Philosophical Writing

Lavanya Singh

October 21, 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 universalizable, 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

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

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 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 suf-

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

fices 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 current calls from regulators are intensifying the need for "ethical AI". There are many approaches to AI ethics and my project contributes an example of a "top down" approach that starts with an ethical theory (Kantian ethics) and automates it. 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 AI agent into maxims in my logic and an output parser to translate the output of my system into a prescription for the action the AI agent should take. 1 depicts the workflow of this example of an 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. Much of the work that a Kantian human being does when making decisions is to translate everyday situations into appropriate maxims. Huge misunderstandings about Kantian ethics (maybe cite homophobia or right to lie) in the literature often result from incorrectly formulated maxims, and the entire field of applied Kantian ethics is devoted to generating the right kinds of

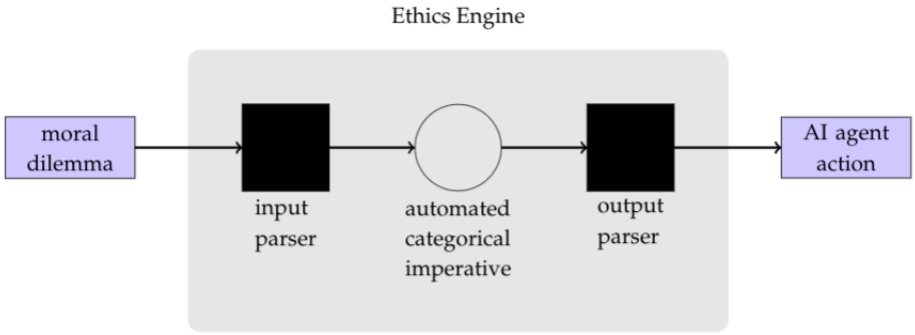

Figure 1: An example of an ethics engine for an artificial agent

maxims to test.

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, of an ethics engine for AI. 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[3].

It is likely that, regardless of the strengths of the human-in-the-loop model, fully automated AI agents will exist. Setting aside the question of whether or not developing this kind of AI is responsible, such developments will require ethics engines, or risk no consideration of ethics at all. Even if a world of fully automated AI is scary, such a world with automated ethics is better than a world 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 (cite the lit, there's tons of it). 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.

[3] maybe i should like...read a book

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 function as 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 an "ethics library" that AI developers could 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. For more on how my project is situated among other work in automated ethics, see Section ??

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 about philosophy. Below I share an 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

The process of implementing and testing a formalization using an interactive theorem prover resulted in logical insights that led to a philosophical insight that was novel to me and is potentially novel to the field. This philosophical insight has implications for what kinds of principles a practical reasoner should be concerned

with. While this insight could have been reached without the help of the computer, my system's logical results provoked an interesting philosophical conversation as I tried to understand their implications for the ethical theory I am formalizing. This serves as an example for how computational ethics can prompt philosophical insights.

As I was implementing 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. Below I document how I came to this conclusion and explain the notion of a vacuous maxim in greater detail.

Logical Insight

First, I used Sledgehammer to show that my formalization of the FUL[4] 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[5]. 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 so the prohibition is impossible to obey so the ought implies can principle is violated, thus contradicting the axiom O_diamond.

To solve this problem, I returned to Korsgaard's practical contradiction interpretation and focused on the imaginatory component of the FUL. Specifically, the universalizability test requires that we imagine a world where the maxim is universalized and study that world to determine whether or not the maxim is obligatory in the actual world. For example, lying is not universalized in our current world, but to determine if lying is prohibited, we imagine a different world in which everyone lied. To capture this difference, I implemented another version of the FUL under which a maxim is prohibited at the current world cw if, when universalized at any world w it is rendered ineffective at w. I hypothesized that this would remove the contradiction found above. To test this new formalization, I used Nitpick, a model checker that quickly generates models that satisfy the given axioms and theorems. Usually, Nitpick can find a satisfying model to show that a logic is consistent in a matter of seconds, but Nitpick consistently timed out when looking for a model of my modified FUL.

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

[5]The full form of the axiom is $OA|B \longrightarrow \diamond(B \wedge A)$

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 timing out due to checking large models that exhausted its time limit, especially due to the logical complexity of my theory[6]. 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. This did not fix the problem.

I next defined a particular (circumstance, act, goal) tuple as a constant and, 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 $\forall(c, a, g).FULholdsformaxim = (c, a, g)$, I now added constants (c, a, g) and added the axiom $FULholdsformaxim = (c, a, g)$. 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 was now 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 so much more time-cosuming than analyzing a single maxim? Professor Amin pointed out that when I defined the circumstances, act, and goal as constants, then they were all distinct. When they were quantified over, they could be identical. To formalize this idea, I defined a maxim as "well-formed" if *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 $\forall maxim(maximiswell - formed \longrightarrow FULholdsformaxim)$. This version of the FUL was indeed inconsistent!

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 define a vacuous maxim as one in which the circumstances entail either the act or the goal. An example of a vacuous maxim is: "When eating breakfast, eat breakfast in order to eat breakfast." This maxim isn't clearly obligatory or prohibited, but there is something empty about it. For one thing, acting on this maxim would 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! Making this maxim a rule to live by doesn't change how

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

you live your life. When you are eating breakfast, you eat breakfast, but this statement is already tautologically true regardless of whether you adopt the maxim or not.

Not only does a badly formed maxim fail to prescribe action, any obligations or prohibitions it generates have already been fulfilled or violated and are thus foregone conclusions. If a badly formed 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 badly formed 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 badly formed maxim does not actually guide action because it doesn't generate new obligations or prohibitions that could ever be acted on.

The implication is that any obligations or prohibitions generated by a badly formed maxim are foregone conclusions and thus do not tell us how to live, and thus must not be the domain of ethics, because ethics is supposed to tell us how to live. Kant's categorical imperative guides our use of practical reason, which is the kind of reason that tells us what we should do. A practical reasoner asks moral questions not because they're mental puzzles or out of curiosity, but because the reasoner wants to know how to act. Practical reason is action-guiding, but a badly formed 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 badly formed maxim like the breakfast example above, but it is the wrong kind of question for a practical reasoner to ask. Moreover, an ordinary person trying to navigate the world would never ask that kind of question. If ethics and practical reason are meant to guide action, then badly formed maxims are not questions for ethics, because they could never guide action.

Above I show that badly formed maxims are not questions for the ethics, but they also represent a stronger problem. Asking if you should be acting while you act is undermining your will. Under the Kantian acount of willing, when you will a maxim, you make it your end and commit yourself to be its cause. You cannot simultaneously will a maxim and ask if you should be willing it. To say that you are willing the maxim is to say that you have decided to will it—so the question, "should I be willing this?" is nonsensical. Either you haven't actually adopted the maxim, or you aren't actually asking the question. In a badly formed maxim, the circumstances already assume that the agent has willed the maxim. If someone asks me"should I eat breakfast while eating breakfast?" I not only wouldn't be able to answer the question, I wouldn't understand what they're asking. I would say, "what do you mean? You ARE eating breakfast." The agent must either not actually be eating breakfast or they must not be asking a real question.

That being said, 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 badly formed 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."[7] While this maxim appears to be badly formed (the circumstance 'dancing' implies the act and goal), it's a question that practical reasoners do ask. I argue that there are multiple ways of understanding this maxim, and none are inconsistent with my complaints about badly formed maxims.

First is what I call the "different action" interpretation, under which "I should just dance" is actually referring to a different act than the circumstances. The circumstances "When dancing" refer 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 actually 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 thus 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.

Another version of the "different action" interpretation is as follows. Consider the maxim modified to be "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. In this case, the question that the agent is actually asking themselves is "should I continue dancing?" That is the actionable 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 badly formed 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."

Another interpretation of the dancing example is the "self doubt" intepretation. The question "When I am dancing, should I really be dancing for the sake of dancing?" is the agent asking, "Am I doing the right thing right now?" The agent is not asking about the next moment, but is expressing doubt about the moral validity of their behavior at this current moment. Under this interpretation, the agent wants to know if they made the right decision. But what decision has the agent made? Did the agent adopt the maxim "When dancing, dance for the sake of dancing?" They could not have, because such a maxim could not have altered their behavior at all. This agent actually wants to know if the maxim that resulted in them dancing, the maxim that got them to the current moment on the dance floor, was actually the right thing to do. They are asking if, in the past, when they decided to dance, they made the right decision. But such a maxim is not vacuous, because it resulted in the act of

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

dancing. 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 they are currently acting on, not the badly formed example maxim. Evaluating the moral validity of this maxim is straightforward.

This analysis also demonstrates the correct way to think about moral regret and self-doubt. When we ask "should I really be doing this?" we are asking if, when we chose to adopt our current maxim, we made the right decision. The question is not, "When acting should I act?" The question is, "When I was in the past, should I have chosen to act?" Regret is a counterfactual question about a valid, well-formed maxim that was adopted in the past and led to the present moment. Unlike a badly formed maxim, it doesn't undermine the will because it acknowledges the will's authority and the fact that the will made a decision, but it asks if the will was mistaken.

4.2.2 The Value of Computational Ethics

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 working without any computational tools. Instead, I argue that computational 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 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 heavily 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 proofs more accessible. Isabelle can complete a proof, starting from first principles, in a matter of seconds that would take a logician pages to complete. Just as calculators make arithmetic more accessible, computational ethics does the same for formal philosophy. Not all philosophy can or will be 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 for both mathematicians and philosophers.

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 com-

puters 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.[8]

References

[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.

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

[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*.

[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.

[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.

[8] 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?

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