longrightarrow g))$. This provoked the philosophical insight that maxims of this form, in which the act or the goal has already been accomplished in the given circumstances are "vacuous" because any prescriptions they generate have already been acted on or violated. The notion of a vacuous maxim has implications for debates about ethical self-doubt and self-confidence or self-respect.

Logical Insight

First, I used Sledgehammer to show that my formalization of the FUL[6] resulted in a contradiction. Sledgehammer was able to tell me which axioms it used to complete this proof, showing me that my formalization contradicted the axiom O_diamond, which states that an obligated term cannot contradict its context[7]. O_diamond formalizes the principle "ought implies can" and requires that if A is obligated in context C, that A is possible in context C. I hypothesized that there was some tension between the antecedent of the FUL, which states that all agents act on the maxim, and the consequent, which states that the maxim is prohibited. If the maxim has already been acted on, then not acting on it is impossible. Thus, the generated prohibition is impossible to obey, so the ought implies can principle and axiom O_diamond are violated.

I then experimented with modifications of the FUL, which I eventually abandoned. I tried universalizing the maxim at a world other than the current world and defining non-contradictory maxims, in which the maxim's circumstances do not contradict the maxim's act. I noticed that, no matter what modifications I made, Nitpick was timing out when looking for a model and Sledgehammer wasn't able to find a proof of inconsistency. Isabelle's proof tools weren't able to tell me if my modifications

[6]The full logical representation is *FUL0* $\equiv \forall c\ a\ g\ s.$ *not-universalizable* $(c,\ a,\ g)\ s \longrightarrow \models$*prohibited* $(c, a, g)\ s$.

[7]The full form of the axiom is $O\{A|B\} \longrightarrow \diamond(B \wedge A)$

were consistent or not. I suspected that something about my implementation was too slow, perhaps due to my liberal use of quantifiers[8].

Isabelle's model checker Nitpick performs an optimized version of a brute force model search, in which it generates many models and checks if they satisfy the given maxims. I suspected that Nitpick was checking large models that exhausted its time limit, especially due to the logical complexity of my theory. To reduce the logical complexity, I decided to specify the exact number of maxims in the system by passing as an argument to Nitpick the cardinality of my desired model. Nitpick no longer timed out, but it could not find a satisfying model with cardinality 1, and thus could not demonstrate that my modified FUL was consisteny. This puzzled me, as I felt that I could construct a pencil-and-paper model with a single world and term in which my modified formalizations were consistent.

Instead of specifying the cardinality of the model, I decided to tell Nitpick exactly how many maxims there were in my system by defining them as constants. I defined a particular (circumstance, act, goal) tuple as a constant. Instead of stating that the FUL held for all maxims, I stated that the FUL held for the specific maxim formed by this tuple. While before I added the axiom $\forall(c, a, g)$*FUL holds for maxim*(c, a, g), I now added constants (c, a, g) and added the axiom *FUL holds for maxim*(c, a, g). By specifying the circumstance, act, and goal as constants, I removed the external universal quantifier, thus removing a layer of logical complexity.

To my surprise, Nitpick not only returned quickly, it was able to show that the FUL was consistent!

This result was counterintuitive—after all, what is the difference between a model of cardinality 1 and a model with one constant object? Why is quantifying over a tiny number of maxims different than analyzing a single maxim? Professor Amin

[8] Benzmueller warned me that as I added quantifiers to the theory, Isabelle's automated proof tools may start to time out.

pointed out that, as constants, the circumstances, act, and goal were all distinct. When they were quantified over, they could be identical. To formalize this idea, I defined a maxim as *well-formed* $\equiv \lambda(c,\ a,\ g)\ s\ w.\ \neg\ c{\rightarrow}g\ w \wedge \neg\ c{\rightarrow}a\ s\ w.$ In propositional logic, a circumstance, act, goal tuple (c, a, g) is well-formed if $(\neg(c \longrightarrow a)) \wedge (\neg(c \longrightarrow g))$. I tested my hypothesis by modifying my axiom to instead read $\forall$*maxim (maxim is well-formed* $\longrightarrow$ *FUL holds for maxim*). This version of the FUL was indeed consistent!

To summarize, I realized that my initial attempt at formalizing the FUL was inconsistent because it required that the FUL hold for badly formed maxims, in which the circumstances entail the act or goal. The logical insight was that if FUL holds for maxims in which $(c \longrightarrow a) \vee (c \longrightarrow g)$, then the logic will be inconsistent.

Philosophical Insight

Once I realized this logical property, I tried to understand its philosophical plausibility. I wanted to philosophically test the hypothesis that maxims in which $(c \longrightarrow a) \vee (c \longrightarrow g)$ are not valid inputs to the FUL. I concluded that because vacuous maxims neither change an agent's behavior nor generate meaningful obligations, they are not the right kinds of questions for practical reasoners to be asking. They cannot be action-guiding and are thus not the domain of ethics. Moreover, under the Kantian account of the will, the very act of asking if a vacuous maxim is prohibited generates a contradiction by undermining the will's authority over itself.

I define a vacuous maxim as one in which the circumstances entail either the act or the goal and argue that such maxims can't meaningfully guide action. Consider the example vacuous maxim, "when eating breakfast, I will eat breakfast in order to eat breakfast." This maxim isn't clearly obligatory or prohibited, but there is something empty about it. Acting on this maxim could never result in any actual action. If an agent adopts this maxim, they decide that, in the circumstances "eating breakfast" they will perform the act "eating breakfast" for the purpose "eating breakfast." In

these circumstances, the act has already been performed! Treating this maxim a law for yourself or a principle to live by doesn't change how you live your life. If you adopt this maxim, when you are eating breakfast, you eat breakfast, but this statement is already tautologically true.

Not only does a vacuous maxim fail to prescribe action, any obligations or prohibitions it generates have already been fulfilled or violated. If a vacuous maxim generates a prohibition, then this prohibition would be impossible to obey. It is impossible to not eat breakfast while eating breakfast, because the circumstances assume that the act has happened. On the other hand, if a vacuous maxim generates an obligation, then the obligation will have already been fulfilled. If you are required to eat breakfast while eating breakfast, then you've already fulfilled your obligation because the circumstances assume that the act has happened. Thus, a vacuous maxim does not actually guide action because it doesn't generate new obligations or prohibitions that could ever be acted on.

Because vacuous maxims can't prescribe or alter action, they are not practically action-guiding and thus are not the right kinds of maxims for practical reasoners to evaluate. Moreover, if ethics is supposed to guide action, then vacuous maxims cannot be the domain of ethics. Practical reason is the kind of reason that helps us decide what we should do. A practical reasoner asks moral questions not as a mental puzzle or out of curiosity, but in order to decide how to act. Practical reason is action-guiding, but a vacuous maxim can never be action-guiding because it prescribes no new actions or obligations. It is not the kind of maxim that a practical reasoner would consider. There is no explicit prohibition against a vacuous maxim like the breakfast example above, but it is the wrong kind of question for a practical reasoner to ask. On ordinary person trying to navigate the world would never ask that kind of question. If ethics is meant to guide action, then badly formed maxims are not questions for ethics, because they could never guide action.

Above I argued that vacuous maxims are not the kind of principle that a practical reasoner would ever ask, and are thus not the right kind of question for ethics. Kantians can make an even stronger claim about vacuous maxims—because maxims are laws that you give to yourself, asking if you should will a maxim as you will it undermines your will's law-giving ability. The circumstances of a vacuous maxim already assume that the agent has willed the maxim. Under the Kantian acount of willing, this act of willing a maxim is equivalent to giving the maxim to yourself as a law. When you will a maxim, you adopt a law to make the maxim your end and commit yourself to be its cause. You cannot simultaneously commit yourself to a maxim and ask if you should be committing to it. To will the maxim is to adopt it as law—so the question, "should I be willing this?" is paradoxical. Either you haven't actually made the maxim your law (and thus haven't yet committed yourself to it), or you aren't actually asking the question (because the decision has already been made). Because a maxim is a law that you give to yourself, you cannot question it absent a sufficient reason (such as a change in the circumstances). To question a law arbitrarily is to not regard it as a law at all. This kind of questioning amounts to questioning the will's authority over itself, but this is impossible. The will definitionally has authority over itself, for that is what it is to be a will.

A skeptic may argue that we do often ask "should I be doing this?" as we do something. What do we mean when we ask this question? In what sense are we trying to evalute the moral status of a vacuous maxim? Can this kind of question ever be valid? To understand this worry, I consider the maxim, "When dancing, I should just dance for the sake of dancing."[9] While this maxim appears to be vacuous (the circumstance 'dancing' implies the act and goal of dancing), it's a question that practical reasoners do ask. I argue that there are this maxim is actually misunderstood and, when interpreted correctly, it no longer poses as a counterexample to

[9]Maybe cite Korgsaard since the dancing thing is her example.

my complaints about vacuous maxims.

Under one reading of this maxim, "I should just dance" is actually referring to a different act than the circumstance "when dancing". The circumstance "when dancing" refers to rythmically moving your body to music, but "I should just dance" refers to dancing without anxiety, completely focused on the joy of dancing itself. More precisely, this maxim should read "When dancing, I should abandon my anxiety and focus on dancing for the sake of dancing." This maxim when so modified is not vacuous at all—abandoning anxiety and focusing on dancing is an entirely different act from moving your body rythmically to music. This maxim is actually well-formed, and thus doesn't pose a problem for my argument. It is entirely plausible to tell yourself "When I am dancing, I should focus on dancing for the sake of dancing itself." The circumstances do not entail the act or the goal because they refer to different meanings of the word dancing. Any valid reading of this maxim will have the structure above, in which the act is actually different from the circumstances. A reasoner cannot accept their will as law-giving or commit themselves to an act and simultaneously question the act. Either they must be questioning a different act or they must have recieved new information to prompt the questioning, modifying the circumstances of the original maxim.

Another related worry has to with maxims that we do in fact think are prohibited. Consider the maxim modified to read "When dancing and seeing a child drowning, I should dance for the sake of dancing." Clearly this maxim is fit for moral evaluation, and we expect a moral theory to prohibit this maxim. The circumstances "When dancing and seeing a child drowning" appear to entail the act of dancing, and the maxim thus appears vacuous. Once again, this maxim is formulated incorrectly. In this case, the question that the agent is actually asking themselves is "should I continue dancing?" That is the maxim that they will adopt or reject. They mean to ask if they should stop dancing and go help the child. Dancing at

the current moment and dancing at the next moment are different acts, and the circumstances imply the former but not the latter. A vacuous maxim would have circumstances and act both "dancing at moment t," but this maxim has circumstances "dancing at moment t" and act "dancing at moment t+1." This is a kind of temporal error that has bearings for other debates in ethics as well. Specifically, the confusion between circumstances and acts that occur at different times (as in this example) and circumstances and acts that occur at the exact same time has bearing on self-doubt, as I will argue next.

Implications for Self Doubt and Self Respect

The dancing maxim can also be understood through the lens of self-doubt. Under this reading, the question "When I am dancing, should I be dancing for the sake of dancing?" is the agent asking, "Am I doing the right thing right now?" Unlike the drowning example, the agent is not asking about the next moment, but is expressing doubt about the moral validity of their behavior at this current moment. I do not want to argue that self-doubt always undermines the will—after all, self-doubt plays an important role in moral reasoning and is often the mark of a thoughtful agent. I argue instead that questions of self-doubt do not actually involve vacuous maxims, for these are not the maxims that the agent is doubting. Indeed, this example demonstrates that the tension between self-doubt and self-respect arises from a mistaken characterization of questions of self-doubt as questions about vacuous maxims. I first explain the tension between self-doubt and self-respect in epistemology, then explain the parallel tension in ethics, and finally present a resolution of this tension.

In epistemology, there is a tension between the rational requirement to believe in yourself and the value of self-doubt, in moderation. Christensen presents the "principle of self-respect," which requires that any rational agent refrain from believing that they have mistaken beliefs [Christensen, 2007, 4]. For example, I cannot ratio-

nally both believe that the sky is blue and believe that I believe that the sky is green. In other words, I cannot disapprove of my own credences. Christensen argues that this principle, which he abbreviates to SR, holds because a perfectly rational agent can make accurate and confident judgements about what they believe. If this is the case, violating SR results in a simple contradiction [Christensen, 2007, 8-9].

While most philosophers accept some version of SR[10], Roush argues that the principle must be modified in order to account for healthy epistemic self-doubt. She argues that, while pathological second-guessing is roundly criticized, we are generally imperfect beings, and some sensitivity to our own limitations is a virtue [Roush, 2009, 2]. Indeed, even Christensen acknowledges that total self-confidence is an epistemic flaw [Christensen, 2007, 1]. Thus, there is tension between the rational requirement to respect our authority as believers and the practical reality that we are often wrong.

This debate between self-respect and self-doubt in epistemology also applies to ethics. When we decide to act and commit ourselves to acting, we cannot simultaneously doubt the validity of our action. If human behavior is purposive, then the very act of committing oneself implies that one has sufficient reasons for committing oneself. These reasons may be flawed, but in making the commitment, the reasoner has accepted them. It is contradictory to claim that someone commits and questions simultaneously, because commitment itself implies a resolution to the question. Either the commitment is not real, or the question is not. I will call the principle that one cannot will a maxim and simultaneously question if they should will that maxim "ethical self-respect" or ESR.

On the other hand, self-doubt is an important part of ethical reasoning. Just as believers are often mistaken, so are practical reasoners. An agent with perfect confidence, who is always sure that they are doing the right thing, is clearly not thinking

[10] Van Fraassen, Vickers, Koons [Christensen, 2007, 5]

deeply enough about their obligations. Some degree of ethical self-doubt is normal and likely desirable. Thus, there is a tension between the rational requirement of ESR and the intuitive validity of ethical self-doubt (ESD).

To resolve this tension, I return to my earlier example of a dancer. Imagine Sara is dancing at a weddding, when, in a moment of angst, she asks herself, "Should I really be dancing right now?" What question is she asking here? The immediate answer is that she asking if the maxim, "When dancing at your friend's wedding, dance for the sake of dancing" is a permissible maxim to act on. Notice that t he maxim in question is vacuous: the circumstance "when dancing at a friend's wedding" implies the act "dance." Because this is a vacuous maxim, it cannot be the maxim that she is questioning, for adopting this maxim could not have changed her behavior at all. Sara is asking a question about her actions and their validity. Any conclusions about the validity of a vacuous maxim would not help her, first because the maxim has no effect on her action, and second because any such validity would be a foregone conclusion as she has already adopted the maxim. As I argued above, no practical reasoner can coherently ask themselves whether a vacuous maxim is valid or not without undermining their will, which is a contradiction. Thus, under the interpretation of self-doubt as a vacuous maxim, the tension between ESR and ethical self-doubt appears irresolvable. Those committed to this interpretation must abandon one principle or the other.

To resolve this issue, I turn to another interpretation of ethical self-doubt. Under this interpretation, when Sara asks, "Should I really be dancing right now?" she wants to know if the maxim that resulted in the current moment when she is on the dance floor was actually the right thing to will. She is asking if she made the right decision in the past, when she decided to dance. The maxim that initiated the dancing would be something like "When at a wedding, dance for the sake of dancing." This is the maxim that she is currently acting on, not the vacuous maxim

"When dancing, dance for the sake of dancing." Under this interpretation, there is no tension at all between self-doubt and self-respect. It is perfectly valid for a reasoner to doubt their prior moral judgements, just as it is perfectly rational for a believer to doubt their past beliefs [Christensen, 2007, 3-4]. Such doubt does not undermine the reasoner's decision-making capacity and is thus perfectly consistent with ethical self-respect.

Not only does this second interpretation resolve the tension between ESR and ESD, it also more accurately tracks the operation of practical reason. As argued above, a practical reasoner would never ask themselves whether or not to will a vacuous maxim, because such a maxim would generate no meaningful obligations. Adopting such a maxim would not alter their behavior in any way. Moreover, the fact that a practical reasoner never adopts a vacuous maxim demonstrates the cause of the tension between ESR and ESD. The tension itself arises from a misreading of questions of self-doubt as questions about the evaluation of vacuous maxims. A question of self-doubt cannot refer to a vacuous maxim and must instead refer to a well-formed maxim about the agent's past decision-making. As seen before, cases where agents appear to ask themselves about vacuous maxims are mistaken about the maxim in question, because such a question could never yield a useful answer for a practical reasoner.

4.2.2 The Value of Computational Ethics

I will now generalize from the personal insight reached above to the methodological value of computational ethics for philosophers. I do not argue that computational ethics, as it stands today, uncovers philosophical insights that humans have not reached or are incapable of reaching. After all, my understanding of a well-formed maxim could very well exist in the literature and certainly could be reached by a philosopher without any computational tools. Instead, I argue that computa-

tional tools prompt philosophers to ask questions that lead to insights. Philosophers already value precision, and the computer forces precision and makes formal reasoning easier. Computational ethics can serve as another tool in a philosopher's arsenal, like a thought experiment or counterexample. While the technology is not yet mature enough and easy enough to use to become widespread in philosophy departments, technical progress could turn computational ethics into an easy-to-use tool for philosophers that doesn't require any specialized progamming or logical knowledge.

The first contribution of computational ethics is precision. Much of analytic philosophy involves making a particular concept precise. Thought experiments, arguments, counterexamples, and examples illustrate features of a concept in the hope of making the concept itself more precise. Computational ethics can help philosophers reach this goal of precision in another, potentially easier, way. Representing a philosophical idea in logic and implementing it in an interactive theorem prover requires making the idea precise to a degree that ordinary discussion can result in, but does not necessarily require. The initial representation of an idea in a logic requires making its form precise. For example, as I formalized the notion of a maxim, I had to understand its components and define it as a circumstance, act, goal tuple. Moreover, Isabelle's strict typing system required that I define coherent, consistent types for each of these entities and for a maxim as a whole. This requires understanding what role each of these components play in the FUL and assigning them each a type. In my example, I concluded that circumstances and goals are terms, which can be true or false at a world, and acts are open sentences, which are true for a particular subject at a particular world. This precision is possible without computational tools, but computational ethics forces a level of precision that ordinary discussion does not demand. Type fuzziness and overloaded definitions are all too common in philosophical writing and discussion (would be cool to cite some

famous debate revolving around this idea), but computers don't allow this kind of imprecision.

Another, related benefit of computational ethics is that it makes formal ethics far less tedious. Certain subfields, such as philosophy of language, see such benefit in precision that they already use symbolic logic to represent philosophical concepts, just as mathematicians use symbolic logic to represent mathematical concepts. Some of this work requres tedious pencil and paper proofs to prove theorems, even when many of these theorems may not generate relevant philosophical insights. Ineractive theorem provers make formal logic more accessible. Isabelle can complete a proof, starting from first principles, in a matter of seconds that would take a logician pages to complete. Similarly, Nitpick can generate examples or counterexamples to a proposition in a brute force manner. The computer can generate hypotheses that philosophers can then think about, just as it did for me to prompt the insight above.

Just as calculators make arithmetic more accessible, computational ethics does the same for formal philosophy. Not all philosophy can or will be formalized or automated—after all, calculators didn't make accountants or mathematicians obsolete. Just as computers reduce the tedium in other aspects of our life, they can reduce the tedium involved in formal logic to allow mathematicians and philosophers to focus their attention on understanding.

4.2.3 Looking Forward

Computational ethics is at its infancy. The use of theorem provers in mathematics is just now beginning to make headway [Buzzard, 2021], even though theorem provers were first invented in the 1960's [Harrison et al., 2014]. In contrast, the first attempts to use theorem provers for ethics occurred in the last decade. The fact that this nascent technology is already helping humans reach non-trivial philosophical

conclusions is reason to, at the very least, entertain the possibility of a future where computational ethics becomes as normal for philosophers.

To the skeptic, the ethical insights uncovered by the computer are not necessarily impressive philosophy. Indeed, the fact that a theorem prover requires specialized knowledge outside of the field of philosophy indicates that the technology is nowhere near ready for universal use in philosophy departments. However, history indicates that as computing power increases and computer scientists make progress, computational ethics will become more usable. Theorem provers in mathematics began as toys incapable of proving that the real number 2 is not equal to the real number 1, but Buzzard showed that moving from such a primitive system to a tool for Fields medal winning mathematics is possible in a matter of years [Buzzard, 2021]. Countless examples from the history of computer science, from the Turing Test to AI game playing to protein folding, demonstrate that progress in computer science can make seemingly obscure computer programs useful and usable in ways that exceed our wildest imaginations. Indeed, programmable computers themselves initially began as unwieldy punch card readers, but their current ubiquity need not be stated. If computer scientists and philosophers invest in computational ethics, it can become as much a tool for philosophy as a calculator is for for arithmetic.[11]

[11] Is this too like, lalalala fantasy of computational philosophy? Would it be less so if I did more work explaining the history of theorem proving for math? Is this even that important for my project?

5 Why Kantian Ethics

In this thesis, I automate Kantian ethics. Deontological theories are attractive candidates for automation because, as Powers argues, rules are generally computationally tractable [Powers, 2006, 1]. Intuitively, algorithms are rules or procedures for problem solving and Kantian ethics, specifically the Formula of Universal Law, is one such procedure for the problem of making ethical judgements. I will make this intuition precise by arguing that Kantian ethics is natural to formalize because evaluating a maxim requires little additional data about the world and a maxim is relatively easy to represent to a computer. Moreover, while all ethical theories have debates that an automated ethical system will need to take a stance on, these debates are less frequent and controversial in the case of Kantian ethics as compared to consequentialism and virtue ethics.

I do not aim to show that Kantian ethics is the only tractable theory to automate or to present a comprehensive overview of all consequentialist or virtue ethical theories. Instead, I present a sample of some approaches in each tradition and argue that Kantian ethics is more straightforward to formalize than these approaches. Insofar as my project serves as an early proof-of-concept for computational ethics, I choose to automate the ethical theory most amenable to automation.

Future work could and likele should address the challenges I outline in this section. The more ethical theories that computational tools can handle, the more valuable computational philosophy becomes both for philosophers and for AI agents. My project merely contributes a formalization of one theory that poses fewer challenges than others.

5.1 Kantian Ethics

5.1.1 Crash Course on Kantian Ethics

Kant's theory is based on the idea of practical reason, which is the kind of reason that we use to decide what to do. In *The Groundwork of the Metaphysics of Morals*, Kant's most influential text on ethics, he explains that rational beings are unique because we can act "in accordance with the representations of laws" [Kant, 1785, 4:412]. A ball thrown into the air acts according to the laws of physics. It cannot ask itself, "Should I fall back to the ground?" It simply falls. A rational being on the other hand, can ask themselves "Should I act on this reason?" As Korsgaard describes it, when choosing which desire to act on, "it is as if there is something over and above all of your desires, somethig which is you, and which chooses which desire to act on" [Korsgaard and O'Neill, 1996, 100]. Rational beings are set apart by this reflective capacity. A rational being's behavior is purposive and their actions are guided by practical reason. They have reasons for acting, even when these reasons may be opaque to them. Kant calls the operation of practical reason the will.

The will operates by willing maxims, which are its percieved reasons for acting. Kant defines a maxim as the "subjective principle of willing," which is the reason that the will *subjectively* gives to itself for acting [Kant, 1785, 16, footnote 1]. There is debate about what exactly must be included in a maxim, but most philosophers agree that a maxim consists of some combination of circumstances, act, and goal[12]. One example of a maxim is "When I am hungry, I will eat a doughnut in order to satisfy my sweet tooth." When a moral agent wills this maxim, they decide to act on it. They commit themselves to the end in the maxim (e.g. satisfying your sweet tooth). They represent their action, to themselves, as following the principle

[12]For more discussion of the definition of a maxim, see Section What Is a Maxim

given by this maxim. It is maxims that Kant evaluates as obligatory or prohibited. He argues that certain maxims have a form or logical structure that requires any rational agent to will them, and these maxims are obligatory.

The form of an obligatory maxim is given by the categorical imperative, Kant's supreme rule of morality. An imperative is a command, such as "Close the door" or "Eat the doughnut in order to satisfy your sweet tooth." A categorical imperative holds unconditionally for all rational agents under all circumstances. Kant argues that the moral law must be a categorical imperative, for otherwise it would not have the force that makes it a moral law [Kant, 1785, 5]. In order for an imperative to be categorical, it must be derived from the will's authority over itself. Our wills are autonomous, so the only thing that can have unconditional authority over a rational will is the rational will itself. Velleman clarifies this argument by claiming that no one can tell you what reasons to act on because you can always ask why you should obey their authority. The only authority that you cannot question is the authority of your own practical reason. To question this authority is to demand a reason for acting for reasons, which concedes the authority of reason itself [Velleman, 2005, 23]. Therefore, the only possible candidates for the categorical imperative are those rules that are required of the will because it is a will. The categorical imperative must be a property of practical reason itself.

Armed with this understanding of practical reason, Kant presents the categorical imperative. He presents three "formulations" or versions of the categorical imperative and goes on to argue that all three formulations are equivalent. In this project, I focus on the first formulation, the Formula of Universal Law, but will briefly present the other two as well[13].

The first formulation of the categorical imperative is the Formula of Universal Law (FUL), which reads, "act only according to that maxim through which you can at

[13]For more on this choice, see Section Why FUL

the same time will that it become a universal law" [Kant, 1785, 34]. This formulation generates the universalizability test, which tests the moral value of a maxim by imagining a world in which it becomes a universal law and attempting to will the maxim in that world. If there is a contradiction in willing the maxim in a world in which everyone universally wills the maxim, the maxim is prohibited. Velleman presents a concise argument for the FUL. He argues that reason is universally shared among reasoners. For example, all reasoners have equal access to the arithmetic logic that shows that "2+2=4" [Velleman, 2005, 29]. The chain of reasoning that makes this statement true is not specific to any person, but is universal across people. Therefore, if I have sufficient reasont to will a maxim, so does every other rational agent. There is nothing special about the operation of my practical reason that other reasoners don't have access to. Practical reason is shared, so in adopting a maxim, I implicitly state that all reasoners across time also have reason to adopt that maxim. Therefore, because I act on reasons, I must obey the FUL. Notice that this fulfills the above criterion for a categorical imperative: the FUL is derived from a property of practical reason itself and thus derives authority from the will's authority over itself, as opposed to some external authority.

The second formulation of the categorical imperative is the formula of humanity (FUH): "So act that you use humanity, in your own person, as well as in the person of any other, always at the same time as an end, never merely as a means." [Kant, 1785, 41]. This formulation is often understood as requiring us to acknowledge and respect the dignity of every other person. The third formulation of the categorical imperative is the formula of autonomy (FOA), which Korsgaard summarizes in her introduction to the Groundwork as, "we should so act that we may think of ourselves as legislating universal laws through our maxims" [Korsgaard, 2012, 28]. While closely related to the FUL, the FOA presents morality as the activity of perfectly rational agents in an ideal "kingdom of ends," guided

by what Kant calls the "laws of freedom."

The above is not meant to serve as a full defense of Kant's ethical theory, as that is outside the scope of this thesis. Instead, I aim to briefly reconstruct an argument for Kant's ethical theory in the hopes of offering context for the implementation of the FUL I present later in the thesis. Additionally, understanding the structure of Kant's argument also reveals facts about his theory that make it an ideal candidate for formalization.

5.1.2 Ease of Automation

Kantian ethics is natural to formalize because the categorical imperative, and the FUL in particular, is a formal principle of practical reason. In other words, the FUL is a property of reason related to the form or structure of a maxim. It has nothing to do with the circumstances of behavior (beyond those included in a maxim), the agent's mental state, or any other contingent facts. Instead, it is purely a property of a proposed principle for action. This formalism makes Kantian ethics an attractive candidate for formalization. While other ethical theories often rely on many facts about the world or the actor, Kantian ethics simply relies on the form of a given maxim. A computer evaluating a maxim doesn't require any knowledge about the world beyond what is contained in a maxim. A maxim is the only input that the computer needs to make a moral judgement. Automating Kantian ethics merely requires making the notion of a maxim precise and representing it to the computer. This distinguishes Kantian ethics from consequentialism and virtue ethics, which, as I will argue below, require far more knowledge about the world or the agent to reach a moral decision.

Not only does evaluating Kantian ethics only require evaluating a maxim, a maxim itself is an object with a thin representation for a computer, as contrasted to more complex objects like states of affairs or moral character. Later in my project, I ar-

gue that a maxim can be represented simply as a tuple of circumstances, act, and goal[14]. This representation is simple and efficient, especially when compared to the representation of a causal chain or a state of affairs or moral character. A maxim is a principle with a well-defined form, so representing a maxim to the computer merely requires capturing this form. This both reduces the computational complexity (in terms of time and space) of representing a maxim, it also make the system easier for human reasoners to interact with. A person crafting an input to a Kantian automated agent needs to reason about relatively simple units of evaluation, as opposed to the more complex features that consequentialism and virtue ethics reqire. I will make the comparision to consequentialism and virtue ethics explicit below.

I do not argue that Kantian ethics is the only theory possible to formalize, but instead that Kantian ethics is relatively easier to formalize. Indeed, like any ethical theory, Kantian ethics contains several debates that philosophers disagree on. In this project, I assume stances on debates about what exactly constitutes the form of a maxim and the correct precise interpretation of the Formula of Universal Law. Those who disagree with my stances will not trust the judgements of my system. Unlike consequentialism or virtue ethics, these debates are considered close to settled in the Kantian literature [Ebels-Duggan, 2012]. Because such debates are fewer and less controversial than those in virtue ethics or consequentialism, Kantian ethics is relatively easier to formalize.

5.2 Consequentialism

A consequentialist ethical theory is, broadly speaking, any ethical theory that evaluates an action by evaluating its consequences[15]. For example, utilitarianism is

[14]For more, see Section What is a Maxim?

[15]There is long debate about what exactly makes an ethical theory consequentialist [Sinnott-Armstrong, 2021]. For this paper, I will focus on theories that place the moral worth of an act in its the consequences.

a form of consequentialism in which the moral action is the action that results in the best consequences or produces the most good [Driver, 2014]. The focus on the consequences of action distinguishes consequentialists from Kantians, who derive the moral worth of an action from the maxim that is acted on as determined at the time of action. Some debates in the consequentialist tradition include discussions of which consequences of an action matter, what exactly constitutes a "good" consequence, and how we can aggregate the consequences of an action for all the individuals involved.

5.2.1 Which Consequences Matter

Because consequentialism evaluates the state of affairs following an action, this kind of ethical reasoning requires more knowledge about the state of the world than Kantian ethics. Under a naive version of consequentialism, evaluating an action requires perfect knowledge of all consequences following an action. This requires that an automated ethical system somehow collect all of the infinite consquences following an action, a likely impossible task. Moreover, compiling this database of consequences requires answering difficult questions about which consequences were actually caused by an action[16].

These challenges also apply to human reasoners, so most consequentialists do not actually think that agents need to calculate all the consequences of their actions. Plausible strategies to avoid this problem include stopping calculation early because constant calculation paralyzes action or only evaluating consequences that the agent could reasonably forsee before acting. Another solution is to adopt the legal notion of "proximate cause," which only holds the agent responsible for the immediate consequences of their acts, but not for consequences resulting from others' voluntary responses to the agent's original act [Sinnott-Armstrong, 2021].

[16] maybe cite the debate about difficulties in determining causation?

Even without understanding the details of these views, it is clear that they require more data than Kantian ethics and scale poorly with the complexity of the act being evaluated. Even if we cut off the chain of causal reasoning at some point based on one of the rules above, evaluating the consequences of an action is still data-intensive. Even evaluating the first or immediate cause of an action requires knowledge about the state of the world before and after an action. Kantian ethics, on the other hand, requires only knowledge about the maxim, not knowledge about the state of the world when the maxim is adopted. Consequentialism requires knowledge about the situation in which the act is performed and following the act, whereas Kantian ethics merely requires knowledge about the act itself. For simple acts, this data does not seem unreasonable, but as acts become more complex and affect more people, the computational time and space required to calculate and store their consequences increases. Kantian ethics, on the other hand, does not suffer as intense of a scaling challenge because maxims that affect 1 person and maxims that affect 1 million people both have the same representation.

The fact that consequentialism requires more knowledge about the world makes it more difficult to formalize. Automated consequentialist ethics would need to represent complex states of the world and causal chains in an efficient manner and reason about them. This both presents a difficult technical challenge and impedes the usability of such a system. Automated consequentialist ethics would need to come equipped with a large enough database of knowledge about the world to extrapolate the consequences of an actions and up-to-date information about the state of the world at the moment of action. Not only does collecting and representing this data pose a technical challenge, it also creates a larger "trusted code base" for the automated system. Trusting my Kantian ethical reasoner merely requires trusting the logical implementation of the categorical imperative, but trusting a consequentialist ethical reasoner requires trusting both the logical machinery that actually

evaluates the consequences of an act and the background/situational knowledge that serves as an input to this machinery.

The challenge of understanding and representing the circumstances of action is not unique to consequentialism, but is particularly acute for consequentialism. Kantian ethicists robustly debate which circumstances of an action are "morally relevant" and should be included in the formulation of the maxim[17]. Those using my system will need to either automatically or manually formulate maxims, a process that involves common-sense reasoning to determine which circumstances, act, and goal tuples the agent is adopting [18]. Because Kantian ethics merely evaluates a single maxim, the surface of this debate is much smaller than the debate about circumstances and consequences in a consequentialist system. An automated consequentialist system must make such judgements about the act itself, the circumstances in which it is performed, and the circumstances following the act. All ethical theories relativize their judgements to the situation in which an act is performed to some extent, but consequentialism requires far more knowledge about the world than Kantian ethics.

5.2.2 Theory of the Good

Another debate that an automated consequentialist reasoner would need to take a stance on is the question of what qualifies as a "good consequence," or what the theory of the good is. Hedonists associate good with the presence of pleasure and the lack of pain. Preference utiliarians believe that good is the satisfaction of desires and is thus related to individuals' preferences, as opposed to some sensation of pleasure or pain. Other consequentialists adopt a pluralistic theory of value, under which many different kinds of things are good for different reasons. For ex-

[17] Powers [Powers, 2006] identifies this as a challenge for automating Kantian ethics and briefly sketches solutions from O'Neill [O'Neill, 1990], Silber [Silber, 1974], and Rawls [Rawls, 1980].

[18] For more on the parsing of ethical dilemmas into maxims, see Section AI Ethics

ample, Moore values beauty and truth and other pluralists value justice, love, and freedom [Moore, 1903]. Welfare utilitarians value a person's welfare and utilitarians of right value states of affairs in which respect for some set of crucial rights is maximized [Sinnott-Armstrong, 2021].

Most of the above theories of good require that a moral reasoner understand complex features about individuals' preferences, desires, or sensations in order to evaluate a moral action, making automated consequentialist ethics difficult. Regardless of the theory of the good, a consequentialist ethical reasoner needs to evaluate a state of affairs, which encompass each involved individual's pleasure, preferences, welfare, freedom, rights, or whatever other criteria make a state good. This requires judgements about whether or not a state of affairs actually satisifes the relevant criteria for goodness. These judgements are difficult and debateable, and any consequentialist decision requires many of these judgements for each individual involved. As systems become more complex and involve more people and more acts, making these judgements quickly becomes difficult, posing a scaling challenge for a consequentialist ethical reasoner. Perfect knowledge of tens of thousands of people's pleasure or preferences or welfare or rights is impossible. Either a human being assigns values to states of affairs, which quickly becomes difficult to scale, or the machine does, which requires massive common-sense, increases room for doubting the system's judgements, and simplifies the judgements. This is a tractable problem, but is much more difficult than the Kantian task of formulating and evaluating a maxim.

5.2.3 Aggregation

Once an automated consequentialist agent assigns a goodness measurement to each person in a state of affairs, it must also calculate an overall goodness measurement for this state of affairs. One approach to assigning this value is to aggregate each

person's individual goodness score into one complete score for a state. For example, under a simple welfare model, each person is assigned a welfare score and the total score for a state of affairs is the sum of the welfare scores for each involved person. The more complex the theory of the good, the more difficult this aggregation becomes. For example, pluralistic theories struggle to explain how different kinds of value can be compared [Sinnott-Armstrong, 2021]. How do we compare one unit of beauty to one unit of pleasure? Subjective theories of the good, such as those focused on the sensation of pleasure or an individual's preferences, present difficulties in comparing subjective measures across people. Resolving this debate requires that the automated reasoner choose one specific aggregation algorithm, but those who disagree with this choice will not trust the reasoner's moral judgements. Moreover, for complex theories of the good, this aggregation algorithm may be complex and may require a lot of data.

To solve this problem, some consequentialists reject aggregation entirely and instead prefer wholistic evaluations of a state of affairs. While this approach no longer requires that a reasoner define an aggregation algorithm, the reasoner still needs to calculate a goodness measurement for a state of affairs. Whereas before the reasoner could restrict analysis to a single person, the algorithm must now evaluate an entire state wholistically. Evaluating the goodness of an entire state of affairs is more complicated than evaluating the goodness of a single person. As consequentialists modulate between aggregation and wholistic evaluation, they face a tradeoff between the difficulty of aggregation and the complexity of goodness measurements for large states of affairs. This tradeoff also holds for an automated consequentialist moral agent. Such an agent either needs to define an aggregation function, thus opening the door to those who disagree with this definition, or needs to evaluate the goodness of entire states of affairs, which is a complex and data-intensive philosophical and technical challenge.

5.2.4 Prior Attempts to Formalize Consequentialism

None of the challenges described above are intractable or capture the full literature of all variations of consequentialism. I do not argue that consequentialism is impossible to automate. Instead, each of the challenges above requires that the developer "plant certain flags" and take a stance on philosophical debates. These difficulties are present in any ethical theory, but consequentialism has more such points of difficulty than Kantian ethics and is thus more difficult to automate both computationally and philosophically.

Because of its intuitive appeal, automated ethicists have tried to formalize consequentialism in the past. These efforts cannot escape the debates outlines above. For example, Abel et. al represent ethics as a Markov Decision Process (MDP), with reward functions customized to particular ethical dilemmas [Abel et al., 2016, 3]. While this appears to be a convenient representation, it either leaves unanswered or takes implicit stances on the debates above. It assumes that consequences can be aggregated just as reward is accumulated in an MDP. It leaves open the question of what the reward function is and thus leaves the theory of the good, arguably the defining trait of a particular consequentialist view, undefined. Similarly, Anderson and Anderson's proposal of a hedonistic act utilitarian automated reasoner chooses hedonism[19] as the theory of the good [Anderson et al., 2004, 2]. Again, their proposal assumes that pleasure and pain can be given numeric values and that these values can be aggregated with a simple sum, taking an implicit stance on the aggregation question. Other attempts to automate consequentialist ethics will suffer similar problems because, at some point, to actually use an automated consequentialist moral agent, the debates above will need to be resolved.

[19] Recall that hedonism views pleasure as good and pain as bad

5.3 Virtue Ethics

5.3.1 What Is Virtue

The virtue ethical tradition generally place the virtues, or those traits that constitute a good moral character, at the center. Virtue ethicists recommend actions based on the character traits that such actions would help cultivate. A virtue is commonly accepted as a character trait that ”makes its possessor good” [Hursthouse and Pettigrove, 2018]. For example, under Aristotlean virtue ethics, the virtues are the traits that enable human flourishing or fulfill the purpose of a human being. Many modern virtue ethicists abandon Aristotle's notion of a “purpose” of human beings, and instead define virtue in terms of the characteristic activity of human beings (in ethical terms, not telelological terms) [Snow, 2017]. Just as consequentialists must offer a view of which consequences are good, virtue ethicists must offer some theory of the virtues which presents and justifices a list of the virtues. Such theories vary from Aristotle's virtues of courage and temperance to the Buddhist virtue of equanimity to Robeyns' summary of Sen's conception of the virtues as capabilites that create “effective opportunities to undertake the actions and activities” an agent wants to engage in [Robeyns, 2005]. An automated virtue ethical agent will need to commit to a particular theory of the virtues, opening itself up to criticism from those who disagree with this theory of the virtues. Unlike Kantian ethicists, who generally agree on the meaning of the FUL, virtue ethicists robustly debate which character traits qualify as virtues.

5.3.2 Evaluating Moral Character

Another difficulty with automating virtue ethics is that the unit of evaluation for a virtue ethical theory is often a person's entire moral character. While Kantians evaluate the maxim of an act and utilitarians evaluate the consequences of an act,

virtue ethicists evaluate the actor's moral character and their disposition towards the act. Virtues are character traits and evaluating an action as virtuous or not requires understanding the agent's character and disposition while acting. If states of affairs require complex representations, an agent's ethical character and disposition is even more difficult to represent. Consequentialism posed a data-collection problem in evaluating and representing states of affairs, but virtue ethics poses a conceptual problem about the formal nature of moral character. Formalizing the concept of character appears to require significant philosophical and computational progress, whereas Kantian ethics automatically presents a formal rule to implement.

5.3.3 Machine Learning and Virtue Ethics

One potential appeal of virtue ethics is that many virtue ethical theories involve some form of moral habit, which seems to suggest a machine learning approach. Artistotle, for example, argued that cultivating virtuous action requires making such action habitual through moral education. Under one view of virtue ethics, the virtuous act is what the virtuous person would do. Both of these ideas imply that ethical behavior can be learned from some dataset of virtuous acts, either those prescribed by a moral teacher or those that a virtuous ideal agent would undertake. Indeed, these theories seem to point towards a machine learning model for computational ethics, in which ethics is learned from a dataset of acts tagged as ethical or not ethical. Such approaches exist in the literature, as in the Delphi system [Jiang et al., 2021].

Just as prior work in consequentialism takes implicit or explicit stances on debates in consequentialist literature, so does work in virtue ethics. For example, the dataset that Delphi uses to learn ethical behavior contains implicit views on what the virtues are and how certain acts impact an agent's moral character. Because

Delphi lacks a specified, explicit theory of virtue, it is impossible to determine which traits the system implicitly classifies as virtues without manually examining the dataset.

Machine learning approaches also may suffer explanability problems that my logical, theorem-prover based approach does not experience. Many machine learning algorithms cannot sufficiently explain their decisions to a human being, and often find patterns or correlations in datasets that don't actually cohere with the trends and causes that a human being would identify. While there is significant activity and progress in explainable machine learning, interactive theorem provers are designed to be explainable at the outset. Indeed, Isabelle can even give the axioms and lemmas it used in constructing a proof, allowing a human being to reconstruct the proof independently if they wish. This is, again, not an intractable problem for machine learning approaches to computational ethics, but is one reason to prefer logical approaches.

Explanability is particularly important in the case of ethics because ethical judgements are often controversial and ethics generally requires reflection. First, the most interesting and important ethical judgements are those resulting from ethical dilemmas. These judgements are usually controversial, and the answers are not obvious because people's intuitions differ and different theories generate different answers. In these cases, explainability is particularly important to convince human beings of the correctness of an ethical judgement. If a machine tells us to kill one person to save five without explaining the justification for this decision, using this judgement becomes difficult. Second, ethics is generally seen as a reflective subject. Practical reason is the exercise of using reason and thought to decide what to do. Someone who believes an automated reasoner's judgements without examining or understanding the reasons for these judgements doesn't seem to be doing

ethics correctly[20]. This does not preclude other uses of automated ethics, such as automated moral agents or hypothesis generation for philosophy, but it does make computer- assisted ethical judgement difficult.

[20] I make this argument precise in Section Is CE Even Good For Us?

References

[Abel et al., 2016] Abel, D., MacGlashan, J., and Littman, M. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*.

[Anderson et al., 2004] Anderson, M., Anderson, S., and Armen, C. (2004). Towards machine ethics.

[Bok, 1998] Bok, H. (1998). *Freedom and Responsibility*. Princeton University Press.

[Buzzard, 2021] Buzzard, K. (2021). How do you convince mathematicians a theorem prover is worth their time? Talk at IOHK.

[Christensen, 2007] Christensen, D. (2007). Epistemic self-respect. *Proceedings of the Aristotelian Society*, 107(1pt3):319–337.

[Dietrichson, 1964] Dietrichson, P. (1964). When is a maxim fully universalizable ? 55(1-4):143–170.

[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.

[Ebels-Duggan, 2012] Ebels-Duggan, K. (2012). *Kantian Ethics*, chapter Kantian Ethics. Continuum.

[Ewing, 1972] Ewing, A. C. (1972). *Philosophy*, 47(180):173–175.

[Harrison et al., 2014] Harrison, J., Urban, J., and Wiedijk, F. (2014). History of interactive theorem proving. In *Computational Logic*.

[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.

[Johnson and Cureton, 2021] Johnson, R. and Cureton, A. (2021). Kant's Moral Philosophy. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition.

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

[Kitcher, 2003] Kitcher, P. (2003). What is a maxim? *Philosophical Topics*, 31(1/2):215–243.

[Korsgaard, 2012] Korsgaard, C. (2012). *Groundwork of the Metaphysics of Morals*, chapter Introduction. Cambridge University Press, Cambridge.

[Korsgaard, 2005] Korsgaard, C. M. (2005). Acting for a reason. *Danish Yearbook of Philosophy*, 40(1):11–35.

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

[Moore, 1903] Moore, G. E. (1903). *Principia Ethica*. Dover Publications.

[O'Neill, 1990] O'Neill, O. (1990). *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge University Press.

[O'Neill, 2013] O'Neill, O. (2013). *Acting on Principle: An Essay on Kantian Ethics*. Cambridge University Press.

[Powers, 2006] Powers, T. M. (2006). Prospects for a kantian machine. *IEEE Intelligent Systems*, 21(4):46–51.

[Rawls, 1980] Rawls, J. (1980). Kantian constructivism in moral theory. *The Journal of Philosophy*, 77(9):515–572.

[Robeyns, 2005] Robeyns, I. (2005). The capability approach: a theoretical survey. *Journal of Human Development*, 6(1):93–117.

[Roush, 2009] Roush, S. (2009). Second guessing: A self-help manual. *Episteme*, 6(3):251–268.

[Silber, 1974] Silber, J. R. (1974). Procedural formalism in kant's ethics. *The Review of Metaphysics*, 28(2):197–236.

[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.

[Snow, 2017] Snow, N. E. (2017). *The Oxford Handbook of Virtue*. Oxford University Press.

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

[Wood, 1999] Wood, A. W. (1999). *Kant's Ethical Thought*. Cambridge University Press.