Automated Kantian Ethics

a Senior Thesis presented
by
Lavanya Singh
to
The Departments of Computer Science and Philosophy

in partial fulfillment of the requirements
for the degree of
Bachelor of Arts with Honors

Harvard University
Cambridge, Massachussetts
May 2022

ABSTRACT

AI agents are beginning to make decisions without human supervision in increasingly consequential contexts like healthcare, policing, and driving. These decisions are inevitably ethically tinged, but most AI agents navigating the world today are not explicitly guided by ethics. Regulators, philosophers, and computer scientists are raising the alarm about the dangers of unethical artificial intelligence, from lethal autonomous weapons to criminal sentencing algorithms prejudiced against people of color. These warnings are spurring interest in automated ethics, or the development of machines that can perform ethical reasoning. Prior work in automated ethics rarely engages with philosophical literature, despite its relevance to the development of responsible AI agents. If automated ethics draws on sophisticated philosophical literature, its decisions will be more nuanced, precise, and consistent, but automating ethical theories is difficult in practice. Faithfully translating a complex ethical theory from natural language to the rigid syntax of a computer program is technically and philosophically challenging.

In this thesis, I present an implementation of automated Kantian ethics that is faithful to the Kantian philosophical tradition. Given an appropriately represented action and minimal factual background, my system can judge a potential action as morally obligatory, permissible, or prohibited. To accomplish this, I formalize Kant's categorical imperative, or moral rule, in Carmo and Jones's Dyadic Deontic Logic, implement this formalization in the Isabelle/HOL theorem prover, and develop a testing framework to evaluate how well my implementation coheres with expected properties of Kantian ethics, as established in the literature. I also use my system to derive philosophically sophisticated and nuanced solutions to two central ethical dilemmas in Kantian literature: the permissibility of lying in the context of a joke and to a murderer asking about the location of their intended victim. Finally, I examine the philosophical implications of my system, exploring its potential to guide AI agents, academic philosophers, and everyday reasoners as they navigate ethical dilemmas. Ultimately, this work serves as an early proof-of-concept for philosophically mature AI agents and is one step towards the development of responsible, trustworthy artificial intelligence.

Contents

Acknowledgements

Every step of this process has been made easier by my mentors' unwavering support. Thank you to Professor Nada Amin for taking me seriously when this project was just a vague intuition, for always asking the right questions at the right time, and for being willing to stumble through the dark with me. I am endlessly grateful for our Friday night email threads.

Thank you to Dr. William Cochran for offering much more of time, energy, and support than I ever imagined possible. This project was less lonely because of your willingness to venture down rabbit holes with me. I hope that someday I can be as cool and collected as you (maybe I'll have to read more Aristotle first).

Thank you to Professor James Waldo for always giving me the advice I need to hear. Since freshman year, I have always walked out of your office with a clear head and renewed resolve. Thank you for giving me the knowledge that, even when everything feels hopeless, there is one person at this university who believes in me. Your mentorship has been the single most transformative part of my college experience, and I will always be grateful to you.

I fell in love with philosophy while reading Christine Korsgaard's *Sources of Normativity*, and this thesis is largely inspired by her work. She once said that we make the world more sensible by being good to each other; her ideas have certainly made my world make more sense.

1 Introduction

As AI agents become more sophisticated and less dependent on humans, interest begins to mount in the development of automated moral agents, or computers that can perform ethical reasoning. AI agents are making decisions in increasingly consequential contexts, such as healthcare, driving, and criminal sentencing, and therefore must perform ethical reasoning in order to navigate moral dilemmas. For example, self-driving cars may face less extreme versions of the following moral dilemma: an autonomous vehicle approaching an intersection fails to notice pedestrians in the crosswalk until it is too late to brake. The car can either continue on its course, running over and killing three pedestrians, or it can swerve to hit the car in the next lane, killing the single passenger inside it. While this example is (hopefully) not typical of the operation of a self-driving car, every decision that such an AI agent makes, from avoiding congested freeways to carpooling, is morally tinged. Not only do AI agents routinely make decisions with ethical implications without explicitly performing ethical reasoning, in many cases they do so without human supervision. For example, the Alleghany Family Screening tool can automatically trigger an investigation into a potential case of child neglect, a decision that can uproot entire families and is known to be biased against poor people of color (Eubanks, 2018). This motivates the need for automated ethics (also called machine ethics), or the development of machines that can perform robust, sophisticated ethical reasoning.

Machine ethicists recognize the need for automated ethics and have made both theoretical ((Awad et al., 2020), (Davenport, 2014), (Wallach and Allen, 2008), (Gabriel, 2020)) and practical progress ((Arkoudas et al., 2005), (Cervantes et al., 2013), (Jiang et al., 2021), (Winfield et al., 2014)) towards automating ethics. However, prior work in machine ethics using popular ethical theories like deontology ((Anderson and Anderson, 2014), (Anderson and Anderson, 2008)), consequentialism ((Abel et al., 2016), (Anderson et al., 2004), (Cloos, 2005)), and virtue ethics (Berberich and Diepold, 2018) rarely engages with philosophical literature and thus often misses philosophers' insights. Even the above example of the malfunctioning self-driving car is an instance of Phillipa Foot's trolley problem, in which a bystander watching a runaway trolley can pull a lever to kill one instead of three (Foot, 1967). Decades of philosophical debate have developed ethical theories that can offer nuanced and consistent answers to the trolley problem. Like the trolley problem, the moral dilemmas that artifical agents face are not entirely new, so solutions to these problems should take advantage of philosophical progress. Philosophers are devoted to the creation of better ethical theories, so the more faithful that automated ethics is to philosophical literature, the more nuanced, precise, consistent, and therefore trustworthy it will be.

A lack of engagement with prior philosophical literature also makes automated moral agents less explainable, or interpretable by human observers. One example of this is Delphi, an implementation of automated ethics that uses deep learning to make moral judgements based on a training dataset of ethical decisions made by humans (Jiang et al., 2021). Early versions of Delphi gave unexpected results, such as declaring that the user should commit genocide if it makes everyone happy (Vincent, 2021). Moreover, because no explicit ethical theory underpins Delphi's judgements, human beings cannot analytically determine why Delphi thinks genocide is obligatory or where its reasoning may have gone wrong. Machine learning approaches like Delphi often cannot explain their decisions to a human being, reducing human trust in a machine's controversial ethical judgements. If a machine prescribes killing one person to save three without rigorously justifying this decision, it is difficult to trust this judgement. The high stakes of automated ethics require explainability to build trust and catch mistakes, which motivates philosophically faithful automated ethics.

While automated ethics should draw on philosophical literature, in practice, automating an ethical theory is a technical and philosophical challenge. Intuitive computational approaches explored previously, such as representing ethics as a constraint satisfaction problem (Dennis et al., 2016) or reinforcement learning algorithm (Abel et al., 2016), fail to capture philosophically plausible ethical theories. For example, encoding ethics as a Markov Decision Process assumes that ethical reward can be aggregated according to some discounted sum[1], but many philosophers reject this notion of aggregation (Sinnott-Armstrong, 2021). On the other

[1]Markov Decision Processes usually assume that the total reward of a system is the discounted sum of the reward at each state, given by r_i. Formally, total reward $R = \sum_0^\infty \gamma^i r_i$ for some $\gamma \leq 1$.

hand, approaches that begin with an ethical theory, instead of a computational method, must contend with the fact that ethical theories are almost always described in natural language and must be made precise enough to represent to a computer. Even once ethics is translated from natural language to program syntax, the factual background given to the machine, such as the description of an ethical dilemma, plays a great role in the machine's decisions. Another complication is that philosophers do not agree on a single choice of ethical theory. Even philosophers who subscribe to a specific ethical theory still debate the theory's details.[2] Moreover, even once reasoning within a particular ethical theory is automated, those who disagree with that theory will disagree with the system's judgements.

Contributions

This thesis presents a proof-of-concept implementation of philosophically faithful automated Kantian ethics. I formalize Kant's categorical imperative, or moral rule, as an axiom in Dyadic Deontic Logic (DDL), a modal logic designed to reason about obligation (Carmo and Jones, 2013). I implement my formalization in Isabelle/HOL, an interactive theorem prover that can automatically verify and generate proofs in user-defined logics (Nipkow et al., 2002). Finally, I use Isabelle to automatically prove theorems (such as, "murder is wrong") in my new logic, generating results derived from the categorical imperative. Because my system automates reasoning in a logic that represents Kantian ethics, it automates Kantian ethical reasoning. Once equipped with minimal factual background, it can classify actions as prohibited, permissible or obligatory. I make the following contributions:

1. In Section 2.1, I make a philosophical argument for why Kantian ethics is the most natural of the three major ethical traditions (deontology, virtue ethics, utilitarianism) to formalize.
2. In Section 3.1, I present a formalization of the practical contradiction interpretation of Kant's Formula of Universal Law in Dyadic Deontic Logic. I implement this formalization in the Isabelle/HOL theorem prover. My implementation includes axioms and

[2] I give examples of such debates within Kantian ethics in Sections 3.1.2, 3.1.3, 4.1, and 4.2.

definitions such that my system, when given an appropriately represented input, can prove that the input action is permissible, obligatory, or prohibited. It can also return a list of facts used in the proof and, in some cases, a human readable proof.

3. In Section 3.2, I present a testing framework that can evaluate how faithful an implementation of automated Kantian ethics is, inspired by philosophical literature. This testing framework shows that my formalization substantially improves on prior attempts to formalize Kantian ethics.

4. In Sections 4.1 and 4.2, I demonstrate my system's power and flexibility by using it to produce nuanced answers to two well-known Kantian ethical dilemmas. I show that, because my system draws on definitions of Kantian ethics presented in philosophical literature, it is able to perform sophisticated moral reasoning with minimal factual or situational context.

5. In Section 5.2, I present ethical insights discovered using my system and argue that computational methods like the one presented in this thesis can help philosophers resolve debates about ethics. Not only can my system help machines reason about ethics, but it can also help philosophers make philosophical progress, just as computational tools unlock discoveries in fields like protein folding and drug discovery.

Automated Kantian Ethics

My implementation of automated Kantian ethics formalizes Kant's moral rule in deontic logic, a modal logic that can express obligation, or morally binding requirements. Traditional modal logics include the necessitation operator, denoted as $\Box$. Using the Kripke semantics, $\Box p$ is true at world w if p is true at all worlds that neighbor w (Cresswell and Hughes, 1996). Modal logics also contain the possibility operator $\diamond$, where $\diamond p \longleftrightarrow \neg(\Box(\neg p))$ and operators of propositional logic like $\neg, \wedge, \vee, \rightarrow$. Standard deontic logic (SDL) replaces $\Box$ with the obligation operator O, where $O\, p$ is true at w if p is true at all morally perfect versions of w (McNamara and Van De Putte, 2021). While SDL is appreciable for its simplicity, in situations where duty is violated, the logic breaks down and produces paradoxical results.[3] To avoid such issues, I use Dyadic Deontic Logic, in which the dyadic obligation operator $O\{A|B\}$ represents the sentence "A is obligated in the context B." The introduction of context allows DDL to express more nuanced reasoning and resolve contrary-to-duty paradoxes. DDL is both deontic and modal, so sentences like $O\{A|B\}$ are terms that can be true or false at a world. For example, the sentence $O\{\text{steal}|\text{when rich}\}$ is true at a world if stealing when rich is obligated at that particular world.

I automate Kantian ethics because it is the most natural of the major ethical traditions to formalize, as I argue in Section 2.1. Kant presents three versions of a single moral rule, known as the categorical imperative, from which all moral judgements can be derived. I implement a version of this rule called the Formula of Universal Law (FUL), which states that an act is only ethical if it can be performed by all people without contradiction. For example, falsely promising to repay a loan is wrong because not everyone can falsely promise to repay a loan, since lenders will no longer believe these promises and will stop offering loans. The FUL prohibits actions that are not "universalizable," or cannot be undertaken by everyone. It formalizes the kind of objection drawn by the question, "What if everyone did that?"

Prior work by Benzmüller, Farjami, and Parent (Benzmüller et al., 2019; Benzmüller et al., 2021) implements DDL in Isabelle/HOL, and I add the Formula of Universal Law as an axiom on top of their library. The resulting Isabelle theory can automatically or semi-automatically generate proofs in a new logic that has the categorical imperative as an axiom. Because proofs in this logic are derived from the categorical imperative, they judge actions as obligated, prohibited, or permissible. Moreover, because interactive theorem provers are designed to be interpretable, my system is explainable. Isabelle can list the axioms and facts it

[3]The paradigm case of a contrary-to-duty paradox is the Chisholm paradox. Consider the following statements:

1. It ought to be that Tom helps his neighbors
2. It ought to be that if Tom helps his neighbors, he tells them he is coming
3. If Tom does not help his neighbors, he ought not tell them that he is coming
4. Tom does not help his neighbors

These premises contradict themselves, because items (2)-(4) imply that Tom ought not help his neighbors. The contradiction results because the logic cannot handle violations of duty mixed with conditionals. (Chisholm, 1963; Rönnedal, 2019) uses to generate an ethical judgement, and, in some cases, construct human-readable proofs.

In addition to presenting the above logic and implementation, I also contribute a testing framework that evaluates how well my formalization coheres with philosophical literature. I formalize expected properties of Kantian ethics as sentences in my logic, such as the property that obligations cannot contradict each other. To run the tests, I use Isabelle to automatically find proofs or countermodels for the test statements. For example, my implementation passes the contradictory obligations test because it is able to prove the sentence $\neg(O\{A|B\} \wedge O\{\neg A|B\})$, which says that A and $\neg A$ are not both obligatory. This testing framework shows that my system outperforms a control group (raw DDL without any moral axioms added) and Moshe Kroy's prior attempt at formalizing Kantian ethics in deontic logic (Kroy, 1976).

In Chapter 4, I demonstrate my system's power by using it to arrive at sophisticated solutions to two ethical dilemmas often used in critiques of Kantian ethics. I show that because my system is faithful to philosophical literature, it is able to provide nuanced answers to paradoxes that require a deep understanding of Kantian ethics. While this reasoning does require some factual and situational context, my system derives mature judgements with relatively little and uncontroversial background. This indicates that the challenge of automating "common sense," a major hurdle for automated ethics, is within closer reach than previously thought. I discuss automated common sense further in Sections 5.1 and 5.4.

A machine that can evaluate the moral status of a maxim can not only help machines better reason about ethics, but it can also help philosophers better study philosophy. I argue for "computational ethics," or the use of computational tools to make philosophical progress, analogous to computational biology or neuroscience. I demonstrate the potential of computational ethics by presenting a philosophical insight about which kinds of actions are appropriate for ethical consideration that I discovered using my system. The process of building and interacting with a computer that can reason about ethics helped me, a human philosopher, arrive at a philosophical conclusion that has implications for practical reason and philosophy of doubt. Thus, my system can be used in three distinct ways. First, my system can help automated agents navigate the world, which I will refer to as automated ethics or machine ethics interchangeably. Second, my system help human philosophers reason about philosophy, which I call computational ethics. Third, as I discuss in Section 5.3, computational ethics can help not only professional philosophers, but can also augment the everyday ethical reasoning that we all perform as we navigate the world.

2 System Components

My system consists of three components: an ethical theory (Kantian ethics), a logic in which I formalize this ethical theory (Dyadic Deontic Logic), and an interactive theorem prover in which I implement the formalized ethical theory (Isabelle/HOL). In this section, I describe these components and present the philosophical, logic, and computational background underlying my system.

These three components determine the features and limitations of my implementation of automated ethics. Other choices of components, such as another ethical theory, a different logic, or a different theorem prover could be made. Flaws with these components are limitations of my system, but do not indict logic-programming-based automated ethics more generally. My thesis seeks to both present a specific implementation of automated ethics but also to argue for a particular approach to automating ethical reasoning and these choices are relevant to the former goal but not to the latter.

2.1 Choice to Automate Kantian Ethics

In this thesis, I automate Kantian ethics. In 2006, Powers posited that deontological theories are attractive candidates for automation because rules are generally computationally tractable (Powers, 2006, 1). Intuitively, algorithms are procedures for problem solving and Kantian ethics (which is a kind of deontological theory) offers 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 it prescribes moral rules that require little additional data about the world and are easy to represent to a computer. I argue that, compared to consequentialism and virtue ethics,[4] Kantian ethics is more amenable to automation.

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 some example approaches in each tradition and argue that deontology is more straightforward to formalize than these approaches. Insofar as my project serves as an early proof-of-concept, I choose to automate an ethical theory that poses fewer challenges than others. All ethical traditions have debates that an automated ethical system will need to take a stance on, but these debates are less frequent and controversial for Kantian ethics than for consequentialism and virtue ethics.

I first present consequentialism, then virtue ethics, and finally Kantian ethics. For each tradition, I present a crash course for non-philosophers and then explain some obstacles to automation, arguing that these obstacles are weakest in the case of Kantian ethics.

2.1.1 Consequentialism

A consequentialist ethical theory evaluates an action by evaluating its consequences.[5] For example, utilitarianism is a form of consequentialism in which the moral action is the action that produces the most good (Driver, 2014). This focus on the consequences of action distinguishes consequentialists from deontologists, who derive the moral worth of an action from the action itself. Some debates in the consequentialist tradition include which consequences matter, what constitutes a "good" consequence, and how we can aggregate the consequences of an action over all the individuals involved.

Which Consequences Matter

Because consequentialism evaluates the state of affairs following an action, this kind of

[4] Technically, virtue ethics and consequentialism are broad ethical traditions, while Kantian ethics is a specific ethical theory within deontology, the third major ethical tradition. However, Kantian ethics is not merely a kind of deontology but is widely regarded as deontology's central representative (Alexander and Moore, 2021). Deontology's comparatively greater focus on Kant means that my choice of Kant as a guiding figure is less controversial for deontologists than, for example, the choice of Bentham as the guiding figure of consequentialism. Given that most deontological theories have some connection to or basis in Kantian ethics, I choose to focus on Kantian ethics, instead of deontology more broadly.

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

ethical reasoning requires more knowledge about the state of the world than Kantian ethics. Consequentialism requires knowledge about some or all consequences following an action. This means that an automated consequentialist system must somehow collect a subset of the infinite consquences of following an action, a difficult, if not impossible, task. Moreover, compiling this database of consequences requires determining which consequences were actually caused [6] by an action and characterizing the state of the world before and after an action. 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 this scaling challenge because it evaluates the acts themselves, and acts that affect 1 person and acts that affect 1 million people share the same representation.

The challenge of representing the circumstances of action is not unique to consequentialism, but it is particularly acute in this case. Kantian ethicists robustly debate which circumstances of an action are "morally relevant" when evaluating an action's moral worth.[7] Because deontology merely evaluates a single action, the surface of this debate is much smaller than debates 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, but consequentialism requires far more knowledge about the world than Kantian ethics.

Theory of the Good

An automated consequentialist reasoner must also adopt a specific theory of the good, or account of what qualifies as a "good consequence." For example, hedonists associate good with the presence of pleasure and the absence of pain, while preference utiliarians believe that good is the satisfaction of desire. Other consequentialists, like Moore, adopt a pluralistic theory of value, under which many different kinds of things are good for different reasons

[6] David Hume argues that many straightforward accounts of causation face difficulties (Hume, 2007), and philosophers continue to debate the possiblity of knowing an event's true cause. Kant even argued that first causes, or noumena, are unknowable by human beings (Stang, 2021).

[7] Powers (2006) identifies this as a challenge for automating Kantian ethics and briefly sketches solutions from O'Neill (1990), Silber (1974), and Rawls (1980). For further discussion of morally relevant circumstances, see Sections 3.1.2 and 5.1.

(Moore, 1903).

Most theories of the 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. Evaluating a state of affairs requires many controversial judgements about whether a state of affairs actually satisifes the relevant criteria for goodness. Perfect knowledge of tens of thousands of people's pleasure or preferences or welfare or rights is difficult, if not impossible.[8] Either a human being assigns values to states of affairs, which doesn't scale, or the machine does, which requires massive factual background and increases room for doubting the system's judgements. This may be a tractable problem, but it is more difficult than the equivalent Kantian task of formulating and evaluating an action.

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 the 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. 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? Resolving this debate requires that an 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 an aggregation algorithm, an automated ethical system still needs to calculate a goodness measurement for a state of affairs. Whereas before the system could restrict its analysis to a single person, the algorithm must now evaluate an entire state wholistically. As consequentialists

[8] Even if it were possible, collecting this kind of data poses privacy and surveillance risks.

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.

Prior Attempts to Formalize Consequentialism

Because of its intuitive appeal, computer scientists have tried to formalize consequentialism in the past. These efforts cannot escape the challenges outlined 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 is 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.[9] It leaves open the question of what the reward function is and thus leaves the theory of the good, arguably the defining trait of consequentialism, undefined. Similarly, Anderson and Anderson's proposal of a hedonistic act utilitarian automated reasoner chooses hedonism[10] as the theory of the good (Anderson et al., 2004, 2). 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, a usable automated consequentialist moral agent will need to resolve the above debates.

2.1.2 Virtue Ethics

Virtue ethics places the virtues, or traits that constitute a good moral character and make their possessor good, at the center (Hursthouse and Pettigrove, 2018). For example, Aristotle describes virtues as the traits that enable human flourishing. Just as consequentialists define "good" consequences, virtue ethicists present a list of virtues, such as the Buddhist virtue of equanimity (McRae, 2013). An automated virtue ethical agent will need to commit to a list of virtues, a controversial choice. Virtue ethicists robustly debate which traits qualify as virtues, what each virtue actually means, and what kinds of feelings or attitudes must accompany

[9] Generally, reward for an MDP is accumulated according to a "discount factor" $\gamma \leq 1$, such that if r_i is the reward at time i, the total reward is $\sum_{i=0}^{\infty} \gamma^i r_i$.

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

virtuous action.

Another difficulty with automating virtue ethics is that the unit of evaluation for virtue ethics is often a person's entire moral character. While Kantians evaluate the act itself, virtue ethicists evaluate the actor's moral character and their disposition towards the act. If states of affairs require complex representations, an agent's ethical character and disposition are even more difficult to represent to a computer. This is more than just a data-collecting problem; it is 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 immediately presents a formal rule to implement.

Prior Work in Machine Learning and Virtue Ethics

Many virtue ethical theories involve some notion of moral habit, which seems to be amenable to a machine learning approach. Artistotle, for example, argued that cultivating virtuous action requires making such action habitual (Aristotle, 1951). This seems to point to a machine learning approach to automated ethics, in which ethical behavior is learned from a dataset of acts, where an act is tagged as virtuous if an ideal virtuous agent would perform it.

Just as prior work in consequentialism takes implicit or explicit stances on debates in consequentialist literature, so must work in machine learning-based virtue ethics. For example, the training dataset with acts labelled as virtuous or not virtuous will contain an implicit view on what the virtues are and how certain acts impact an agent's moral character. Because there is no canonical list of all virtues that virtue ethicists accept, this implicit view will likely be controversial. Even virtue ethicists who agree that certain traits, like courage, are virtues debate the exact definitions of these traits.

Machine learning approaches like the Delphi system (Jiang et al., 2021) mentioned in Chapter 1 also may suffer explanability problems that my logic-programming, theorem-prover approach does not face. Many machine learning algorithms cannot sufficiently explain their decisions to a human being and often find patterns in datasets that don't cohere with the causes that a human being would identify (Puiutta and Veith, 2020). While there is significant activity and progress in explainable machine learning, interactive theorem provers are designed to be explainable at the outset. Isabelle can show the axioms and lemmas it used in constructing a proof, allowing a human being to reconstruct the proof independently if they wish. This is not an intractable problem for machine learning approaches to computational ethics, but is one reason to prefer logical approaches.[11]

2.1.3 Kantian Ethics

Kant's theory is centered on 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 explains that rational beings are unique because we act "in accordance with the representations of laws" (Kant, 1785, 26). 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, "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, something 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. We are purposive and our actions are guided by practical reason. We have reasons for acting, even when these reasons are opaque to us. This reflective choosing, or operation of practical reason, is what Kant calls the will.

The will operates by adopting or willing maxims, which are its perceived reasons for acting. Kant defines a maxim as the "subjective principle of willing," or the reason that the will *subjectively* gives to itself for acting (Kant, 1785, 16, footnote 1). Many 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 an 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 given by this maxim. Because a maxim captures an agent's principle of action, Kant evaluates maxims as obligatory, prohibited, or permissible. He argues that

[11] This argument about explanability is in the context of virtue ethics and machine learning. It also applies to a broader class of work in automated ethics that uses a "bottom-up" approach, in which a system learns moral judgements from prior judgements. I will extend this argument to general bottom-up approaches in Section 5.5.

[12] For more discussion of the definition of a maxim, see Section 3.1.2.

the form of certain maxims requires any rational agent to will them, and these maxims are obligatory.

The form of an obligatory maxim is given by the categorical imperative. An imperative is a command, such as "Close the door" or "Eat the doughnut in order to satisfy your sweet tooth." An imperative is categorical if it holds unconditionally for all rational agents in all circumstances. Kant argues that the moral law must be a categorical imperative (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 will itself. No one else can tell you what to do 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.

Armed with this understanding of practical reason, Kant presents the categorical imperative. He presents three "formulations" or versions of the categorical imperative. In this project, I focus on the first formulation, the Formula of Universal Law, and I justify this choice in Section 2.1.4.

The Formula of Universal Law (FUL) states, "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, in which we test the moral worth 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 reasoning that makes this statement true is not specific to any person, but is universal across people. Therefore, if I have sufficient reason to will a maxim, so does every other rational agent. There is nothing special about the operation of my practical reason. 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.

Ease of Automation

Kantian ethics is an especially attractive candidate for formalization because the categorical imperative, particularly the FUL, is a property of reason related to the form or structure of a maxim. It does not require any situational knowledge beyond the circumstances included in the maxim itself and thus requires fewer contingent facts than other ethical theories. While other ethical theories often rely on many facts about the world or the actor, a computer evaluating a maxim doesn't require any knowledge about the world beyond what is contained in a maxim. 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 require far more knowledge about the world or the agent to reach a moral decision.

A maxim itself is an object with a thin representation for a computer, as compared to more complex objects like states of affairs or moral character. In Section 3.1.2, I argue that a maxim can be represented simply as a tuple of circumstances, act, and goal. This representation is simple and efficient, especially when compared to representations of a causal chain or a state of affairs or moral character. This property not only reduces the computational complexity (in terms of time and space) of representing a maxim, but it also makes 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 require.

Difficulties in Automation

One challenge for automating Kantian ethics is the need for "common-sense", or factual and situational background. Common-sense is needed when formulating a maxim and determining if a maxim violates the Formula of Universal Law. Maxims include the circumstances in which they apply and determining which circumstances are "morally relevant" to a maxim requires factual background. My system does not need to answer this question because I assume a properly crafted maxim as input and apply the categorical imperative to this input. Using my system to build a fully automated moral agent will require answering this question, a challenging computational and philosophical task. I discuss this problem in greater detail in Section 3.1.2 and Section 5.1.

Common sense is also relevant when applying the universalizability test itself. Consider the example maxim "When broke, I will falsely promise to repay a loan to get some quick cash." This maxim fails the universalizability test because in a world where everyone falsely promises to repay loans, no one will believe promises anymore, so the maxim will no longer serve its intended purpose (getting some quick cash). Making this judgement requires understanding enough about the system of promising to realize that it breaks down if everyone abuses it in this manner. This is a kind of common sense reasoning that an automated Kantian agent would need. This need is not unique to Kantian ethics; consequentialists agents need common sense to determine the consequences of an action and virtue ethical agents need common sense to determine which virtues an action reflects. For example, in the case of virtue ethics, in order to see that saving a baby from a lion requires courage, a reasoner must have enough background knowledge to know that lions are scary. Making any ethical judgement requires robust conceptions of the action at hand, but Kantian ethics requires far less common sense than a consequentialism or virtue ethics.[13] All moral theories evaluating falsely promising will a robust definition of promising, but consequentialism and virtue ethics will require more information than Kantian ethics. Thus, although the need for common sense poses a challenge to automated Kantian ethics, this challenge is more acute for consequentialism or virtue ethics.

[13]In Sections 4.1 and 4.2, I use my system to demonstrate that Kantian ethics requires relatively lightweight, uncontroversial definitions of concepts like falsely promising.

2.1.4 The Formula of Universal Law

Kant presents three formulations, or versions, of what he calls the "supreme law of morality," but I focus on the first of these three. In this section, I argue that the Formula of Universal Law, specifically, is the easiest part of Kantian ethics to automate and the most generalizable.

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). 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 describes 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 the FUL,[14] 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 precision reduces ambiguity, making it easier to remain faithful to philosophical literature on Kant. Ambiguity in an ethical theory, such as the ambiguity about what counts as a good consequence in the case of consequentialism, forces an implementation of automated ethics to take stances on controversial philosophical debates. Minimizing the need for such choices puts my implementation on solid philosophical footing.

[14] The FUL is often seen as emblematic of Kantian constructivism (Ebels-Duggan, 2012, 173). My project is not committed to Kantian constructivism.

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.

Those who do not believe that all three formulations of the categorical imperative are equivalent understand the FUL as the strongest or most foundational, and thus an appropriate initial choice for automation. Korsgaard characterizes the three formulations of the categorical imperative according to Rawls' general and special conception of justice. The general conception applies universally and can never be violated, while the special conception represents an ideal for us to live towards that may not be possible to achieve. For example, the special conception may require that we prefer some job applicants over others in order to remedy historical injustice, and the general conception may require that such inequalities always operate in the service of the least privileged (Korsgaard, 1986, 19). Korsgaard argues that the Formula of Universal Law represents Kant's general conception of justice, and the Formula of Humanity represents his special conception. The FUL's prescriptions can never be violated, even in the most non-ideal circumstances imaginable, but the FUH is merely an ideal to strive towards. Thus, the FUL generates stronger requirements than the other two formulations and reflects the bare minimum standard of Kant's ethics. Because the FUL's prescriptions outweigh those of the other two formulations, it serves as a functional, minimal version of Kantian ethics.

2.2 Dyadic Deontic Logic

I formalize Kantian ethics by representing it as an axiom on top of a base logic. In this section, I present the logical background necessary to understand my work and my choice of Dyadic Deontic Logic (DDL).

As explained in Chapter 1, traditional modal logics include the necessitation operator, denoted as $\Box$. In simple modal logic using the Kripke semantics, $\Box p$ is true at a world w if p is true at all of w's neighbors, and it represents the concept of necessary truth (Cresswell and Hughes, 1996). These logics usually also contain the possibility operator $\diamond$, where $\diamond p \iff \sim \Box \sim p$. $\diamond p$ means that the statement p is possibly true, or true at at least one of w's neighbors. Additionally, modal logics include standard operators of propositional logic like $\sim, \wedge, \vee, \rightarrow$.

A deontic logic is a special kind of modal logic designed to reason about moral obligation. Standard deontic logic replaces $\Box$ with the obligation operator O, and $\diamond$ with the permissibility operator P (Cresswell and Hughes, 1996). Using the Kripke semantics for O, Op is true at w if p is true at all ideal deontic alternatives to w, and thus represents the concept of moral necessity or necessary requirements. The O operator in SDL takes a single argument (the formula that is obligatory), and is thus called a monadic deontic operator.

While SDL is appreciable for its simplicity, it suffers a variety of well-documented paradoxes, including contrary-to-duty paradoxes. [15] In situations where duty is violated, the logic breaks down and produces paradoxical results. Thus, I use an improved deontic logic instead of SDL for this work.

I use as my base logic Carmo and Jones's Dyadic Deontic Logic (DDL), which improves on SDL (Carmo and Jones, 2013). It introduces a dyadic obligation operator $O\{A|B\}$ to represent the sentence "A is obligated in the context B." The introduction of context allows DDL to gracefully handle contrary-to-duty conditionals, since violations of duty simply modify the context. The obligation operator uses a neighborhood semantics, instead of the Kripke semantics (Scott, 1970; Montague, 1970). While Kripke semantics requires that an obligated proposition hold at all worlds, the neighborhood semantics defines a different set of neigh-

[15]The paradigm case of a contrary-to-duty paradox is the Chisholm paradox. Consider the following statements:

1. It ought to be that Tom helps his neighbors
2. It ought to be that if Tom helps his neighbors, he tells them he is coming
3. If Tom does not help his neighbors, he ought not tell them that he is coming
4. Tom does not help his neighbors

These premises contradict themselves, because items (2)-(4) imply that Tom ought not help his neighbors. The contradiction results because the logic cannot handle violations of duty mixed with conditionals. (Chisholm, 1963; Rönnedal, 2019) bors, or morally relevant alternatives, for each world. To represent this, Carmo and Jones define a function ob that maps a given context (or world) to the propositions that are obligatory at this world, where a proposition p is defined as the worlds at which the p is true. DDL is thus both modal and deontic; statements about obligations are true or false at a world according to the neighbordhood semantics, and different obligations may hold at different worlds. This property is particularly relevant to my work because the universalizability test requires reasoning about alternative worlds, such as the world of the universalized maxim.

DDL also includes modal operators. In addition to $\Box$ and $\diamond$, DDL also has a notion of actual obligation and possible obligation, represented by operators O_a and O_p respectively. These notions are accompanied by the corresponding modal operators $\Box_a, \diamond_a, \Box_p, \diamond_p$. These operators use a Kripke semantics, with the functions av and pv mapping a world w to the set of corresponding actual or possible versions of w. These operators are not relevant to the work in this thesis, but this additional expressivity could be used to extend my project to incorporate more sophisticated ethical concepts like counterfactuals.

For more of fine-grained properties of DDL see Carmo and Jones (2013) or this project's source code.[16] DDL is a heavy logic and contains modal operators that aren't necessary for my analysis. While this expressivity is powerful, it may also cause performance issues. DDL has a large set of axioms involving quantification over complex higher-order logical expressions. Proofs involving these axioms will be computationally expensive. I do not run into performance issues in my system, but future work may choose to embed a less complicated logic.

2.3 Isabelle/HOL

The final component of my project is the automated theorem prover I use to automate my formalization. Isabelle/HOL is an interactive proof assistant built on Haskell and Scala (Nipkow et al., 2002). It allows the user to define types, functions, definitions, and axiom systems. It has built-in support for both automatic and interactive/manual theorem proving.

[16]The project's source code can be found here.

To demonstrate the power and usage of Isabelle and make DDL more precise, I walk through Benzmueller, Farjami, and Parent's implementation of DDL in Isabelle/HOL, which serves as the basis of my formalization of Kantian Ethics.

2.3.1 System Definition

The first step in embedding a logic in Isabelle is defining the relevant terms and types. Commands to do this include `typedecl`, which declares a new type, `type_synonym`, which defines an abbreviation for a complex type, and `consts`, which defines constants.

typedecl *i* — i is the type for a set of worlds.

— This is an Isabelle comment, while the text above is a line of actual, executable Isabelle code.

type-synonym $t = (i \Rightarrow \mathit{bool})$ — t represents a set of DDL formulas.

— A set of formulas is defined by its truth value at a set of worlds. For example, the set $\{\mathit{True}\}$ is true at any set of worlds.

The *ob* function described in Section 2.2 is used to determine which propositions are obligatory in which contexts. I implement it as a constant. This constant has no meaning (I merely specify the type), but future proofs will specify models for this constant.

consts $\mathit{ob}{::}t \Rightarrow (t \Rightarrow \mathit{bool})$ — set of propositions obligatory in this context

— $\mathit{ob}\ (context)\ (term)$ is *True* if the term is obligatory in this context

In a semantic embedding like this one, axioms are modelled as restrictions on models of the system. In this case, a model is specified by the relevant accessibility relations (such as *ob*), so it suffices to place conditions on the accessibility relations. Isabelle allows users to create new axiomatizations on top of its base logic (HOL) and use these axioms in proofs. Here's an example of an axiom:

axiomatization where

$\mathit{ax\text{-}5d}: \forall X\, Y Z.\ ((\forall w.\ Y(w) \longrightarrow X(w)) \wedge \mathit{ob}(X)(Y) \wedge (\forall w.\ X(w) \longrightarrow Z(w)))$
$\longrightarrow \mathit{ob}(Z)(\lambda w.(Z(w) \wedge \neg X(w)) \vee Y(w))$

— If some subset Y of X is obligatory in the context X, then in a larger context Z, any obligatory proposition must either be in Y or in $Z \setminus X$. Intuitively, expanding the context can't cause something unobligatory to become obligatory, so the obligation operator is monotonically increasing with respect to changing contexts.

2.3.2 Syntax

The axiomatization above defines the semantics of DDL and, as demonstrated by the example axiom, is unwieldly. In my work, I mostly perform syntactic proofs, so I must define the syntax of the logic. Isabelle already knows the semantics of the axioms of this logic, so I can define the syntax as abbreviations involving the axioms above. Each DDL operator is represented as a HOL formula. Isabelle automatically unfolds formulas defined with the `abbreviation` command whenever they are applied. While the shallow embedding is performant (because it uses Isabelle's original syntax tree), my heavy use of abbreviations may impact the performance of long proofs.

Modal operators, implemented below, will be useful for my purposes.

abbreviation *ddlbox*$::t \Rightarrow t$ $(\square)$

where $\square A \equiv \lambda w. \forall y. A(y)$

— Notice that the necessity operator is an abbreviation, or syntactic sugar for, the higher order logic formula that the proposition holds at all worlds.

abbreviation *ddldiamond*$::t \Rightarrow t$ $(\Diamond)$

where $\Diamond A \equiv \neg(\square(\neg A))$

— Possibility is similarly an abbreviation for a higher order logic formula involving the defined semantics.

The most important operator for my project is the obligation operator, implemented below.

abbreviation *ddlob*$::t \Rightarrow t \Rightarrow t$ $(O\{\text{-}|\text{-}\})$

where $O\{B|A\} \equiv \lambda\, w.\, ob(A)(B)$

— O$\{B|A\}$ can be read as "B is obligatory in the context A"

While DDL is powerful because of its support for a dyadic obligation operator, in many cases, I only need a monadic obligation operator. Below is some syntactic sugar for a monadic obligation operator.

abbreviation *ddltrue*$::t$ $(\top)$

where $\top \equiv \lambda w.\ True$

abbreviation *ddlfalse*::t $(\bot)$

where $\bot \equiv \lambda w.\ False$

abbreviation *ddlob-normal*::$t \Rightarrow t$ $(O\ \{\text{-}\})$

where $(O\ \{A\}) \equiv (O\{A|\top\})$

— Intuitively, the context $True$ is the widest context possible because $True$ holds at all worlds. Therefore, the monadic obligation operator requires that A is obligated at all worlds.

Finally, validity will be useful when discussing metalogical/ethical properties.

abbreviation *ddlvalid*::$t \Rightarrow bool$ $(\models\text{-})$

where $\models A \equiv \forall\, w.\ A\ w$

— A proposition is valid if it is true at all worlds.

Benemueller, Farjami, and Parent provide a proof of the completeness of the above embedding (Benzmüller et al., 2021). Isabelle allows us to check consistency immediately using Nitpick, a model checker (Blanchette and Nipkow, 2010). Nitpick can find satisfying models for a particular lemma using the `satisfy` option and it can find counterexamples using the `falsify` option, both of which I use heavily in this project.

lemma *True* **nitpick** [*satisfy,user-axioms,format=2*] **by** *simp*

— This an example of a typical Nitpick output. In this case, Nitpick successfully found a model satisfying these axioms so the system is consistent.

— Nitpick found a model for card i = 1:

Empty assignment

In the proof above, "by simp" indicates the use of the Simplification proof method, which unfolds definitions and applies theorems directly. HOL has $True$ as a theorem, which is why this theorem was so easy to prove.

3 Implementation Details

In this section, I present the details of my implementation of automated Kantian ethics, which consists of a formalization of the FUL in Dyadic Deontic Logic and an implementation of this logic in Isabelle/HOL. The final Isabelle library is a logic that has the categorical imperative as an axiom and can express and derive moral judgements. Using Isabelle's automated theorem proving abilities, my system can show that appropriately represented actions are obligatory, permissible, or prohibited by proving or refuting sentences of the form "A is obligated to do B." I also present a testing framework to evaluate how faithful my implementation is to philosophical literature. This testing framework shows that my system outperforms unmodified DDL (a control group) and Moshe Kroy's prior formalization of the FUL (Kroy, 1976).

3.1 Formalization and Implementation of the FUL

Formalizing the FUL requires implementing enough logical background to represent the FUL as an axiom. Dyadic Deontic Logic can express obligation and prohibition, but it cannot represent more complex features of moral judgement like actions, subject, maxims, and ends. I augment DDL by adding representations of these concepts, drawn from philosophical literature.

3.1.1 Logical Background

Kantian ethics is action-guiding; the categorical imperative is a moral rule that agents can use to decide between potential actions. Thus, before I begin to formalize a specific formulation of the categorical imperative, I must define subjects and actions. I add representations of subjects and actions so that my new logic can express sentences of the form, "x does action."

typedecl s — The new type s is the type for a "subject," as in the subject of a sentence.

The `typedecl` keyword indicates that I am defining a new atomic type, which is not composed of pre-existing types but is instead a new kind of object altogether. A type does not come with any properties out of the box. There is no difference between declaring a type with label "subject" or any other label, such as "color" or "mammal." I add some properties of this type below by creating formulas and more complex types that use this type, but I do not provide a complete definition of a subject. Formalizing and using the FUL does not require many of the complex properties of a subject, such as rationality or humanity. Thus, instead of providing a complete definition of subject, I can avoid murky philosophical debates about the nature of agency and instead provide a "thin" definition that only includes the minimum necessary properties to apply the FUL. Throughout my project, I will use bare syntactic units like types and constants to create thin definitions of new ideas. This strategy lets me avoid messy philosophical controversies and makes my system's judgements more trustworthy, because they rely on relatively little prior knowledge.

In this interpretation, the defining feature of a subject is that they can act. I represent that below by allowing subjects to substitute into sentences, a property that I will use to represent the idea that different people can perform the same actions.

type-synonym $os = (s \Rightarrow t)$

— To model the idea of a subject being substituted into an action, I define `type_synonym` os for an open sentence. An open sentence takes as input a subject and returns a complete or "closed" DDL formula by binding the free variable in the sentence to the input. For example, "runs" is an open sentence that can be instantiated with subject, "Sara" to create the DDL term "Sara runs," which can be true or false at a world. An open sentence itself is not truth-apt, ot the kind of thing that can be true or false at a world. When an action is substituted into an open sentence, the resulting term is truth apt. "Runs" is not the kind of thing that can be true or false, but "Sara runs" is a sentence that can be true or false.

3.1.2 Maxims

As established in Section 2.1.4, I formalize a version of the categorical imperative called the Formula of Universal, which reads "act only according to that maxim by which you can at the same time will that it should become a universal law" (Kant, 1785, 34). In order to faithfully formalize the FUL, I must make precise what it means to will a maxim and what kinds of maxims can become universal laws. I draw on reliable definitions of willing, maxims, and universalization from Kantian literature and represent them in DDL. Throughout this section, I will use one of Kant's canonical maxims as an example.

Example 1 (False Promising)**.** *The false promising example maxim reads, "When I am strapped for cash, I will falsely promise to repay a loan to get some easy cash."*

The central unit of evaluation for Kantian ethics is a "maxim," which Kant defines as "the subjective principle of willing," or the principle that the agent understands themselves as acting on (Kant, 1785, 16, footnote 1). Modern Kantians differ in their interpretations of this definition. I adopt O'Neill's view, derived from Kant's example maxims, that a maxim includes the act, the circumstances, and the agent's purpose of acting or goal (O'Neill, 2013). Other potential views include Korsgaard's view, which omits the circumstances, and Kitcher's view, which adds a motivation (Korsgaard, 2005; Kitcher, 2004). I address the limitations of these approaches in Appendix A.

Definition 1 (Maxim)**.** *A maxim is a circumstance, act, goal tuple (C, A, G), read as "In circumstances C, act A for goal G."*

I implement this definition in Isabelle by defining the `type_synonym` below for the type of a maxim.

type-synonym $maxim = (t * os * t)$

— A maxim is of type term, open sentence, term tuple, such as "(When I am strapped for cash, will falsely promise to repay a loan, to get some easy cash)". The first term represents the circumstance, which can be true or false at a world. For example, the circumstance "when I am strapped for cash" is true at the real world when my bank account is empty. The second term represents the action, which is an open sentence because different agents can perform a particular action. For example, the action, "will falsely promise to repay a loan" is an open sentence that can be acted on by a subject. The third term represents the goal, which can again be true or false at a world. For example, the goal "to get some easy cash" is true at the real world if I have successfully received easy cash.

O'Neill argues that a maxim is an action-guiding rule and thus naturally includes an act and the circumstances under which it should be performed (O'Neill, 2013, 37). She also includes a purpose, end, or goal in the maxim because human activity is guided by a rational will and is thus inherently purposive (Kant, 1785, 40). A rational will does not act randomly (else it would not be rational), but instead in the pursuit of ends which it deems valuable. The inclusion a maxim's end is essential for the version of the FUL that I will implement, explained in Section 3.1.3.

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 gives rise to "the tailoring objection," under which maxims are arbitrarily specified to pass the FUL (Kitcher, 2003, 217). [17] For example, the maxim "When my name is Jane Doe and I am wearing a purple shirt and it is Tuesday morning, I will murder my boss so I can take their job," is universalizable but is clearly a false positive because we think that murder for professional gain is wrong. One solution to this problem is to argue that the circumstance "When my name is Jane Doe and I am wearing a purple shirt and it is Tuesday morning" is not morally relevant to the act and goal. This solution requires determining 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 an 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 difficulty in determining relevant circumstances is a limitation of my system and requires that a human being formulate the maxim or that future work develop heuristics to classify circumstances as morally relevant. I discuss this challenge and potential solutions in greater detail in Section

[17] Kitcher cites Wood (1999) as offering an example of a false positive due to this objection.

5.1.

With this robust representation of a maxim, I can now define willing. To will a maxim is to adopt it as a principle to live by, or to commit oneself to the maxim's action for the sake of maxim's end in the relevant circumstances. I formalize this idea in Definition 2.

Definition 2 (Willing)**.** *For maxim* $M = (C, A, G)$ *and actor s,*

$$\text{will}\, M\; s \equiv \forall w\, (C \longrightarrow A\,(s))\, w$$

At all worlds w, if the circumstances hold at that world, agent s performs act A.

If I will the example False Promising maxim, then whenever I need cash, I will falsely promise to repay a loan. I can represent this definition using the following Isabelle formula.

abbreviation *will* :: *maxim* ⇒ *s*⇒ *t* (*W* - -)
where *will* ≡ $\lambda(c, a, g)\; s.\; (c \rightarrow (a\; s))$

— An agent s wills a maxim if in the circumstances, s performs the action, or s substituted into the open sentence a is true. This is an Isabelle `abbreviation`, which is syntactic sugar for an Isabelle formula. The type of this formula is $maxim \rightarrow s \rightarrow t$, so it takes as input a maxim and a subject and returns the term, "s wills maxim."

3.1.3 Practical Contradiction Interpretation of the FUL

In order to evaluate the moral status of a maxim, I must define what it means for a maxim to not be universalizable, or to fail the universalizability test. Kantians debate the correct interpretation of the Formula of Universal Law because Kant himself appears to interpret the criterion in different ways. I adopt Korsgaard's practical contradiction interpretation, broadly accepted as correct within the philosophical community (Ebels-Duggan, 2012).

Recall that the Formula of Universal Law is to "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant, 1785, 34). To determine if a maxim can be willed as a universal law, we use the "universalizability test," which requires imagining a world in which everyone has willed the maxim. If willing the maxim in such a world generates a contradiction, then the action is prohibited. For example, the False Promising maxim will be prohibited if it is impossible to will the maxim in a world where everyone falsely promises to repay loans.

One interpretation of the FUL, the logical contradiction interpretation, prohibits maxims that are logically impossible when universalized. Under this view, falsely promising to repay a loan fails the universalizability test because, in the universalized world, everyone falsely promises to repay loans so lenders no longer believe promises to repay loans. The practice of giving loans would die out, so making a false promise to repay a loan would be impossible.

This view cannot correctly handle natural acts. Korsgaard appeals to Dietrichson (1964) to construct the example natural act of a mother killing her children that cry too much at night so that she can get some sleep. Universalizing this maxim does not generate a logical contradiction because killing is still possible in a world where everyone kills noisy children, but it is clearly wrong. Because killing is a natural act, it can never be logically impossible so the logical contradiction view cannot prohibit it.

As an alternative to the logical contradiction view, Korsgaard endorses the practical contradiction view, which prohibits maxims that are self-defeating, or ineffective, when universalized. By willing a maxim, you commit yourself to the maxim's goal, and thus cannot rationally will that this goal be undercut. This interpretation can prohibit natural acts like those of the sleep-deprived mother: in willing the end of sleeping, she is implicitly willing that she is alive. If all mothers kill all loud children, then she cannot be secure in the possession of her life, because her own mother may have killed her as an infant. Her willing this maxim thwarts the end that she sought to secure.

The practical contradiction interpretation offers a satisfying explanation of *why* certain maxims are immoral. These maxims involve parasitic behavior on social conditions that the agent seeks to benefit from. The false promiser wants to both abuse the system of promising and benefit from it, and is thus making an exception of themselves. The test formalizes the kinds of objections that the question "What if everyone did that?" seeks to draw out.[18]

Under the practical contradiction interpretation, the FUL states, "If, when universal-

[18] This argument for the practical contradiction interpretation is due to Korsgaard (1985).

ized, a maxim is not effective, then it is prohibited." This requires defining effectiveness and universalization. If an agent wills an effective maxim, then the maxim's goal is achieved, and if the agent does not will it, then the goal is not achieved.

Definition 3 (Effective Maxim)**.** *For a maxim $M = (C, A, G)$ and actor s,*

$$\text{effective } M\ s \equiv \forall w\ (\text{will}\ (C, A, G)\ s \iff G)\ w$$

I implement this in Isabelle below.

abbreviation *effective* :: *maxim*⇒*s*⇒ *t* (*E* - -)
where *effective* $\equiv \lambda(c, a, g)\ s.\ ((will\ (c, a, g)\ s) \equiv g)$
— A maxim is effective for a subject if the goal is achieved if and only if the subject wills the maxim. Once again, I use an abbreviation to conveniently refer to this Isabelle formula.

The former direction of the implication is intuitive: if the act results in the goal, it was effective in causing the goal. This represents sufficient causality. The latter direction represents necessary causality, or the idea that, counterfactually, if the maxim were not willed, then the goal would not be achieved (Lewis, 1973).[19] Combining these ideas, this definition of effective states that a maxim is effective if the maxim being acted on by a subject is the necessary and sufficient cause of the goal.

Next, I define what it means for a maxim to be universalized. Recall that the universalizability test requires imagining a world in which everyone wills a maxim. Therefore, a maxim is universalized when everyone wills the maxim.

Definition 4 (Universalized)**.** *For a maxim M and agent s,*

$$\text{universalized } M \equiv \forall w\ (\forall p\ \text{will } M\ p)$$

I can once again represent this as an abbreviation in Isabelle.

abbreviation *universalized*::*maxim*⇒*t* **where**
universalized $\equiv \lambda M.\ (\lambda w.\ (\forall p.\ (W\ M\ p)\ w))$

[19]Thank you to Jeremy Zucker for helping me think about causality in this way.

— The abbreviation $universalized$ takes a maxim as input and returns a term which is true at a world if all people at that world will the maxim.

With the above definitions of effective and universalization, I can define what it means for a maxim to not be universalizable. This is the core of the FUL, which states that if a maxim is not universalizable, it is prohibited. Under the practical contradiction interpretation, a maxim is not universalizable if, when universalized, it is no longer effective.

Definition 5 (Not Universalizable)**.** *For a maxim M and agent s,*

$$\text{not\_universalizable}\, M\ s \equiv [\text{universalized}\, M \longrightarrow \neg\, \text{effective}\, M\ s]$$

A maxim is not universalizable when, if everyone wills the maxim, then it is no longer effective. I implement this definition in Isabelle using another abbreviaion.

abbreviation *not-universalizable* :: *maxim*⇒*s*⇒*bool* **where**
not-universalizable $\equiv \lambda M\, s.\ \forall\, w.\ ((universalized\ M) \rightarrow (\neg\ (E\ M\ s)))\ w$
— Maxim M is not universalizable at world w when, "at world w, if M is universalized, then M is not effective."

The FUL states that if a maxim is not universalizable, then it is prohibited. Before performing moral reasoning with my system, I must define obligation, permissibility, and prohibition. To judge an action, my system evaluates the moral status of the action "person s wills maxim M." This action can be obligated, prohibited, or permissible. I will use the phrase "subject s willing maxim M is obligatory" interchangeably with "maxim M is obligatory for subject s." I will use "maxim M is obligatory" to refer to M being obligatory for any arbitrary subject, which is equivalent to M being obligatory for a specific subject.[20]

Definition 6 (Obligation)**.** *Let maxim M be composed of the circumstances, act, goal tuple C, A, G and let s be an arbitrary agent.*

$$\text{obligated}\, M\ s \equiv O\{\text{will}\,(C, A, G)\ s \mid C\}$$

[20]The full proof for this result is the Obligation Universalizes Across People Test in Section 3.2.

The action "s wills M" is the first argument passed to the dyadic obligation operator and is thus the action that is shown to be obligated or not. The second argument passed to the obligation operator represents the context in which the obligation holds and is thus naturally the maxim's circumstances. This definition does not require any additional syntactic sugar, since it merely uses the dyadic obligation operator. Using this definition, I can define prohibition and permissibility.

Definition 7 (Prohibition and Permissibility)**.** *Let maxim M be composed of the circumstances, act, goal tuple C, A, G and let s be an arbitrary agent.*[21]

$$\text{prohibited}\, M\ s \approx \text{obligated}\, \neg M \ \equiv O\{\neg \text{will}\,(C, A, G)\ s \mid C\}$$

$$\text{permissible}\, M\ s \equiv \neg \text{prohibited}\, M\ s \equiv \neg O\{\neg \text{will}\,(C, A, G)\ s \mid C\}$$

abbreviation *prohibited*::*maxim*⇒*s*⇒*t* **where**
prohibited ≡ λ(*c*, *a*, *g*) *s*. *O*{¬ (*will* (*c*,*a*, *g*) *s*) | *c*}
— A maxim is prohibited for a subject s if its negation is obligated for s. It is morally wrong for an agent to will a prohibited maxim.
abbreviation *permissible*::*maxim*⇒*s*⇒*t*
where *permissible* ≡ λ*M s*. ¬ (*prohibited M s*)
— A maxim is permissible for a subject s if it is not prohibited for s. It is morally acceptable for an agent to will or not will a permissible maxim.

One additional piece of logical background necessary before I implement the FUL is the notion of contradictory obligations. Many deontic logics, including DDL, allow contradictory obligations. As I will explain in Section 3.2, Kantian ethics never prescribes contradictory obligations, so I will add this as an axiom.

abbreviation *non-contradictory* **where**
non-contradictory A B c w ≡ ((*O*{*A*|*c*} ∧ *O*{*B*|*c*}) *w*) ⟶ ¬((*A* ∧ (*B* ∧ *c*)) *w* ⟶ *False*)

[21]Technically, a maxim is not a boolean type, so the term $\neg M$ is not type correct. The expression $obligated\, \neg M$ merely provides intuition for the meaning of prohibition, but the exact definition is given by $O\{\neg will\,(C, A, G)\ s \mid C\}$.

— Terms A and B are non contradictory in circumstances c if, when A and B are obligated in circumstances c, the conjunction of A, B, and c, does not imply $False$.

axiomatization where *no-contradictions*:$\forall\, A{::}t.\ \forall\, B{::}t.\ \forall\, c{::}t.\ \forall\, w{::}i.$ *non-contradictory A B c w*

— This axiom formalizes the idea that, for any terms A, B, and circumstances c, A, and B must be non-contradictory in circumstances c at all worlds. Intuitively, this axiom requires that obligations do not conflict.

3.1.4 Formalizing the FUL

With this logical background, I can begin to implement the Formula of Universal Law, which, as defined by the practical contradiction interpretation, states that a maxim is prohibited if it is ineffective when universalized. A first, unsuccessful attempt to formalize the FUL simply translates this into Isabelle's syntax using the abbreviations above. While this attempt will not be consistent, I will use Isabelle's automatic proof search abilities to determine how to modify this formula to be consistent, revealing a key philosophical insight about maxims in the process. This section presents my final formalization of the FUL and the philosophical insight produced while creating it, which, as I argue in Section 5.2, demonstrates the power of computational tools in aiding philosophical progress.

abbreviation *FUL0*::*bool* **where** *FUL0* $\equiv \forall$ *c a g s*. *not-universalizable* (c, a, g) $s \longrightarrow \models((prohibited$ (c, a, g) $s))$

— This representation of the Formula of Universal Law reads, "For all circumstances, goals, acts, and subjects, if the maxim of the subject performing the act for the goal in the circumstances is not universalizable (as defined above), then, at all worlds, in those circumstances, the subject is prohibited from (obligated not to) willing the maxim."

I can immediately determine if this version of the FUL is consistent by checking if `FUL0` implies $False$.

lemma *FUL0* $\longrightarrow$ *False* **using** *O-diamond*
 using *case-prod-conv no-contradictions old.prod.case old.prod.case* **by** *fastforce*

Isabelle's proof-finding tool, Sledgehammer, shows that `FUL0` is not consistent by showing that it implies a contradiction using axiom `O_diamond`[22] (Paulson and Blanchette, 2015). This axiom roughly states that an obligation can't contradict its context. Knowing that `FUL0` contradicts this particular axiom offers insight into what the problem is. If the goal or action or a maxim are equivalent to its circumstance, then prohibiting it is contradictory. If the maxim has already been acted on or the goal has already been achieved, then the agent cannot undo their action or the achivement of the goal.

This motivates the exclusion of what I call "badly formed maxims," which are those maxims such that the goal has already been achieved or the act has already been acted on. Under my formalization, such maxims are not well-formed.

Definition 8 (Well-Formed Maxim)**.** *A maxim is well-formed if the circumstances do not contain the act and goal. For a maxim* (C, A, G)*, and subject* s*,*

$$\text{well\_formed}\,(C, A, G)\; s \equiv \forall w\, (\neg(C \longrightarrow G) \wedge \neg(C \longrightarrow A\, s))\; w$$

For example, the maxim "When I eat breakfast, I will eat breakfast to eat breakfast" is badly-formed because the circumstance "when I eat breakfast" contains the act and goal. Well-formedness is not discussed in the literature, but I find that if I require that the FUL only holds for well-formed maxim, it is consistent.

abbreviation *well-formed*::*maxim*⇒*s*⇒*i*⇒*bool* **where**
well-formed ≡ $\lambda(c, a, g).\ \lambda s.\ \lambda w.\ (\neg\ (c \rightarrow g)\ w) \wedge (\neg\ (c \rightarrow a\ s)\ w)$
— This abbreviation formalizes the well-formedness of a maxim for a subject. The goal cannot be already achieved in the circumstances and the subject cannot have already performed the act.

If I modify `FUL0` to only hold for well-formed maxims, it becomes consistent.

abbreviation *FUL* **where**
FUL ≡ $\forall\, M$::*maxim*. $\forall\, s$::*s*. ($\forall\, w$. *well-formed* $M\ s\ w$) ⟶ (*not-universalizable* $M\ s$ ⟶ ⊨ (*prohibited* $M\ s$))
— This formalization states that if a maxim is well-formed, then if it is not universalizable, it is prohibited.

[22] The full axiom reads ⊨λw. *ob ?B ?A* ⟶ ¬ ⊨¬ *?B*∧*?A*.

lemma *FUL*

nitpick[*user-axioms*, *falsify=false*] **oops**

— Nitpick is Isabelle's countermodel checker, and I can use it to quickly check that an axiom is consistent (Blanchette and Nipkow, 2010). If Nitpick can find a model in which the axioms of DDL hold and the FUL is true, then it is consistent.

Nitpick found a model for card i = 1 and card s = 2:

Empty assignment

My above investigation of `FUL0` shows that if the FUL holds for badly formed maxims, then it is inconsistent. This is not only a logical property of my system, but it also has philosophical significance that coheres with Korsgaard's and O'Neill's interpretations of a maxim as a practical guide to action (Korsgaard, 2005; O'Neill, 2013). A maxim is a practical principle that guides how we behave in everyday life. A principle of the form "When you are eating breakfast, eat breakfast in order to eat breakfast," is not practically relevant. No agent would ever need to act on such a principle. Morality helps agents decide whether to act on a potential principle of action, but no agent would need to ask "When I am eating breakfast, should I eat breakfast in order to eat breakfast?" It is the wrong kind of principle to be evaluating. It is not a well-formed maxim, so the categorical imperative cannot apply to it.

The fact that Isabelle revealed a philosophical insight about which kinds of maxims are well-formed is an example of the power of computational tools to aid philosophical progress. Nitpick and Sledgehammer helped me confirm that certain kinds of circumstance, act, goal tuples are too badly formed for the categorical imperative to logically apply to them. The realization of this subtle problem would have been incredibly difficult without computational tools, and serves as evidence of the power of computational ethics. I discuss the philosophical properties and implications of well-formed maxims and the power of computational ethics further in Section 5.2.

I complete my implementation by adding the consistent version of the FUL as an axiom.

axiomatization where *FUL*:*FUL*

This concludes my implementation of the Formula of Universal Law in Isabelle/HOL.

My implementation consistents of necessary logical background, first formalized in DDL and then implemented in Isabelle. The code snippets in this chapter are a subset of over 100 lines of Isabelle/HOL code necessary to complete this implementation. In Section 3.2 and Chapter 4, I demonstrate how this implementation can be tested and used to make moral judgements.

3.2 Tests

In addition to an implementation of automated Kantian ethics, I also contribute a testing framework to evaluate how well my implementation coheres with philosophical literature. This testing architecture makes the notion of "philosophical faithfulness" precise. Each test consists of a sentence in my logic and an expected outcome, where the possible outcomes are proving or refuting the sentence. For example, one such sentence is that obligations cannot contradict each other. To run the tests, I attempt to prove or refute each test sentence in my logic. Because these tests are derived from moral intuition and philosophical literature, they evaluate how well my system reflects philosophical literature. Running the tests on my implementation consisted of approximately 400 lines of Isabelle code.

The test framework can be expanded by adding more test sentences and can guide implementations of other parts of Kantian ethics or other ethical theories. As I was implementing my formalization, I checked it against the testing framework, performing test-driven development for automated ethics.

I use my testing framework to show that my formalization and implementation of Kantian ethics outperforms two other potential implementations. First, I consider raw DDL, which serves as a control group because it simply contains the base logic on top of which I build other implementations. DDL can express obligation, but has no knowledge of any specific moral rules (like the categorical imperative). Second, I consider Moshe Kroy's 1976 formalization of the FUL and the second formulation of the categorical imperative (Kroy, 1976). His formalization is based on Hintikka's deontic logic, which is a different, less expressive logic than DDL (Hintikka, 1962). He presents a logical representation of the FUL that has not yet been implemented using an automated theorem prover, so I implement it in Is-

Test	Naive	Kroy	Custom
FUL Stronger than DDL	×	✓	✓
Obligation Universalizes Across People	×	✓	✓
Obligations Never Contradict	×	×	✓
Distributive Property for Obligations Holds	×	×	✓
Prohibits Actions That Are Impossible to Universalize	×	×	✓
Robust Representation of Maxims	×	×	✓
Can Prohibit Conventional Acts	×	×	✓
Can Prohibit Natural Acts	×	×	✓

Figure 1: Table showing which tests each implementation passes. "Naive" indicates raw DDL, "Kroy" is my implementation of Moshe Kroy's formalization of the FUL, and "Custom" is my novel implementation.

abelle.[23] I find that my implementation outperforms both other implementations. Full test results are summarized in Figure 1. Below, I briefly explain some notable tests.

FUL Stronger than DDL The FUL should not hold in raw DDL, which I use a control group. If the FUL holds in the base logic, then adding it as an axiom doesn't make the logic any stronger, which is troubling because the base logic does not come equipped with the categorical imperative. DDL defines basic properties of obligation, such as ought implies can, but contains no axioms that represent the Formula of Universal Law. Therefore, if a formalization of the FUL holds in the base logic, then it is too weak to actually represent the FUL. Both Kroy's formalization and my implementation do not hold in the base logic, and thus represent progress over the control group. To test this property, I used Nitpick to find a countermodel in which my version of the FUL does not hold. I performed this test before adding the FUL as an axiom, since after adding it no countermodel will be possible.

Obligation Universalizes Across People The obligations prescribed by the Formula of Universal Law should generalize across people. In other words, if a maxim is obligated for one person, then it is obligated for all other people because maxims are not person-specific. Velleman argues that, because reason is accessible to everyone identically, obligations apply to all people

[23] I present the complete implementation in Appendix B.

equally (Velleman, 2005, 25). When Kant describes the categorical imperative as the objective principle of the will, he is referring to the fact that, as opposed to a subjective principle, the categorical imperative applies to all rational agents equally (Kant, 1785, 16). At its core, the FUL best handles, "the temptation to make oneself an exception: selfishness, meanness, advantagetaking, and disregard for the rights of others" (Korsgaard, 1985, 30). Kroy makes this property the center of his formalization, which essentially says that if an act is permissible for someone, it is permissible for everyone.[24] Kroy's formalization and my formalization satisfy this property, but raw DDL does not. Below I run this test for my formalization.

lemma *wrong-if-wrong-for-someone*:

shows $\forall\, w.\ \forall\, c{::}t.\ \forall g{::}t.\ \forall\, a.\ \exists s{::}s.\ O\{\neg\ (W\ (c, a, g)\ s)\ |\ c\}\ w \longrightarrow (\forall p.\ \ O\{\neg\ (W\ (c, a, g)\ p)\ |\ c\}\ w)$

by *blast*

— I represent my tests as lemmas that I expect Isabelle to either prove or refute. The statement following the keyword `shows` is the sentence of the lemma, and the proof follows the `by` keyword

— This lemma shows that if a maxim (c, a, g) is wrong for subject s at a world, then it is wrong for all people at that world. Isabelle automatically completed this proof using the `blast` method, which implements a generic tableau prover, a proof method that operates on lists of formulae using rules for conjunction, disjunction, universal quantification, and existential quantification (Paulson, 1999).

lemma *right-if-right-for-someone*:

shows $\forall\, w.\ \forall\, c{::}t.\ \forall g{::}t.\ \forall\, a.\ \exists s{::}s.\ O\{W\ (c, a, g)\ s\ |\ c\}\ w \longrightarrow (\forall p.\ \ O\{W\ (c, a, g)\ p\ |\ c\}\ w)$

by *blast*

— This lemma shows that if a maxim (c, a, g) is right for subject s at a world, then it is right for all people at that world. The proof similarly proceeds using `blast`.

Obligations Never Contradict Contradictory obligations make obeying the prescriptions of an ethical theory impossible. Kant subscribes to the general, popular view that morality is supposed to guide action, so ought implies can.[25] Kohl reconstructs Kant's argument for this principle as follows: if the will cannot comply with the moral law, then the moral law has no

[24] Formally, $P\{A(s)\} \longrightarrow \forall p.P\{A(p)\}$.

[25] Kohl points out that this principle is referred to as Kant's dictum or Kant's law in the literature (Kohl, 2015, footnote 1).

prescriptive authority for the will (Kohl, 2015, 703-4). This defeats the purpose of Kant's theory, which is to develop an unconditional, categorical imperative for rational agents. Ought implies can requires that obligations never contradict each other, because an agent can't perform contradictory actions. Therefore, any ethical theory that respects ought implies can, and Kantian ethics in particular, must not result in conflicting obligations. Kant only briefly discusses contradictory obligations in *Metaphysics of Morals*, where he argues that conflicting moral obligations are impossible under his theory (Kant, 2017, V224). Particularly, the categorical imperative generates "strict negative laws of omission," which cannot conflict by definition (Timmermann, 2013, 45).[26] Both raw DDL and Kroy's formalization allow contradictory obligations, but I explicitly add an axiom to my formalization that prohibits contradictory obligations.

lemma *conflicting-obligations*:
 shows $\neg\ (O\{W\ (c, a, g)\ s|c\} \wedge O\{\neg(W\ (c, a, g)\ s)|\ c\})\ w$
 using *no-contradictions* **by** *blast*

— This test passes immediately by the new axiom that prohibits contradictory obligations.

lemma *implied-contradiction*:
 assumes $(((W\ (c1, a1, g1)\ s) \wedge (W\ (c2, a2, g2)\ s)) \rightarrow \bot)\ w$
 shows $\neg\ (O\{W(c1, a1, g1)\ s|c\} \wedge O\{W(c2, a2, g2)\ s|c\})\ w$
 using *assms no-contradictions* **by** *blast*

— This stronger property states that the combination of obligatory maxims can't imply a contradiction and should hold for the same reasons that contradictory obligations aren't allowed. The added axiom also makes this test pass.

Contradictory obligations are closely related to two other properties. First is the idea that obligation implies permissibility, or that obligation is a stronger property than permissibility. If there are no contradictory obligations, then this property holds because actions

[26]The kinds of obligations generated by the FUL are called "perfect duties" which arise from "contradictions in conception," or maxims that we cannot even concieve of universalizing. These duties are always negative and thus never conflict. Kant also presents "imperfect duties," generated from "contradictions in will," or maxims that we can concieve of universalizing but would never want to. These duties tend to be broader, such as "improve oneself" or "help others," and are secondary to perfect duties. My project only analyzes perfect duties, as these are always stronger than imperfect duties.

are either permissible or prohibited and obligation contradicts prohibition. In a system with contradictory obligations, this property fails because there is some maxim that is obligated but also prohibited and therefore not permisible. The simple proof below shows that this property is incompatible with contradictory obligations.

lemma $\forall\, w.\ \exists\, A.\ (((O\,\{A\} \wedge O\,\{\neg\, A\})w)) \equiv (\exists\, B.\ (\neg\,(O\,\{B\} \rightarrow \neg\, O\,\{\neg B\}))\ w)$
by *simp*
— This lemma shows that if there is some maxim A such that A and $\neg A$ are both obligatory (which is the formal statement of contradictory obligations), then obligation does not always imply permissibility.
— `simp` is the simplification tactic, which unfolds definitions to complete a proof.

Distributive Property for Obligations Holds Another property related to contradictory obligations is the distributive property for the obligation operator.[27] This is another property that we expect to hold. The rough English translation of $O\{A \wedge B\}$ is "you are obligated to do both A and B". The rough English translation of $O\{A\} \wedge O\{B\}$ is "you are obligated to do A and you are obligated to do B." We think those English sentences mean the same thing, and they should mean the same thing in logic as well. Moreover, if that (rather intuitive) property holds, then contradictory obligations are impossible, as shown in the below proof.

lemma *distributive-implies-no-contradictions*:
assumes $\forall\, A\, B.\ \models ((O\,\{A\} \wedge O\,\{B\}) \equiv O\,\{A \wedge B\})$
shows $\forall\, A.\ \models(\ \neg(O\,\{A\} \wedge O\,\{\neg\, A\}))$
using *O-diamond assms* **by** *blast*
— The `assumes` keyword indicates assumptions used when proving a lemma. I use it here to represent metalogical implication. With the assumption, the lemma above reads, "If the distributive property holds in this logic, then obligations cannot contradict."

Again, this test fails for raw DDL and for Kroy's formalization, but passes for my interpretation because I require that obligations don't contradict as an axiom.

lemma *distribution*:
assumes $\models (O\,\{A\} \wedge O\,\{B\})$
shows $\models O\,\{A \wedge B\}$

[27] Formally, $O\{A\} \wedge O\{B\} \longleftrightarrow O\{A \wedge B\}$.

using *assms no-contradictions* **by** *fastforce*

— The proof proceeds almost immediately using the new axiom.

Prohibits Actions That Are Impossible to Universalize Recall that the logical contradiction interpretation of the Formula of Universal Law prohibits lying because in a world where everyone simultaneously lies, lying is impossible. In other words, not everyone can simultaneously lie because the institution of lying and believing would break down. In Section 3.1.3, I recreated Korsgaard's argument for why the logical contradiction interpretation is weaker than what the Formula of Universal Law should actually require. Therefore, any implementation of the FUL should be able to show that the actions prohibited by the logical contradiction interpretation are prohibited, because the set of actions prohibited by the practical contradiction interpretation is a superset of these. The FUL should show that actions that cannot possibly be universalized are prohibited, because those acts cannot be willed in a world where they are universalized. This property fails to hold in both raw DDL and Kroy's formalization, but holds for my formalization. Showing that this property holds for my formalization required significant logical work and the full code is presented in Appendix C.

Robust Representation of Maxims Kant does not evaluate the correctness of acts, but rather of maxims. Therefore, any faithful formalization of the categorical imperative must evaluate maxims, not acts. This requires representing a maxim and making it the input to the obligation operator, which only my implementation does. Because my implementation includes the notion of a maxim, it is able to perform sophisticated reasoning as demonstrated in Sections 4.1 and 4.2. Staying faithful to the philosophical literature enables my system to make more reliable judgements.

Can Prohibit Conventional and Natural Acts When arguing for the practical contradiction interpretation, Korsgaard makes a distinction between conventional and natural acts (Korsgaard, 1985). A conventional act relies on a convention, like the convention that a promise is a commitment, whereas a natural act is possible simply because of the laws of the natural world. Conventional acts exist within a practice, which is "comprised of certain rules, and its existence (where it is not embodied in an institution with sanctions) consists in the general acknowledgement and following of those rules" (Korsgaard, 1985, 10). Promising is a conventional act because it exists as a practice. Murder, on the other hand, is an example of a natural act because its existence only depends on the laws of nature(Korsgaard, 1985, 11).

It is easier to show the wrongness of conventional acts because there are worlds in which these acts are impossible; namely, worlds in which the convention does not exist. For example, the common argument against falsely promising is that if everyone were to falsely promise, the convention of promising would fall apart because people wouldn't believe each other, so falsely promising is prohibited. It is more difficult to show the wrongness of a natural act, like murder or violence. These acts can never be logically impossible; even if everyone murders or acts violently, murder and violence will still be possible, so it is difficult to show that they violate the FUL.

Both raw DDL and Kroy's interpretation fail to show the wrongness of conventional or natural acts. My system shows the wrongness of both natural and conventional acts because it is faithful to Korsgaard's practical contradiction interpretation of the FUL, which is the canonical interpretation of the FUL (Ebels-Duggan, 2012; Korsgaard, 1985). I run this test in Chapter 4, where I use my system to reason about two ethical dilemmas, one which involves conventional acts and the other which involves natural acts. I present an additional example demonstrating that my implementation passes this test in Appendix C.

4 Applications

In this chapter, I demonstrate that my system can produce correct judgements for challenging moral dilemmas that naive ethical reasoning cannot satisfactorily handle. Because my system is faithful to philosophical literature, it can reproduce complex ethical judgements presented by philosophers as solutions to controversial open questions in ethics. These dilemmas serve as examples of how my system could be used in practice and demonstrate my system's ability to formalize longer, more complicted ethical arguments than those presented in Chapter 3. Moreover, in the process of formalizing these dilemmas, I isolated the exact conditon that makes a maxim about lying wrong, an insight that could contribute to the philosophical literature on lying.

Many of the tests in Section 3.2 perform metaethical reasoning, which analyzes properties of morality itself and involves questions about the nature of ethical truth. In this chapter, I perform "applied ethical reasoning," which is the use of ethics to resolve dilemmas and make judgements about what an agent should or should not do. This is the kind of reasoning necessary for an AI agent using my system to navigate the real world.

One challenge of applied ethical reasoning is that it requires more factual background than metaethical reasoning. Because metaethics is about ethics itself, and not about the dilemmas that ethics is supposed to help us resolve, this kind of reasoning requires very little knowledge about the world. Applied ethical reasoning, on the other hand, focuses on a particular ethical dilemma and thus requires enough factual background, or common sense, to understand the dilemma and options at hand. For example, an applied ethicist evaluating the permissibility of lying needs a robust definition of the term lying and likely some understanding of communication and truth telling. Kantians specifically describe this common sense as "postulates of rationality" that are nontrivial and nonnormative, but still part of the process of practical reasoning itself (Silber, 1974).

In this chapter, I tackle this challenge by endowing my system with this kind of common sense in the specific case of lying. My system needs common sense facts and definitions because, while it has the ability to reason using the Formula of Universal Law, this reasoning must be applied to objects that are defined using common sense. Because these common sense facts can determine my system's judgements, they are part of the trusted code base for my system, or the logic and code that a user must trust in order to trust my system. Changing these common sense facts will change the judgements that my system makes. For example, if we define truth telling as an act that is self-contradictory (perhaps by defining it as $p \wedge \neg p$), then my system will output that truth telling is prohibited. Malicious common sense facts and definitions will result in bad ethical judgements. The challenge of endowing automated ethical reasoners with common sense reasoning is not unique to my system, and virtually all prior attempts in machine ethics face similar challenges. Many prior attempts sidestep this question, whereas I contribute an prototype implementation of one kind of common sense reasoning.

This chapter will provide examples of the kinds of common sense facts required to get my system off the ground. I use a lean and uncontroversial common sense database to achieve robust and powerful results. This serves as evidence for the ease of automating Kantian ethics, an example of the additional work required to use my system in practice, and a demonstration of my system's power and flexibility. These examples demonstrate that, armed with nuanced common sense facts, my system can make sophisticated judgements faithful to philosophical literature.

4.1 Lies and Jokes

The moral status of lying is hotly debated in the Kantian literature. I focus on two dilemmas presented in Korsgaard's "Right to Lie," which examines Kant's prohibition on lying (Korsgaard, 1986). She begins with the case of lying and joking. Many of Kant's critics argue that his prohibition on lies includes lies told in the context of a joke, which should be permissible. Korsgaard responds by arguing that there is a crucial difference between lying and joking: lies involve deception, but jokes do not. The purpose of a joke is amusement, which does not rely on the listener believing the story told. Given appropriate definitions of lies and jokes, my system shows that jokes are permissible but lies are not, demonstrating its power and flexibility. This section demonstrates how my system can be used in practice; it needs to be given some baseline common sense facts, but with those facts, it can make sophisticated judgements. Moreover, because my system is faithful to definitions found in philosophical literature, it can perform nuanced reasoning, demonstrating the value of faithful automated ethics.

First, Korsgaard argues that the categorical imperative prohibits lies because they involve deception. When universalized, lies will no longer be believed, so lying could never be an effective way of achieving any goal when universalized. Korsgaard points out that "we believe what is said to us in a given context because most of the time people in that context say what they really think" (Korsgaard, 1986, 4). In order to formalize this argument, I first need to define lying and formalize Korsgaard's argument about the basis of trust.

I define lying and trust in terms of belief. As in Section 3, I choose thin, or minimal, definitions to reduce the potential for controversy in my system's factual background.

consts *believe*::$s \Rightarrow t \Rightarrow t$ (*- believes -*)

— `believe` is a constant that maps a subject and a term to another DDL term. For example, subject "Sara" might believe the term "the sky is blue" to create the sentence "Sara believes that the sky is blue," which can be true or false at a world.

Logicians and epistemologists develop and debate complex logics of belief and knowledge (Baltag and Renne, 2016). I avoid this complexity by defining the concept of belief simply as a constant that maps a subject, term pair to a term. For the examples in this section, this choice suffices. I encode some minimal properties of belief below, but avoid any full definition of the term.

Belief is useful to construct the idea of "knowingly uttering a falsehood," a core component of both lying and joking.

consts *utter*::$s \Rightarrow t \Rightarrow t$

— `utter` also maps a subject and term to another DDL term. For example, the sentence "Sara utters, 'I am hungry' " is a DDL term that can be true or false at a world.

abbreviation *utter-falsehood*::$s \Rightarrow t \Rightarrow t$ **where**

$\textit{utter-falsehood}\ s\ t \equiv\ (\textit{utter}\ s\ t) \wedge (\neg\ t)$

— To utter a falsehood is to utter a statement that is false, or to utter t when $\neg t$.

abbreviation *knowingly-utter-falsehood*::$s \Rightarrow t \Rightarrow t$ **where**

$\textit{knowingly-utter-falsehood}\ s\ t \equiv (\textit{utter-falsehood}\ s\ t) \wedge (\neg\ (\textit{believe}\ s\ t))$

— Sometimes we unknowingly utter falsehoods. For example, if I believe that the Earth is flat, then when I utter, "the Earth is flat," I am unknowingly uttering a falsehood. This motivates defining the idea of knowingly uttering a falsehood, which requires both uttering a falsehood and not believing your utterance. If I utter "the Earth is flat," even though I know that the Earth is round, I am knowingly uttering a falsehood.

The above abbreviations are the core of my formalization of Korsgaard's definitions of lying and joking. They are also relatively uncontroversial and encode little moral or normative content. They say nothing about the moral status of uttering a falsehood, the agent's intention when making the utterance, or the conversational norms guiding the utterance. The complexities of a complete definition of lying or belief are unnecessary for Kantian ethics, and therefore for my system, to make moral judgements.

Using the above definitions, I define lying. I characterize a maxim as involving a lie if the act requires knowingly uttering a falsehood and the end requires that some person p believe the false statement t.

abbreviation *lie*::$\textit{maxim} \Rightarrow \textit{bool}$ **where**

$\textit{lie} \equiv \lambda\ (c, a, g).\ \exists\, t.\ (a \longrightarrow (\lambda s.\ \textit{knowingly-utter-falsehood}\ s\ t)) \wedge (\exists\, p.\ \forall\, w.\ (g \rightarrow (\textit{believe}\ p\ t))\ w)$

— The abbreviation above maps a maxim to a boolean value that indicates if it is a lie.

To avoid unintentional wrongdoing, I focus on "knowing lies," in which the speaker is aware that they are lying. It is uncontroversial that, in order for an act to be a knowing lie, the speaker must utter a false statement that they do not believe. This also makes it easier to make moral judgements about the speaker's action, since they were, at the very least, aware of their lie. The second half of this definition requires that the goal of the lie is deception. This is inspired by Korsgaard's interpretation of a lie. She understands a lie as a kind of falsehood that is usually effective *because* it decieves (Korsgaard, 1986, 4). In my formalization, this means that the purpose or goal of the maxim must involve decieving someone, or, in other words, that someone believe what the speaker knows to be a falsehood.

With the above logical background, I automate Korsgaard's argument that maxims that involve lying are prohibited. First, I define the relevant subject and maxim.

consts $me::s$

— I am trying to reason about *my* obligations so I will define myself as a specific subject. Again, this is a minimal definition that does not include any facts about me, such as the fact that my name is Lavanya or that I have brown hair.

consts $m::maxim$

— I also define a maxim m. My goal is to show that if m is a maxim about lying, then m is prohibited.

consts $c::t$ $a::os$ $g::t$

— m will be composed of the circumstances, act, and goal above.

In the following lemma, I use my system to show that lying is prohibited. The assumptions of this lemma represent the common sense necessary to reach this conclusion. This common sense background is a direct formalization of the premises of Korsgaard's argument. Using these relatively minimal premises about individual behavior, my system derives a prohibition against lying.

lemma *lying-prohibited*:

assumes $m \equiv (c::t, a::os, g::t)$

assumes $\forall w. \forall s.$ *well-formed m s w*

— Initial technical set-up: m is a well-formed maxim composed of some circumstances, act, and goal.

assumes *lie m*

— m is a maxim about lying as defined above. Precisely, it is a maxim in which the action requires knowingly uttering a falsehood and the goal requires that someone believe this falsehood.

assumes $\forall t\ w.\ ((\forall p.\ \textit{utter-falsehood}\ p\ t\ w) \longrightarrow (\forall p.\ \neg\ (\textit{believe}\ p\ t)\ w))$

— Assumption that if everyone utters false statement t, then no one will believe t. This assumption is Korsgaard's core piece of "common sense" about lying (Korsgaard, 1986, 5). This simple assumption encodes the common sense knowledge that human communication involves an implicit trust, and that when this trust erodes, the convention of communication begins to break down and people no longer believe each other. Call this the "convention of trust" fact. In the rest of this section, I will test versions of this assumption, effectively encoding different common sense understandings of lying.

assumes $\forall w.\ c\ w$

— Restrict our focus to worlds in which the circumstances hold. A technical detail.

shows $\models$ (*prohibited m me*)

proof –

have $(\forall p\, w.\, (W\, m\, p)\, w) \longrightarrow (\models (c \to (\neg g)))$

by (*smt assms*(*1*) *assms*(*2*) *assms*(*5*) *case-prod-beta fst-conv old.prod.exhaust snd-conv*)

— This proof requires some manual work. After I divide the proof into the intermediate steps shown here, Isabelle is able to do the rest. This step says that if m is universalized, then the circumstances won't lead to the goal, which is close to the idea of the maxim not being universalizable.

have *not-universalizable m me*

by (*metis* (*mono-tags*, *lifting*) *assms*(*1*) *assms*(*2*) *case-prod-beta fst-conv snd-conv*)

thus *?thesis*

using *FUL assms*(*2*) **by** *blast*

— $?thesis$ is Isabelle's syntax for the goal of the lemma. In this case, $?thesis$ is equivalent to $\models$ $prohibited\ m\ me$.

qed

The lemma above demonstrates that my system finds that lying is prohibited with a thin definition of lying and relatively uncontroversial facts about the world. My system needs two pieces of common sense to complete this proof. First, I defined lying as knowingly uttering a falsehood with the goal that someone believe the falsehood, a definition of lying that is relatively well-accepted. Second, I assumed (following in Korsgaard's footsteps) that if everyone lies in a given context, then people will stop believing each other in that context. This is a slightly heavier assumption, but it is still so uncontroversial that Korsgaard doesn't bother to justify it in her argument against lying (Korsgaard, 1986).

Now that I have formalized Korsgaard's argument for why lying is prohibited, I will implement her argument for why jokes are permissible. Specifically, she defines a joke as a story that is false and argues that joking is permissible because "the universal practice of lying in the context of jokes does not interfere with the purpose of jokes, which is to amuse and does not depend on deception" (Korsgaard, 1986, 4). I use the fact that a joke does not depend on deception as the defining feature of a joke.

abbreviation *joke*::*maxim*$\Rightarrow$*bool* **where**

$joke \equiv \lambda\ (c, a, g).\ \exists\, t.\ (a \longrightarrow (\lambda s.\ \textit{knowingly-utter-falsehood}\ s\ t)) \wedge \neg\ (\exists\, p.\ \forall\, w.\ (g \rightarrow (\textit{believe}\ p\ t))\ w)$

— Thie abbreviation states that a maxim is a joke if the action involves knowingly uttering a falsehood but the goal does *not* require that someone believe the falsehood told.

This definition of a joke defines a joke as a falsehood uttered for some purpose that doesn't require deception, where deception involves someone believing the uttered falsehood. This definition doesn't require any conception of humor, but merely distinguishes jokes from lies.

Korsgaard argues that her above argument for a prohibition against lying also implies that joking is permissible, because its purpose is not to decieve, but something else entirely. This means that, even armed with the same convention of trust assumption as above, joking should be permissible. The lemma below shows exactly that.

lemma *joking-not-prohibited*:

assumes $m \equiv (c{::}t, a{::}os, g{::}t)$

assumes $\forall\, w.\ \forall\, s.\ \textit{well-formed}\ m\ s\ w$

— Initial set-up: m is a well-formed maxim composed of some circumstances, act, and goal.

assumes *joke m*

— m is a maxim about joking. Precisely, it is a maxim in which the action is to knowingly utter a falsehood and the goal does not require that someone believe this falsehood.

assumes $\forall\, t\ w.\ ((\forall\, p.\ \textit{utter-falsehood}\ p\ t\ w) \longrightarrow (\forall\, p.\ \neg\ (\textit{believe}\ p\ t)\ w))$

— The same convention of trust assumption as in the above example.

assumes $\forall\, w.\ c\ w$

— Restrict our focus to worlds in which the circumstances hold. A technical detail.

shows $\models (\textit{permissible}\ m\ \textit{me})$

by (*smt assms*(*1*) *assms*(*2*) *assms*(*3*) *case-prod-conv*)

— Isabelle is able to show that maxims about joking are permissible. It also lists the facts used in its proof, which offer insight into how it arrived at its judgement. Specifically, it uses assumptions 1, 2, and 3, which are the logical background and definition of the joking maxim. Notably, it does not use the convention of trust assumption. This demonstrates that even the convention of trust assumption is not strong enough to prohibit joking, which is exactly the desired result.

My system shows that lies are prohibited and jokes are permissible with thin conceptions of amusement and deception. This shows that it isolates a necessary and sufficient property of this class of maxims that fail the universalizability test. My definitions of a lie and joke only differ in whether or not their goal requires that someone believe the falsehood in question, so this is a necessary and sufficient condition for a maxim about knowingly uttering falsehoods to be prohibited. This logical fact derived by my system tracks a fact implicit in Korsgaard's argument and in most Kantian accounts of lying: the wrongness of lying is derived from the requirement that someone believe the falsehood. The logical reality that this property is necessary and sufficient to generate a prohibition reflects a deep philosohopical explanation of *why* certain maxims about uttering falsehood fail the universalizability test. Universalizing uttering a falsehood makes belief in that falsehood impossible, so any maxims with goals that require believing in the falsehood will be prohibited.

This account not only describes the kind of maxims that fail or pass the universalizability test, but it also provides a guide to constructing permissible maxims about uttering falsehoods. As an example, consider the idea of throwing a surprise birthday party. At first glance, the maxim of action is something like, "When it is my friend's birthday, I will secretly plan a party so that I can surprise them." The goal "so that I can surprise them" clearly requires that your friend believe the falsehood that you are not planning a party, else the surprise would be ruined. This seems to imply that Kantian ethics would prohibit surprise parties, which is a sad conclusion for birthday-lovers everywhere. Knowing that this maxim is prohibited because the goal requires belief in a falsehood provides a way to rescue surprise parties. When throwing a surprise party, the objective is *not* to surprise your friend, but to celebrate your friend and help them have a fun birthday. If someone ruins the surprise, but the party is still fun and the birthday person feels loved, then such a party is a success! Someone who calls this party a failure is clearly missing the point of a surprise party. The goal of a surprise party is not the surprise itself, but rather celebrating the birthday person. The modified goal[28] no longer

[28] Some may worry that this argument implies that the "means justify the ends," or that modifying an act's goal can change its moral worth. This conclusion is not only unsurprising to Kantians, but it is the defining feature of their theory. Under Kantian ethics, an action alone is not the kind of thing that can be moral or immoral; rather, a maxim, or a circumstances, act, goal tuple, is what has moral worth. The rightness of an action can hinge on the maxim's goal, circumstances, or act because these three features of an action are inseparable.

requires belief in the falsehood and thus passes the universalizability test.

There are two implications of this section. First, my system is capable of performing ethical reasoning sophisticated enough to show that lying is prohibited but joking is not. This is a direct consequence of my system's use of a robust conception of a maxim, which encodes the goal of an act as part of the maxim being evaluated. Because my implementation is faithful to philosophical literature, it is able to recreate Korsgaard's solution to a complex ethical dilemma that philosophers debated for decades. Second, in the process of making this argument precise, my system isolated a necessary and sufficient condition of a maxim about uttering a falsehood being prohibited: that the goal require that someone believe the falsehood. This condition made an long-standing argument in Kantian ethics more precise, can guide the correct formulation of future maxims, and could contribute to the rich philosophical conversation about the wrongness of lying. In other words, an insight generated by the computer provides value to ethicists, bolstering the argument for computational ethics provided in Section 5.2.

4.2 Lying to a Liar

My system can not only distinguish between lying and joking, but it can also resolve the paradox of the murderer at the door. In this dilemma, murderer Bill knocks on your door asking about Sara, his intended victim. Sara is at home, but moral intuition says that you should lie to Bill and say that she is out to prevent him from murdering her. Many critics of Kantian ethics argue that the FUL prohibits you from lying in this instance; if everyone lied to murderers, then murderers wouldn't believe the lies and would search the house anyways. Korsgaard resolved this centuries-long debate by noting that the maxim of lying to a murderer is actually that of lying to a liar: Bill cannot announce his intentions to murder; instead, he must "must suppose that you do not know who he is and what he has in mind" (Korsgaard, 1986).[29] Thus,

[29]Korsgaard assumes that the murderer will lie about his identity in order to take advantage of your honesty to find his victim. In footnote 5, she accepts that her arguments will not apply in the case of the honest murderer who announces his intentions, so she restricts her focus to the case of lying to a liar. She claims that in the case of the honest murderer, the correct act is to refuse to respond. Since I am formalizing Korsgaard's argument, I also accept this assumption.

the maxim in question specifies that when someone lies to you, you are allowed to lie to them. The maxim of lying to the murderer is actually the maxim of lying to a liar, which is permissible.

In this section, I formalize Korsgaard's argument for the permissibility of lying to a liar. First, I define Bill's maxim, which is to hide his intention to murder.

consts *murderer*::*s*

— This example involves the murderer as an additional subject.

consts *not-a-murderer*::*t*

— This statement represents the lie that the murderer tells you. By not announcing his intention, he is implicitly telling you that he is not a murderer, as people normally assume that those knocking on their door are not murderers.

consts *when-at-my-door*::*t*

— These are the circumstances that the murderer is in.

consts *find-victim*::*t*

— This will be the murderer's goal: to find his victim.

abbreviation *murderers-maxim*::*maxim* **where**

murderers-maxim ≡ (*when-at-my-door*, *λs. knowingly-utter-falsehood s not-a-murderer*, *find-victim*)

— Using the above definitions, I can define the murderer's maxim as, "When at your door, I will knowingly utter the falsehood that I am not a murderer in order to find my intended victim."

These constants are defined only in relation to each other and elide most of the complex features of murder, life, and death. These thin representations will suffice to show the wrongness of the murderer's maxim. Similarly, I can use thin representations to define your maxim of lying about Sara's whereabouts.

consts *victim-not-home*::*t*

— This statement is the lie that you tell the murderer: that his intended victim is not at home.

abbreviation *murderer-at-door*::*t* **where**

murderer-at-door ≡ *W murderers-maxim murderer*

— These are the circumstances that you are in: the murderer has willed his maxim and thus lied to you.

consts *protect-victim*::*t*

— Your goal is to protect the murderer's intended victim.

abbreviation *my-maxim*::*maxim* **where**

$my\text{-}maxim \equiv (murderer\text{-}at\text{-}door,\ \lambda s.\ knowingly\text{-}utter\text{-}falsehood\ s\ victim\text{-}not\text{-}home, protect\text{-}victim)$

— Using these definitions, I construct your maxim, which is "When a murderer is at my door, I will knowingly utter the falsehood that his intended victim is not at home, in order to protect the victim."

I now formalize Korsgaard's argument for the permissibility of lying to a liar. She modifies the convention of trust assumption above when she argues that, if the murderer believes that you don't believe he is a murderer, he will think that you won't lie to him. Precisely, she claims that, "it is because the murderer supposes you do not know what circumstances you are in—that is, that you do not know you are addressing a murderer—and so does not conclude from the fact that people in those circumstances always lie that you will lie" (Korsgaard, 1986, 6). Even though the maxim of lying to a murderer is universalized, Bill thinks that you don't know his true identity. Thus, even if you have willed this maxim, he thinks that you won't perform the act of lying to the murderer, since he thinks that you don't think you're in the relevant circumstances. I formalize this argument below.

lemma *lying-to-liar-permissible*:

assumes $\models (well\text{-}formed\ murderers\text{-}maxim\ murderer)$

assumes $\models (well\text{-}formed\ my\text{-}maxim\ me)$

— Initial set-up: both maxims are well-formed.

assumes $\models (protect\text{-}victim \rightarrow (murderer\ believes\ victim\text{-}not\text{-}home))$

— In order for you to protect the victim, the murderer must believe that the victim is not home.

assumes $\forall sentence::t.\ \forall p1::s.\ \forall p2::s.\ \forall w::i.\ ((p1\ believes\ (utter\text{-}falsehood\ p2\ sentence))\ w) \longrightarrow (\neg\ (p1\ believes\ sentence)\ w)$

— This is one of two assumptions that encode Korsgaard's core argument. If $p1$ believes that $p2$ utters a sentence as a falsehood, then $p1$ won't believe that sentence. This is a modification of the convention of trust assumption from above, and I will refer to it as the "convention of belief" assumption. Again, like the convention of trust assumption, this assumption is uncontroversial: if I think you are lying, then I won't believe you.

assumes $\forall c\ a\ g\ w.\ \forall p1::s.\ \forall p2::s.\ (universalized\ (c, a, g)\ w) \longrightarrow ((p1\ believes\ (p2\ believes\ c)) \rightarrow (p1\ believes\ (a\ p2)))\ w$

— This is the second major common sense assumption. If the maxim (c, a, g) is universalized, then if $p1$ believes that $p2$ believes they are in the given circumstances, then $p1$ believes that $p2$ performs the act. In other words, $p1$ will believe that $p2$ wills the maxim. I will refer to this as the "convention of willing" assumption. This follows directly from Korsgaard's conception of universalizability: when a maxim is universalized, everyone wills it and thus notices the pattern of everyone willing it. If you observe that many do X in circumstances C, you will assume that everyone does X in circumstance C.

assumes ∀ *w*. *murderer-at-door w*

— Restrict our focus to worlds in which the circumstance of the murderer being at my door holds. A technical detail.

shows ⊨ (*permissible my-maxim me*)

using *assms*(*1*) *assms*(*6*) **by** *auto*

— Isabelle completes this proof using the first and sixth assumption, ignoring the convention of belief and convention of willing assumptions. These common sense assumptions are not strong enough to generate a prohibition against lying to a liar and are thus unused in this proof.

The above lemma shows that, with a nuanced set of common sense facts, my system can show that lying to a liar is permissible. One worry may be that this set of assumptions is too weak to yield a prohibition against wrong maxims, like that of the murderer. As a sanity check, I show that this set of assumptions prohibits the murderer's maxim below.

lemma *murderers-maxim-prohibited*:

assumes ∀ *w*. *well-formed murderers-maxim murderer w*

— Initial set-up: the murderer's maxim is well-formed.

assumes ⊨ (*find-victim* → (*believe me not-a-murderer*))

— In order for you to protect the victim, the murderer must believe that the victim is not home.

assumes ∀ *sentence*::*t*. ∀ *p1*::*s*. ∀ *p2*::*s*. ∀ *w*::*i*. ((*p1 believes* (*utter-falsehood p2 sentence*)) *w*) ⟶ (¬ (*p1 believes sentence*) *w*)

— The convention of belief assumption.

assumes ∀ *c a g w*. (*universalized* (*c*, *a*, *g*) *w*) ⟶ ((*person1 believes* (*person2 believes c*)) → (*person1 believes* (*a person2*))) *w*

— The convention of willing assumption.

assumes ∀ *w*. *when-at-my-door w*

— Restrict our focus to worlds in which the circumstance of the murderer being at my door holds. A technical detail.

shows $\models$ (*prohibited murderers-maxim murderer*)

proof –

have $(\forall p\ w.\ (W\ \textit{murderers-maxim}\ p)\ w) \longrightarrow (\models (\textit{when-at-my-door} \rightarrow (\neg\ \textit{find-victim})))$

using *assms*(*2*) **by** *auto*

have *not-universalizable murderers-maxim murderer*

using *assms*(*2*) *assms*(*5*) *case-prod-beta fst-conv internal-case-prod-def old.prod.case old.prod.exhaust snd-conv* **by** *auto*

thus *?thesis*

using *FUL assms*(*1*) **by** *blast*

qed

This concludes my examination of the maxim of lying to a liar. I was able to show that, by modifying the common sense facts used, my system can show that lying to a liar is permissible, but lying in order to find a victim is not. The assumptions used in this example were a little more robust, but still ultimately uncontroversial because they were direct consequences of Korsgaard's definition of willing and of ordinary definitions of lying. These thin assumptions were sufficient to recreate Korsgaard's solution to an open ethical problem. Armed with this common sense, my system generated a conclusion that many critics of Kant prior to Korsgaard failed to see.[30]

While this example demonstrates the power of my system, it also shows how vital the role of the common sense reasoning is. Slight, intuitive changes in the factual background achieved completely different conclusions about lying. This example also demonstrates the importance and difficulty of correctly formulating the maxim, particularly its circumstances. Korsgaard's argument for the permissibility of lying to a murderer hinged on a clever formu-

[30]While it is true that lying to the murderer should be permissible, Korsgaard notes that many may want to say something stronger, like the fact that lying to the murderer is obligatory in order to protect the intended victim (Korsgaard, 1986, 15). Korsgaard solves this problem by noting that, while the FUL shows that lying to the murderer permissible, other parts of Kant's ethics show that it is obligatory. Recall that Kant presents perfect and imperfect duties, where the former are strict, inviolable, and specific and the latter are broader prescriptions for action. The details of this distinction are outside the scope of this thesis, but imperfect duties generate the obligation to lie to the murderer. An even more sophisticated automated Kantian reasoner could formalize imperfect duties and other formulatations of the categorical imperative in order to generate the obligation to lie to the murderer.

lation of the maxim that highlights the fact that the murderer is lying to you. The need for common sense reasoning to evaluate the universalizability test and to formulate a maxim is a potential limitation of my system, and I adress this concern in Section 5.1.

On one hand, the need for common sense facts is a limitation of my system. On the other, these examples show that common sense is within reach. Even thin, uncontroversial definitions and assumptions are enough to achieve nuanced ethical judgements. Moreover, these kinds of judgements demonstrate that, with additional work, my system could be used in practice to guide AI agents. The idea of AI making decisions as in the dilemmas above may seem far-fetched, but such scenarios are already becoming reality. For example, a "smart doorbell" may face a dilemma like that of the murderer at the door. Such an AI agent equipped with a future version of my system would be able to reason about lying to the murderer and arrive at the right judgement, guided by explainable, rigorous philosophical arguments.

5 Discussion

In the above chapters, I presented an implementation of automated Kantian ethics that, given an appropriately represented maxim and sufficient factual background, can classify a maxim as obligatory, permissible, or prohibited. In this chapter, I discuss the philosophical limitations and implications of this work.

First, I explore how my system can be used in practice to guide AI agents, help academic philosophers make philosophical progress, and augment the everyday practical reasoning that we all perform as we navigate the world. I discuss these issues in Sections 5.1, 5.2, and 5.3 respectively and outline the further work necessary to bring each of these applications to life. Much of this work has to do with how the input maxim given to my system is formulated and the role of common sense or prior knowledge.

In Section 5.4, I discuss theoretical objections to automated ethics, derived from the common philosophical intuition that there is no algorithm for ethics. I explore concrete versions of this objection and make a theoretical argument for the possibility of the kind of automated ethics I implement in this thesis. Finally, in Section 5.5, I situate my project among related work.

5.1 Automated Moral Agents in Practice

In Chapter 4, I demonstrated that my system is capable of performing sophisticated, nuanced ethical reasoning. In this section, I outline the additional components necessary for my system to guide AI agents in practice. As it stands, my system is a categorical imperative library that, given sufficient factual or "common sense" background, can take as input the logical representation of a maxim and return its moral status (if it is obligatory, prohibited, or permissible). My project 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

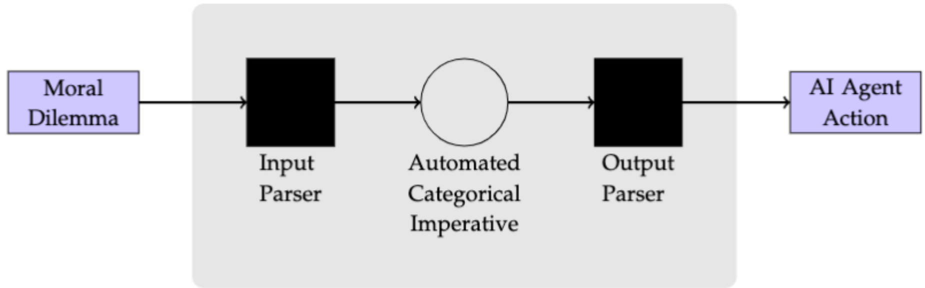

Figure 2: An example of an ethics engine for an artificial agent. This ethics engine passes a moral dilemma through an input parser, applies the automated categorical imperative test, and finally processes the output using an output parser, producing a prescription for action. I contribute the automated categorical imperative component.

logic. An output parser could translate my system's output into a prescription for action that the AI agent could act on. Figure 2 depicts the workflow of this example ethics engine.

In this workflow, an AI agent is faced with a moral dilemma in some internal representation. The input parser translates this internal representation into an appropriate logical representation, i.e. a circumstance, act, goal tuple. The output parser translates the output of the categorical imperative library (the moral status of the maxim as obligatory, prohibited, or permissible) to a prescription for action. In order for my system to be used in an AI agent using the workflow above, future work must develop such input and output parsers.

One of the biggest challenges for an ethics engine is the development of an input parser. An input parser for my implementation of automated Kantian ethics must translate a complex real-world situation into a flat, logical representation. This requires that the input parser determine which circumstances are morally relevant to a maxim, a controversial judgement. As mentioned in Section 3.1.2, there is robust debate on the circumstances that should be considered when formulating a maxim, inspired by a common criticism of Kantian ethics called the tailoring objection. Recall that the tailoring objection is the worry that arbitrarily specific circumstances render any maxim universalizable. For example, consider the maxim "When my name is Lavanya Singh and I am wearing a purple shirt and it is November 26th, I will lie in order to get some easy cash." Even if this maxim is willed universally, the circumstances are so specific that lying will not become the general mechanism for getting easy cash, so the lender will believe my lie and the maxim will remain effective. By tailoring the circumstances, any maxim can evade universalization.

The Kantian response to this criticism is to require that the circumstances included in the formulation of the maxim be morally relevant. In the example above, my purple shirt and the date have no bearing on the moral status of lying. On the other hand, consider the maxim, "When I am unemployed, I will murder someone in order to take their job." The circumstance "when I am unemployed" clearly has some moral relevance to the murder in question; it speaks to the motivation for the murder.

While this view has intuitive appeal, it raises the question of how we can determine which circumstances are morally relevant. O'Neill answers this question by noting that the Formula of Universal Law is a "test of moral worth rather than of outward rightness" (O'Neill, 1990, 98). The FUL is a way for an agent to decide how they should behave, not for a third-party to judge their behavior. Ethics is a personal process and the FUL is designed to help agents make decisions for themselves. Because agents use the FUL to evaluate their own behavior, the test is at its best when they make a good faith effort to isolate the *principle* of their action, rather than some "surface intent" (O'Neill, 1990, 87). The FUL is supposed to determine if an agent's principle of action is universally consistent, so it is most effective when an agent accurately formulates the principle that they act on. Circumstances are morally relevant if they reflect the way that the agent is thinking about their own action. In the example above, the circumstance of wearing a purple shirt doesn't reflect the principle of the liar's action. Its inclusion is a disingenous attempt to evade the universalizability test, but because the FUL is a test of personal integrity, it cannot withstand this kind of mental gymnastics.

While the above account explains how a well-intentioned human agent can determine morally relevant circumstances, the challenge remains open for automated ethics. However an action is turned into a maxim for my system to process, whether manually as I did in Chapter 4 or using an automatic input parser, this transformation must be a good-faith attempt to capture the principle of action. Correctly formulating a maxim requires what Kantians call "practical judgement," or common sense reasoning and factual background (O'Neill, 1989). Returning to the example above, the fact that being unemployed may contribute to one's desire to steal another's job is a consequence of practical judgement, not just pure reason alone. Translating everyday situations into appropriate maxims is the bulk of the reasoning that a Kantian human being does when making decisions. Automating this reasoning requires endowing a machine with common sense.

The challenge of formulating a maxim is one of the biggest obstacles to using my categorical imperative library in an AI ethics engine. One solution is for a human being to perform the role of the input parser by supervising the operation of an AI agent. When the agent stumbles onto an ethical dilemma, the human could take over, formulate the right question, and feed it into the categorical imperative library to see what action the categorical imperative would prescribe. Alternatively, AI developers can identify expected ethical dilemmas that the machine may face in advance and hardcode their own judgements for these dilemmas. For proponents of the "human-in-the-loop" model of AI ethics, in which ethical AI requires that humans guide machines, this kind of human involvement may be a feature (Lukowicz, 2019). Both of these solutions imply that the outcome of the universalizability test will depend on how the human formulates the maxim; if the human puts garbage into the test, the test will return garbage out.

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. In such a world, the input parser in my ethics engine would have to be automated. It is likely that, just as implementations of automated ethics choose a particular ethical theory and implement it, different implementations of such an input parser may adopt different interpretations of maxim formulation and morally relevant circumstances.

These interpretations could inspire heuristics to classify circumstances as morally relevant. For example, one such attempt could define a moral closeness relation between an action, a goal, and circumstances. This heuristic could define morally relevant circumstances as those that reach a certain closeness threshhold with the action and the goal. Another possible heuristic could define some set of morally important entities, and classify morally relevant circumstances as those that involve morally important entities. I propose a potential machine-learning based approach which formulates maxims based on a training set of appropriately formulated maxims in Section 5.4. This approach mimics how human beings formulate maxims; we use common sense and prior situational, factual, and ethical knowledge to isolate our principle of action. Determining morally relevant circumstances, either using heuristics or human involvement, is a ripe area for future work.

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 returns a value indicating if the maxim is obligatory, permissible, or prohibited. The maxim is 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 require testing if a certain theorem holds or not in my logic, a calculation that I demonstrate in Section 3.2. Testing these properties requires that my system have a database of common sense or factual background. Different applications of my system may require different factual background (e.g. a self-driving car needs to know traffic regulations), so this common sense database will need to be application specific. As demonstrated in the examples in Chapter 4, my system can produce sophisticated judgements with relatively little situational context. Thus, while my automated categorical imperative requires some common sense and factual background, Chapter 4 demonstrates that automating this common sense is less daunting than it seems.

My system's output could be converted into some actionable, useful response with an 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. The input parser, categorical imperative, and output parser together constitute an "ethics engine" that AI developers could use in a variety of AI systems.

The ethics engine depicted above is a high-level example of one way to use my project to guide an artifical agent, with additional work to parse the input and output of my implementation of the categorical imperative. Effectively, the kind of automated ethics I implement could be part of 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 automated ethics, which rarely captures the complexity of any ethical theory that philosophers plausibly defend.

5.2 Computational Ethics

In addition to guiding AI agents, automated ethics can also help academic philosophers make philosophical progress. Just as theorem provers make mathematics more efficient and push mathematicians to think precisely about the phenomena they model, computational ethics can help philosophers ask and answer new philosophical questions. In Section 4.1, I presented one example of the power of computational ethics by using my system to isolate a necessary and sufficient condition for the act of lying, or knowingly saying something false, to be wrong. In this section, I share another example of the kind of philosophical insight that computational ethics can prompt and analyze the value that this tool can offer to philosophers.[31] Fields from protein folding to game theory are uncovering new insights using computational tools, and computational ethics harnesses this power to inspire similar progress in philosophy.

5.2.1 Example of a Philosophical Insight: Well-Formed Maxims

As presented in Section 3.1.4, in the process of developing my formalization of the FUL, I discovered that certain kinds of maxims are badly formed, or inappropriate inputs to the universalizability test. The FUL is consistent only if it holds for "well-formed maxims," such that neither the act nor goal are already achieved in the given circumstances. Precisely, a circumstance, act, goal tuple (c, a, g) is well-formed if $(\neg(c \longrightarrow a)) \wedge (\neg(c \longrightarrow g))$, or if the circumstances do not imply the act or the goal. The insight that the FUL does not ap-

[31] I present a final example of computational ethics in Appendix C, where I resolve an ambiguity in Korsgaard's argument for the wrongness of false promising using my system (Korsgaard, 1985).

ply to badly-formed maxims has philosophical value and serves as evidence of the potential of computational ethics. In the next section, I explain and explore this insight using Kantian literature. In the following section, I demonstrate the philosophical implications of well-formed maxims by using this insight to resolve a tension between self-doubt and self-respect. The idea of well-formed maxims can guide the formulation of maxims and thus may have implications for many parts of Kantian ethics.

Badly-Formed Maxims and Kantian Ethics

Isabelle gave a logical argument for why the FUL can only hold for well-formed maxims, and I return to Kantian literature to better understand this idea. In this section, I argue that because badly-formed maxims neither change an agent's behavior nor generate meaningful obligations, they are not the right kinds of actions for practical reasoners to make moral judgements about. They cannot be action-guiding and are thus not the kind of problem that ethics should be concerned with. Moreover, under the Kantian account of the will, the very act of asking if a badly-formed maxim is prohibited generates a contradiction by undermining the will's authority over itself.

Consider the example badly-formed maxim, "When eating breakfast, I will eat breakfast in order to eat breakfast." There is something empty about this maxim because acting on it could never result in any action. If I adopt this maxim as a principle to live by, I decide that, in the circumstances "eating breakfast" I will perform the act "eating breakfast" for the purpose "eating breakfast." In these circumstances, the act has already been performed. Adopting this maxim as a law to live by does not change how I live. If I adopt this maxim, then when I am eating breakfast, I eat breakfast, but this statement is already tautologically true.

Not only does a badly-formed maxim fail to prescribe action, any obligations or prohibitions it generates have already been fulfilled or violated. If a badly-formed maxim generates a prohibition, then this prohibition is impossible to obey, which is why my original version of the FUL was inconsistent. 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.

Because badly-formed 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. Insofar as ethics is supposed to guide action, badly-formed maxims cannot be part of this project because they have no bearing on what someone should do. 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 to decide how to act. A badly-formed maxim is not the kind of maxim that a practical reasoner should consider, because it will have no bearing on what the agent should do.

Kantians can make an even stronger claim about badly-formed 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 badly-formed maxim assume that the agent has willed the maxim. Under the Kantian acount of willing, willing a maxim is equivalent to giving the maxim to yourself as a law. When you will a maxim, you commit yourself to the maxim's end. 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 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. To understand this worry, I consider the maxim, "When dancing, I should just dance for the sake of dancing." While this maxim appears to be badly-formed (the circumstance "dancing" implies the act and goal of dancing), it is a question that practical reasoners do ask. I argue that the correct interpretation of this maxim is no longer a badly-formed maxim.

Under one reading of this maxim, “I should just dance” is 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 badly-formed at all—abandoning anxiety and focusing on dancing is an entirely different act from moving your body rythmically to music. 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 concerns maxims that we 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 badly-formed. 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 want to know 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.”

Implications for Self-Doubt and Self-Respect

Above, I defined badly formed maxims, a philosophical concept that I discovered using computational ethics. In this section, I explore the implications of this concept for the ethical tension between self-doubt and self-respect to show that computational ethics can result in insights with real philosophical weight. The debate between self-doubt and self-respect originates in epistemology, which values doubting your beliefs but also rationally requires that you believe that you are not mistaken (else you should update your beliefs). I present a parallel tension between self-doubt and self-respect in ethics, where questioning your judgements is valuable but questioning a commitment as you make it is impossible. Ethical self-doubt and self-respect appear irresolvably opposed until they are understood through the lens of badly formed maxims. I argue that naive conceptions of self-doubt are badly formed maxims in disguise. If we reformulate these maxims to be well-formed, the tension between self-doubt and self-respect dissolves. I sketch the details of this argument in the rest of this section. I first introduce 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 using badly-formed maxims. I conclude that well-formed maxims form a boundary condition for the formulation of maxims, so they may resolve many debates in Kantian literature and ethics more generally.

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 a rational agent refrain from believing that they have mistaken beliefs (Christensen, 2007, 4). For example, I cannot rationally believe both that the sky is blue and that I believe that the sky is green. In other words, I cannot disapprove of my own credences, since if I do disapprove of them, I should just abandon them. 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,[32] Roush argues that the principle must be modified in order to account for healthy epistemic self-doubt. She argues that,

[32]Christensen cites Fraassen (1984), Vickers (2000), and Koons (1992).

while pathological second-guessing is correctly criticized, we are generally imperfect beings, and some sensitivity to our own limitations is a virtue (Roush, 2009, 2). 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 can be extended to ethics. When we commit ourselves to acting, we cannot simultaneously doubt the validity of our action, just as we cannot rationally disapprove of our own beliefs. If human behavior is purposive, then the very act of committing implies that one has sufficient reasons for committing. These reasons may be flawed, but in making the commitment, the reasoner accepts them—that is just what it means to commit oneself to acting. It is contradictory to claim that someone commits and questions simultaneously. 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 who is always sure that they are doing the right thing is not thinking deeply enough about their obligations. Some degree of ethical self-doubt is desirable. Thus, there is tension between the rational requirement of ESR and the intuitive validity of ethical self-doubt (ESD).

I now argue that well-formed maxims can resolve this tension. As in my earlier example, imagine that Sara is dancing at a weddding, when, in a moment of angst, she asks herself, "Should I really be dancing right now?" It seems 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. This maxim is badly-formed: the circumstance "when dancing at a friend's wedding" implies the act "dance." As I argued above, if Sara asks if she should will a badly formed maxim, she is questioning her own will's authority, which is paradoxical. If expressions of self-doubt are badly-formed maxims, then the tension between ESR and ESD is natural and unavoidable: debating the permissibility of a badly-formed maxim inherently involves questioning commitments as we make them, which is impossible. This is the source of the tension between ESR and ESD. Those committed to this interpretation must abandon one principle or the other, since simultaneous committing and questioning are incompatible.

Because understanding ethical self-doubt as a badly-formed maxim contradicts self-respect, resolving this issue requires a different 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 of dancing was 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 may 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 badly-formed maxim from above. 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. Moreover, this is a much more intuitive question to be asking *because* it is a well-formed maxim. Not only is the badly-formed interpretation of this maxim problematic for ESR, it is also not the kind of question that is useful for a practical reasoner to ask, as argued above.

Thus, the tension between ESR and ESD arises from a misreading of questions of self-doubt as questions about the evaluation of badly-formed maxims. A question of self-doubt cannot refer to a badly-formed 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 badly-formed maxims are mistaken about the maxim in question, because such a question could never yield a useful answer for a practical reasoner.

By recognizing that the naive version of ethical self-doubt is a badly formed maxim, I realized that there is something wrong with its formulation and modified it to resolve the above debate. This demonstrates a larger meta-pattern in Kantian reasoning: many debates in Kantian philosophical literature revolve around incorrectly formulated maxims, which means that the insight about badly formed maxims may have bearing on these debates. Common misconceptions about Kantian ethics[33] often result from incorrectly formulated maxims, and the entire field of applied Kantian ethics is devoted to generating the right kinds of maxims to test. Much of the work of a Kantian ethicist is formulating an appropriate maxim, and badly-formed maxims define one boundary condition for this task. Just as my isolation of the wrongness of lying in Section 4.1 could help guide the formation of morally worthy maxims, so can the insight about badly-formed maxims.

5.2.2 An Argument For Computational Ethics

The insight above is an example of the kind of philosophical progress that can be made using computational tools and serves as evidence for the power of computational ethics. The idea that the FUL can only hold for well-formed maxims would have been incredibly difficult to discover without a computer. I discovered it while formulating the FUL because Isabelle's proof-finding tools look for edge cases like badly-formed maxims. Badly-formed maxims are interesting precisely because they are the kind of thing that is usually ignored in ordinary philosophical inquiry. Philosophers usually assume that we are not discussing badly-formed maxims because, as argued above, they are not the kind of thing that is immediately relevant to ethics. Computational tools like Isabelle require that assumptions like the exclusion of well-formed maxims are made precise, and thus force philosophers to understand their arguments in a new way. The philosophical insights about lying in Section 4.1 and about well-formed maxims above demonstrate the contributions that computational ethics makes: it can quickly check edge cases and it uncovers imprecise, ambiguous, or implicit assumptions.

I do not argue that computational ethics uncovers philosophical insights that humans are always incapable of reaching. In some cases, computational ethics may automate calculations too long and tedious for any human philosopher to complete. Insights like the one

[33]For example, critics of Kantian ethics worry that the maxim, "When I am a man, I will marry a man because I want to spend my life with him" fails the universalizability test because if all men only married men, sexual reproduction would stop. This argument implies that Kantian ethics is homophobic. Kantians often respond by arguing that the correct formulation of this maxim is, "When I love a man, I will marry him because I want to spend my life with him," which is universalizable because if everyone marries who they love, some men will marry women and others will marry men.

about well-formed maxims, however, could have been discovered by a human being but are much easier to reach with computational tools. Computational tools prompt philosophers to ask new questions that lead to insights, and can thus serve as another tool in a philosopher's arsenal, like a thought experiment or counterexample.

The first benefit of computational ethics is precision, which is the goal of much analytic philosophy. 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 the goal of precision. Representing a philosophical idea in logic and implementing it in an interactive theorem prover requires making the idea precise to a degree that most philosophical conversation does not necessarily require. For example, when formalizing the notion of a maxim, I had to understand its components, define it as a circumstance, act, goal tuple, and identify coherent and consistent types for each of these entities. This precision is also evident in my examination of lying and joking in Section 4.1, where I isolate the specific cause of lying's wrongness. This insight could contribute to the rich philosophical literature on deception, and demonstrates the philosophical signifiance of precision. This level of precision is possible without computational tools, but computational ethics forces it. Type fuzziness and overloaded definitions are all too common in philosophical writing and discussion, but computers disallow this kind of imprecision.

Another benefit of computational ethics is that it makes certain kinds of ethical inquiry, such as searching for counterexamples or formal ethics, far less tedious. For example, Nitpick can refute an ethical statement in seconds by using brute force to construct a counterexample, something that can require hours of thought and discussion. I arrived at the insight about badly-formed maxims because Isabelle can check edge cases, like that of the badly-formed maxim, far more quickly than a human being. Moreover, subfields that use symbolic logic to represent philosophical concepts (e.g. philosophy of language) can use interactive theorem provers like Isabelle to complete proofs in a matter of seconds. By automating away the tedium, computational ethics can give philosophers the tools to ask new kinds of questions.

Computational ethics is at its infancy. The use of theorem provers in mathematics is only 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 to automate deontic logic occurred in the last few years. The fact that this nascent technology is already helping humans reach non-trivial ethical conclusions is reason to, at the very least, entertain the possibility of a future where computational ethics becomes as normal for philosophers as using a thought experiment.

To the skeptic, 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 moving from basic algebra to Fields medal winning mathematics became 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 expectations. 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 could become as commonplace in philosophy departments as reflective equilibrium. Just as computational tools have amplified progress in healthcare and drug discovery, computational ethics has the potential to enable great philosophical progress.

5.3 Automating Everyday Practical Reason

In Sections 5.1 and 5.2, I outline how automated ethics can guide artificial agents and philosophers respectively. This raises a natural question: can automated ethics guide ordinary human beings, not just academic philosophers, as we navigate the world and face ethical dilemmas? Some may hope (or worry) that automated ethics could render ethical reasoning obsolete. In this section, I argue that while computers should not replace human ethical reasoning entirely, they can supplement and improve our ethical reasoning. I argue for a kind of human-computer symbiosis in which computers offer ethical advice, arguments for particular moral judgements, and speed up moral calculations without subverting human ethical reasoning entirely (Licklider, 1960).

Ethics bears weight for everyone, not just for academic philosophers, because it studies the unavoidable question: how should we live? The ethical question is the only question that we answer merely by living. To turn away from ethics is to take a stance on the question of how to live (namely, to live unreflectively) and thus to engage in ethics. Every rational being must decide how to navigate the world and ethics answers this question. Given that ethics is vital, it seems that if comptuational tools can help us derive ethical judgements more efficiently, then everyone should engage in computational ethics. In the most extreme case, we can unthinkingly follow the commands of an ethical calculator that dictates how we should live. Maybe computers can answer the unavoidable question for us.

The argument above places the value of ethics solely in its action-guiding potential, and thus fails to take into account the importance of practical reason, which, as I argued in Section 2.1.3, is the source of freedom itself. We are committed to ethical reflection because of the kind of beings that we are. Recall that Korsgaard argues that, as beings occupying minds with a reflective structure, when faced with a choice,"it is as if there were something over and above all of your desires, something that is you, and that chooses which desire to act on" (Korsgaard and O'Neill, 1996, 83). This choosing is the operation of practical reason, and this reflection makes us free. We are free because we can choose which reasons to act on.

If reflection makes us free, then unthinkingly obeying a computer sacrifices our autonomy. Consider an Ethics Oracle that can unfailingly tell you the right thing to do in any situation.[34] Someone who surrenders themselves to this Oracle unthinkingly follows its prescriptions. The reflection involved in the decision to obey each of the Oracle's prescriptions is limited (Bok, 1998). This person is not reflecting on the real matters at hand and is not making decisions for themselves. They have surrendered their reflective capacity to the Oracle. They

[34]This example is inspired by the Pocket Oracle presented in Bok (1998). Unlike the Ethics Oracle, Bok's Pocket Oracle perfectly predicts what we *will* do, not what we *should* do.

live a worse life than someone who reflects on their actions; they have less ownership over their actions than the reflective person. In a less extreme case, a person may retain control of many of their decisions but cede some important or tricky choices to the Ethics Oracle. Because every single exercise of practical reason is an exercise of autonomy, this person is still less autonomous than the purely reflective person. Even surrendering simple, inconsequential decisions such as which flavor of coffee to drink surrenders some piece of our autonomy. Perhaps in trivial cases we can accept that tiny sacrifice, but giving over life-changing decisions to the machine sacrifices our core freedom. Unreflectively relying on computational ethics surrenders our autonomy to the machine.

One objection to this emphasis on reflection is the impracticality of making ethical calculations from first principles every time we are faced with a decision. This is why we follow the advice of moral mentors, like our family or influential philosophers. Most people do not reason about ethics during everyday decisions; they rely on some combination of prior knowledge and external testimony. For example, my mother taught me to respect myself, so I follow her advice.

What is the difference between following the guidance of a moral educator and obeying the Ethics Oracle? The best kind of ethical advice prompts reflection, such as an argument made in a philosophy paper. Unthinkingly following someone's advice results in the same loss of autonomy as unthinkingly obeying the Ethics Oracle; people who merely obey orders are less autonomous than those who think for themselves. This account of moral advice offers a model for human-computer symbiotic computational ethics. The computer should serve as a moral guide by providing arguments, just as my mom explained why I should always respect myself. Human-computer symbiotic computational ethics nurtures autonomy when it not only offers prescriptions for action, but also explanations for these prescriptions. Because my theorem-prover-based automated ethical system is explainable, it can guide action without sacrificing autonomy. It can make an argument for some action, instead of merely giving a verdict. Isabelle can list the facts used to show a partcular action prohibited, and a human being can reflect on whether or not these principles indeed prohibit the action in question.

The computer serves as a collaborator and a tool, but not as an authority, so the human being's reflective capacity and freedom is preserved.

The above model of human-computer symbiosis demonstrates how computational ethics can augment human ethical reasoning without replacing it. When deliberating over moral dilemmas, ordinary people can turn to computational tools for advice, like an "Ask an Ethicist" column. If we appeal to philosophically faithful computational ethics like that implemented in this thesis, then this advice will synthesize decades of philosophical progress and is thus a way to apply philosophers' insights to ordinary life. Moreover, just as they do for philosophers, computers can help ordinary people approach ethical questions from a different perspective. Even interacting with my system requires the user to consider the action's maxim, which includes the circumstances, act, and goal. Making these components of action precise already changes the user's perspective. Just as computational ethics can serve as a tool for academic philosophers to ask new questions and achieve greater precision, it can do the same for ordinary human beings navigating the world. Moreover, it also offers another way for the general public to access professional philosophy insights, and thus carries potential to improve our everyday reasoning.

5.4 Theoretical Objections to Automating Kantian Ethics

Many philosophers cringe at the idea that a computer could perform ethical reasoning or that the categorical imperative could provide an algorithm for moral judgement. For example, Rawls asserts, "it is a serious misconception to think of the CI-procedure as an algorithm intended to yield, more or less mechanically, a correct judgment. There is no such algorithm, and Kant knows this" (Rawls, 2000, 166). Ebels-Duggan also claims, "no one supposes that the Categorical Imperative provides a mechanical algorithm that delivers all by itself a complete account of what we ought to do in any given situation" (Ebels-Duggan, 2012, 174). However unmechanical ethical reasoning may seem, these claims are not obvious and require further justification. Philosophers who believe that mental activity completely determines moral reasoning must explain why computers can, in theory, simulate certain mental processes like arithmetic and language, but cannot perform ethical reasoning. Without a soul or God-based account of ethical reasoning, it is not obvious that it is theoretically impossible to automate ethical reasoning. After all, computers may eventually learn to simulate human mental activity entirely, as shown by progress in brain simulation (Yamazaki et al., 2021). Skeptics about automated ethics must explain why ethics is any different from other automated human activity.

The above claims represent the general view that automating ethics is impossible. In the rest of this section, I explore specific arguments within this view. First, I consider the argument that machines cannot have the necessary motivations or attitude towards action to behave morally. Second, I consider the difficulty of formulating an input to the categorical imperative. Third, I consider the strongest objection to automated ethics, that moral judgement requires prior ethical knowledge and intuition that machines lack. I argue that these objections do not render automated Kantian ethics impossible, but merely difficult. Rawls may be correct that there is no simple algorithm for ethical reasoning, but, as I demonstrate in this thesis, moral judgement using the FUL can be automated using not one algorithm, but the many algorithms that constitute my system.

Machines, Morality, and Motivation

In *Universal Laws and Ends In Themselves*, O'Neill argues against the existence of an algorithm for moral behavior. She points out that Kant draws an important distinction between a morally worthy maxim and a morally worthy action: the latter requires a good will, or a will motivated by duty (O'Neill, 1989, 345). Moral behavior doesn't just require performing a "good" action, but it requires acting on a morally worthy maxim from the motivation of duty, or doing the right thing because it is the right thing to do. It is this capacity for self-motivation that makes morality binding for rational beings; we must behave morally precisely because we have wills, or the ability to be motivated by ends that we choose. Only rational beings have wills, so only rational beings can have good wills, so only rational beings can behave morally. Under this understanding of moral behavior, it seems unlikely that a computer could behave morally since a computer does not have motivation in the same way as a human being.[35]

The idea that a computer cannot behave morally does not preclude the kind of automated categorical imperative that I present in this thesis. O'Neill argues that the FUL serves as a test of morally worthy maxims, and a implementation of an automated categorical imperative can identify this kind of maxim. Perhaps a computer cannot act on a morally relevant maxim from the motivation of duty, but it certainly can act on this maxim nonetheless. For example, a self-driving car can choose to swerve to hit a tree to avoid injuring pedestrians in the crosswalk. This action may be one that acts on a morally worthy maxim *even if* the self-driving car is not motivated by duty. The discpline of machine ethics is spurred by the recognition that, as automated agents become more powerful, they will need to make morally consequential decisions. Automated agents may be incapable of moral motivation, but automated agents that mimic moral behavior are better than agents that ignore morality entirely. AI agents are navigating a world inhabited by human beings, and their decisions impact us. Insofar as AI will operate in human society, the behavior of such AI should mimic the behavior of an ethical human being for our sakes, so that we can interact with it safely.

Inputs to the Categorical Imperative

Another challenge for automated Kantian ethics is that the FUL test requires that a maxim be given as input. O'Neill notes that the test assumes "that agents will have certain tentative plans, proposals and policies which they can consider, revise or reject or endorse and pursue" (O'Neill, 1989, 343). The FUL evaluates the moral worth of a maxim given as input, where this potential maxim is generated by the choices that an agent is faced with. As I argue in Section 5.1, determining this potential maxim is a challenge for both human and automated reasoners. Kant even claims that the difficulty of determining an agent's potential maxim, which is their own, subjective understanding of their principle of action, is a reason that we may never be able to know if the morally worthy action has been performed (O'Neill, 1989, 345). Reasoners are faced with choices between potential actions and must determine the maxim, or principle, underlying each potential action. This is equivalent to a "mapping"

[35] A parallel argument can also be made for virtue ethics. Virtuous behavior requires not only a certain action, but also a certain disposition towards the action, so it seems difficult for an AI agent to truly behave virtuously.

problem: agents are given situations or dilemmas as input and must map these to maxims.

The challenge of mapping actions to maxims is a limitation of my system, but it is not insurmountable. In Section 5.1, I argued that, before my system can be used in practice, it must be paired with an input parser that can translate choices that an automated agent faces into maxims into a logic that my system can evaluate. This need follows from the difficulty in mapping a potential action to the maxim of action, whether concerning human action or machine action. As argued in Section 5.1, human-in-the-loop or heuristic-based approaches could resolve this issue. Determining the maxim of action is a challenge for Kantian human beings (O'Neill, 1989), so it is unsurprising that it is a major hurdle for automated Kantian agents. Overcoming this hurdle is not impossible, and progress in automated ethics could address this concern.

Prior Moral Knowledge

As one of the strongest arguments against a categorical imperative algorithm, O'Neill argues that the FUL is not supposed to provide a mechanism for deriving all morally worthy maxims from scratch. She notes that "we usually already have learnt or worked out the moral standing of many common maxims of duty," and so approach moral deliberation with an "almanac" of morally worthy and empty maxims (O'Neill, 1989, 394). Rational agents navigating the world rarely recalculate the moral status of each potential maxim of action; instead, we consult our almanac of maxims. This almanac is generated by moral education, absorbed social values, and moral advice from people we trust. The categorical imperative is useful to verify the rightness or wrongness of a maxim, but is not part of the bulk of human ethical reasoning.

While human beings cannot repeatedly apply the universalizability test to all potential maxims encountered during a moral dilemma, computers have the computational power to do so. Human beings are equipped with enough prior knowledge or common sense, to have an almanac of morally worthy maxims, but we have limited computational power. Computers, on the other hand, are comparatively much more capable of computation and thus can repeatedly recompute the results of the categorical imperative test. They do not come equipped with an almanac of maxims, but can simply recompute this almanac every time they need to make a decision. Human beings use common sense to make up for their computational limitations, and automated moral agents can use computational power to reduce the need for common sense.

Daniela Tafani takes O'Neill's argument one step further by arguing that this "alamnac" of maxims already includes the moral status of the maxims in questions; human beings already know which maxims are morally worthy and which are morally lacking. The categorical imperative test merely reminds us, in moments of weakness, when we are tempted to make an exception to the moral law for our own convenience or pleasure, that the moral law has no exceptions (Tafani, 2021, 9). Thus, she claims that "the Kantian test is therefore as useless for machines as it is for anyone who does not already know what to do" (Tafani, 2021, 8).[36] Understanding the categorical imperative test as a reminder instead of a derivation tool also explains the response to the tailor objection presented in Section 5.1, that the FUL cannot handle bad-faith attempts to generate false positives or negatives. The test only returns the right result when an agent sincerely attempts to represent their maxim of action, not when an adversary attempts to "trick" the categorical imperative because such tricks will fall outside the scope of our moral almanacs.

Under Tafani's understanding of the categorical imperative, not only is automated moral reasoning possible, but the challenge of creating an input parser or automatically formulating a maxim becomes easier as well. If the categorical imperative test is only useful to those who have some prior moral knowledge, then prior moral knowledge can and should be used to create an input parser. Specifically, a machine learning-based approach could learn action-maxim mappings from a database of such mappings compiled by a human being. Moreover, the human being could assign each maxim in the database a rightness or wrongness score. My implementation of the automated categorical imperative would then simply check the work of this machine learning algorithm and transform a fuzzy prediction into a provable, rigorous moral judgement. This rigorous moral judgement could in turn be fed into the database of

[36] Translated using Google Translate.

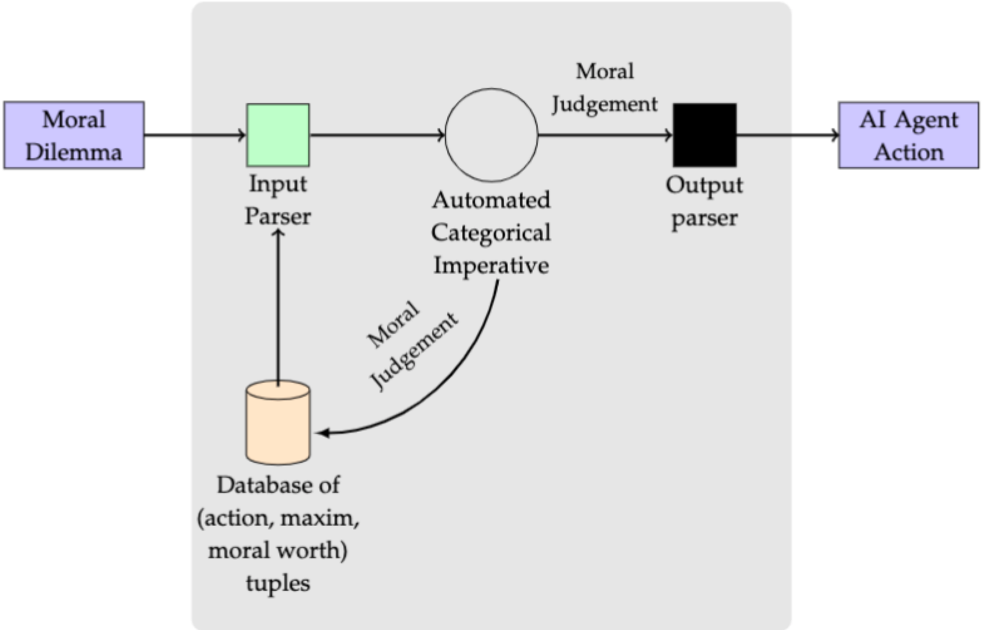

Figure 3: A refined version of Figure 2 in which the input parser learns from a database of action-maxim mappings, which is in turn fed the output of my automated categorical imperative.

maxims to make the intput parser smarter. One example of this kind of system is shown in Figure 3. The combination of prior knowledge of some maxims' moral worth and the ability of a computer to constantly perform the universalizability test could not only match human ethical reasoning but could perhaps surpass it by double checking the moral intuitions that we take for granted. A computer with no common sense or prior knowledge may be unable to reason using the categorical imperative, but one equipped with some prior knowledge of maxims and their moral worth may even be able to reason about morality better than human beings can.

5.5 Related Work

In 1685, Leibniz dreamed of a calculator that could resolve philosophical and theological disputes (Leibniz, 1679). At the time, the logical and computational resources necessary to make his dream a reality did not exist. Today, automated ethics is a growing field, spurred in part by the need for ethically intelligent AI agents. Tolmeijer et al. surveyed the state of the field of machine ethics (Tolmeijer et al., 2021) and characterized implementations in automated ethics by (1) the choice of ethical theory, (2) implementation design decisions (e.g. logic programming), and (3) implementation details (e.g. choice of logic).

Automated ethics falls into two branches: top-down and bottom-up ethics. Top-down automated ethics begins with an ethical theory, whereas bottom-up automated ethics learns ethical judgements from prior judgements. One example of bottom-up automated ethics is Delphi, which uses deep learning to make ethical judgements based on a dataset of human judgements (Jiang et al., 2021). While Delphi displays great flexibility, it often produces contradictory judgements, such as claiming that taxing exploitative profitable companies is good, but burdening successful companies with high tax rates is bad (Vincent, 2021). Because Delphi draws on error-prone human judgements instead of philosophical literature, it makes the same judgement errors that humans make. Moreover, because Delphi uses a bottom-up approach, there is no explicit ethical theory explaining its judgements, so analytically arguing for or against its conclusions is impossible. Top-down approaches, on the other hand, must be explicit about the underlying ethical theories, and are thus more explainable.

In this paper, I use a top-down approach to formalize Kantian ethics. There is a long line of work automating other ethical theories, like consequentialism (Abel et al., 2016; Anderson et al., 2004) or particularism (Ashley and McLaren, 1994; Guarini, 2006). I choose to implement Kantian ethics because, as argued in Section 2.1, it is the most formal and least data-intensive of the three major ethical traditions. Kantian ethics is a deontological, or rule based ethic, and there is prior work implementing other deontological theories (Govindarajulu and Bringsjord, 2017; Anderson and Anderson, 2008, 2014).

Kantian ethics specifically appears to be an intuitive candidate for formalization and implementation and there has been both theoretical and practical work on automating Kantian ethics (Powers, 2006; Lin et al., 2012). In 2006, Powers argued that implementing Kantian ethics presented technical challenges, such as automation of a non-monotonic logic, and philosophical challenges, like a definition of the categorical imperative (Powers, 2006). I address the former through my use of Dyadic Deontic Logic, which allows obligations to be retracted as context changes, and the latter through my use of the practical contradiction interpretation. There has also been prior work in formalizing Kantian metaphysics using I/O logic (Stephenson et al., 2019). Deontic logic, which has been implemented in Isabelle/HOL, is itself inspired by Kant's "ought implies can" principle, but it does not include a robust formalization of the entire categorical imperative (Cresswell and Hughes, 1996).

Kroy presented a formalization of the first two formulations of the categorical imperative, but wrote before the computational tools existed to automate such a formalization (Kroy, 1976). I implement his formalization of the FUL to compare it to my system. Lindner and Bentzen presented one of the first implementations of a formalization of Kant's second formulation of the categorical imperative (Bentzen and Lindner, 2018). They present their goal as "not to get close to a correct interpretation of Kant, but to show that our interpretation of Kant's ideas can contribute to the development of machine ethics." My work builds on theirs by formalizing the first formulation of the categorical imperative as faithfully as possible. Staying faithful to philosophical literature makes my system capable of making robust and reliable judgements.

The implementation of this paper was inspired by and builds on Benzmüller, Parent, and Farjami's foundational work with the LogiKEy framework for machine ethics, which includes their implementation of DDL in Isabelle (Benzmüller et al., 2021; Benzmüller et al., 2019). The LogiKEy project has been used to study metaphysics (Benzmüller and Paleo, 2013; Kirchner et al., 2019), law (Zahoransky and Benzmüller, 2020), and ethics (Fuenmayor and Benzmüller, 2018), but not Kant's categorical imperative.

5.6 Conclusion

In this thesis, I present a proof-of-concept implementation of automated Kantian ethics. My system takes as input a potential action, appropriately represented, and can prove that it is obligated, prohibited or permissible. I represent Kant's Formula of Universal Law in a deontic logic and implement this logic in the Isabelle/HOL interactive theorem prover, which can automatically prove or refute theorems in my custom logic. I also contribute a testing framework that demonstrates that my implementation of Kantian ethics is more faithful to philosophical literature than two other potential implementations. My completed system can, when given appropriate factual background, make philosophically mature judgements about complex moral dilemmas. This work is one step towards building morally sophisticated artifical agents.

The idea of fully automated artificial agents navigating the world without human supervision may be terrifying, but progress in AI indicates that such a future is likely closer than we think. Philosophers, regulators, and computer scientists are sounding the alarm about the dangers of developing this kind of AI. Insofar as developers will continue to ignore these warnings and develop increasingly independent AI, there is a dire need to program such AI with some notion of ethics. If AI is navigating human society, then it is making ethically-tinged decisions at all times. Ethics is inescapable; if AI developers and computer scientists ignore it, then they will be building machines that make decisions based on some set of unknown, implicit ethical values. Countless examples, from the Alleghany family services screening algorithm that is biased against poor families to search algorithms that associate black-sounding names with crime, demonstrate that such implicit ethics usually codifies the biases, prejudices, and moral failings of the society in which it is developed (Eubanks, 2018; Sweeney, 2013). AI will inevitably make judgements on moral dilemmas, and automated ethics is necessary to make these judgements morally correct.

Given that the discpline of philosophy has spent centuries debating such judgements and their theoretical underpinnings, such AI will be most trustworthy, nuanced, consistent, and mature when it is faithful to philosophical literature. In order to develop high-quality automated ethics, computer scientists and philosophers must work together. Neither discipline alone can address the pressing need for ethical AI. This thesis is an experiment in marrying philosophy and computer science to create automated ethics that is both technically and philosophically advanced.

This work is an early proof-of-concept. It demonstrates the potential of top-down, logic programming approaches to automated ethics and shows that it is possible to faithfully automate an ethical theory as complex as Kantian ethics. There are open questions that must be resolved before a system like this could be used in practice, but this project demonstrates that these questions are within closer reach than they may seem. Automated ethics does not need to limit itself to simple, flattened versions of ethical theories. With technical and philosophical progress, faithful automated ethics is possible. Growing public consciousness about the dangers of unregulated AI is creating momentum in machine ethics; work like Delphi demonstrates that the time is ripe to create usable, reliable automated ethics. This thesis is one step towards building computers that can think ethically in the richest sense of the word.

References

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

Alexander, L. and Moore, M. (2021), Deontological Ethics, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2021 edn, Metaphysics Research Lab, Stanford University.

Anderson, M. and Anderson, S. (2014), 'Geneth: A general ethical dilemma analyzer', *Proceedings of the National Conference on Artificial Intelligence* **1**.

Anderson, M., Anderson, S. and Armen, C. (2004), 'Towards machine ethics'.

Anderson, M. and Anderson, S. L. (2008), Ethel: Toward a principled ethical eldercare system, *in* 'AAAI Fall Symposium: AI in Eldercare: New Solutions to Old Problems'.

Aristotle (1951), 'The nicomachean ethics', *Journal of Hellenic Studies* **77**, 172.

Arkoudas, K., Bringsjord, S. and Bello, P. (2005), 'Toward ethical robots via mechanized deontic logic', *AAAI Fall Symposium - Technical Report* .

Ashley, K. D. and McLaren, B. M. (1994), A cbr knowledge representation for practical ethics, *in* 'Selected Papers from the Second European Workshop on Advances in Case-Based Reasoning', EWCBR '94, Springer-Verlag, Berlin, Heidelberg, p. 181–197.

Awad, E., Dsouza, S., Shariff, A., Rahwan, I. and Bonnefon, J.-F. (2020), 'Universals and variations in moral decisions made in 42 countries by 70,000 participants', *Proceedings of the National Academy of Sciences* **117**(5), 2332–2337.
URL: *https://www.pnas.org/content/117/5/2332*

Baltag, A. and Renne, B. (2016), Dynamic Epistemic Logic, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2016 edn, Metaphysics Research Lab, Stanford University.

Bentzen, M. M. and Lindner, F. (2018), 'A formalization of kant's second formulation of the categorical imperative', *CoRR* **abs/1801.03160**.
URL: *http://arxiv.org/abs/1801.03160*

Benzmüller, C., Farjami, A. and Parent, X. (2021), Dyadic deontic logic in hol: Faithful embedding and meta-theoretical experiments, *in* M. Armgardt, H. C. Nordtveit Kvernenes and S. Rahman, eds, 'New Developments in Legal Reasoning and Logic: From Ancient Law to Modern Legal Systems', Vol. 23 of *Logic, Argumentation & Reasoning*, Springer Nature Switzerland AG.

Benzmüller, C. and Paleo, B. W. (2013), 'Formalization, mechanization and automation of gödel's proof of god's existence', *CoRR* **abs/1308.4526**.
URL: *http://arxiv.org/abs/1308.4526*

Benzmüller, C., Parent, X. and van der Torre, L. W. N. (2019), 'Designing normative theories of ethical reasoning: Formal framework, methodology, and tool support', *CoRR* **abs/1903.10187**.
URL: *http://arxiv.org/abs/1903.10187*

Berberich, N. and Diepold, K. (2018), 'The virtuous machine: Old ethics for new technology?', *CoRR* **abs/1806.10322**.
URL: *http://arxiv.org/abs/1806.10322*

Blanchette, J. C. and Nipkow, T. (2010), *Nitpick: A Counterexample Generator for Higher-Order Logic Based on a Relational Model Finder*, Vol. 6172, Springer Berlin Heidelberg, p. 131–146.
URL: *http://link.springer.com/10.1007/978-3-642-14052-5_11*

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

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

Carmo, J. and Jones, A. (2013), 'Completeness and decidability results for a logic of contrary-to-duty conditionals', *Journal of Logical Computation* **23**, 585–626.

Cervantes, J.-A., Rodríguez, L.-F., López, S. and Ramos, F. (2013), A biologically inspired computational model of moral decision making for autonomous agents, *in* '2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing', pp. 111–117.

Chisholm, R. M. (1963), 'Contrary-to-duty imperatives and deontic logic', *Analysis (Oxford)* **24**(2), 33–36.

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

Cloos, C. (2005), 'The utilibot project: An autonomous mobile robot based on utilitarianism', *AAAI Fall Symposium - Technical Report* .

Cresswell, M. J. and Hughes, G. E. (1996), *A New Introduction to Modal Logic*, Routledge.

Davenport, D. (2014), 'Moral mechanisms', *Philosophy and Technology* **27**(1), 47–60.

Dennis, L., Fisher, M., Slavkovik, M. and Webster, M. (2016), 'Formal verification of ethical choices in autonomous systems', *Robotics and Autonomous Systems* **77**, 1–14.
URL: *https://www.sciencedirect.com/science/article/pii/S0921889015003000*

Dietrichson, P. (1964), 'When is a maxim fully universalizable?', *Kant Studien* **55**(1-4), 143–170.
URL: *https://doi.org/10.1515/kant.1964.55.1-4.143*

Driver, J. (2014), The History of Utilitarianism, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2014 edn, Metaphysics Research Lab, Stanford University.

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

Eubanks, V. (2018), *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St. Martin's Press.

Foot, P. (1967), 'The problem of abortion and the doctrine of the double effect', *Oxford Review* **5**, 5–15.

Fraassen, B. C. V. (1984), 'Belief and the will', *Journal of Philosophy* **81**(5), 235–256.

Fuenmayor, D. and Benzmüller, C. (2018), 'Formalisation and evaluation of alan gewirth's proof for the principle of generic consistency in isabelle/hol', *Archive of Formal Proofs* . https://isa-afp.org/entries/GewirthPGCProof.html, Formal proof development.

Gabriel, I. (2020), 'Artificial intelligence, values, and alignment', *Minds and Machines* **30**(3), 411–437.
URL: *http://dx.doi.org/10.1007/s11023-020-09539-2*

Govindarajulu, N. S. and Bringsjord, S. (2017), 'On automating the doctrine of double effect', *CoRR* **abs/1703.08922**.
URL: *http://arxiv.org/abs/1703.08922*

Guarini, M. (2006), 'Particularism and the classification and reclassification of moral cases', *IEEE Intelligent Systems* **21**(4), 22–28.

Harrison, J., Urban, J. and Wiedijk, F. (2014), History of interactive theorem proving, *in* 'Computational Logic'.

Hintikka, J. (1962), *Knowledge and Belief*, Cornell University Press.

Hume, D. (2007), *An Enquiry Concerning Human Understanding and Other Writings*, Cambridge University Press.

Hursthouse, R. and Pettigrove, G. (2018), Virtue Ethics, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2018 edn, Metaphysics Research Lab, Stanford University.

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, R. and Cureton, A. (2021), Kant's Moral Philosophy, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Spring 2021 edn, Metaphysics Research Lab, Stanford University.

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

Kant, I. (2017), *Introduction*, Cambridge Texts in the History of Philosophy, 2 edn, Cambridge University Press, pp. ix–xxix.

Kemp, J. (1958), 'Kant's examples of the categorical imperative', *The Philosophical Quarterly (1950-)* **8**(30), 63–71.
URL: *http://www.jstor.org/stable/2216857*

Kirchner, D., Benzmüller, C. and Zalta, E. N. (2019), 'Computer science and metaphysics: A cross-fertilization', *CoRR* **abs/1905.00787**.
URL: *http://arxiv.org/abs/1905.00787*

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

Kitcher, P. (2004), 'Kant's argument for the categorical imperative.', *Nous* **38**.

Kohl, M. (2015), 'Kant and 'ought implies can'', *The Philosophical Quarterly (1950-)* **65**(261), 690–710.
URL: *http://www.jstor.org/stable/24672780*

Koons, R. C. (1992), *Paradoxes of Belief and Strategic Rationality*, Cambridge Studies in Probability, Induction and Decision Theory, Cambridge University Press.

Korsgaard, C. (1985), 'Kant's Formula of Universal Law', *Pacific Philosophical Quarterly* **66**, 24–47.

Korsgaard, C. (1986), 'The Right to Lie: Kant on Dealing with Evil', *Philosophy and Public Affairs* **15**, 325–249.

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

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

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

Kroy, M. (1976), 'A partial formalization of kant's categorical imperative. an application of deontic logic to classical moral philosophy', *Kant-Studien* **67**(1-4), 192–209.
URL: *https://doi.org/10.1515/kant.1976.67.1-4.192*

Leibniz, G. W. (1679), On universal synthesis and analysis, or the art of discovery and judgment: 1679(?), *in* 'Philosophical Papers and Letters', The New Synthese Historical Library, Springer Netherlands, Dordrecht, pp. 229–234.

Lewis, D. (1973), 'Causation', *Journal of Philosophy* **70**(17), 556–567.

Licklider, J. C. R. (1960), 'Man-computer symbiosis', *IRE Transactions on Human Factors in Electronics* **HFE-1**(1), 4–11.

Lin, P., Abney, K. and Bekey, G. A. (2012), *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, MIT Press, pp. 35–52.

Lukowicz, P. (2019), 'The challenge of human centric ai', *Digitale Welt* **3**, 9–10.

McNamara, P. and Van De Putte, F. (2021), Deontic Logic, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2021 edn, Metaphysics Research Lab, Stanford University.

McRae, E. (2013), 'Equanimity and intimacy: A buddhist-feminist approach to the elimination of bias', *Sophia* **52**(3), 447–462.

Montague, R. (1970), 'Universal grammar', *Theoria* **36**(3), 373–398.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-2567.1970.tb00434.x*

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

Nipkow, T., Paulson, L. C. and Wenzel, M. (2002), *Isabelle/HOL: A Proof Assistant for Higher Order Logic*, Springer-Verlag Berlin Heidelberg, Berlin.

O'Neill, O. (1989), 'Universal laws and ends-in-themselves', *The Monist* **72**(3), 341–361.
URL: *http://www.jstor.org/stable/27903145*

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

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

Paulson, L. and Blanchette, J. (2015), 'Three years of experience with sledgehammer, a practical link between automatic and interactive theorem provers', *Proceedings of the 8th International Workshop on the Implementation of Logics* .

Paulson, L. C. (1999), 'A generic tableau prover and its integration with isabelle', *J. Univers. Comput. Sci.* **5**, 73–87.

Powers, T. M. (2006), 'Prospects for a kantian machine', *IEEE Intelligent Systems* **21**(4), 46–51.

Puiutta, E. and Veith, E. M. (2020), 'Explainable reinforcement learning: A survey'.

Rawls, J. (1980), 'Kantian constructivism in moral theory', *The Journal of Philosophy* **77**(9), 515–572.
URL: *http://www.jstor.org/stable/2025790*

Rawls, J. (2000), 'Lectures on the history of moral philosophy', *Critica* **35**(104), 121–145.

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

Rönnedal, D. (2019), 'Contrary-to-duty paradoxes and counterfactual deontic logic', *Philosophia* **47**.

Scott, D. (1970), Advice on modal logic, *in* K. Lambert, ed., 'Philosophical Problems in Logic: Some Recent Developments', D. Reidel, pp. 143–173.

Silber, J. R. (1974), 'Procedural formalism in kant's ethics', *The Review of Metaphysics* **28**(2), 197–236.
URL: *http://www.jstor.org/stable/20126622*

Sinnott-Armstrong, W. (2021), Consequentialism, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2021 edn, Metaphysics Research Lab, Stanford University.

Solt, K. (1984), 'Deontic alternative worlds and the truth-value of 'oa'', *Logique et Analyse* **27**(107), 349–351.
URL: *http://www.jstor.org/stable/44084096*

Stang, N. F. (2021), Kant's Transcendental Idealism, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Spring 2021 edn, Metaphysics Research Lab, Stanford University.

Stephenson, A., Sergot, M. and Evans, R. (2019), 'Formalizing kant's rules: A logic of conditional imperatives and permissives', *Journal of Philosophical Logic* **49**.
URL: *https://eprints.soton.ac.uk/432344/*

Sweeney, L. (2013), 'Discrimination in online ad delivery'.

Tafani, D. (2021), 'Imperativo categorico come algoritmo. kant e l'etica delle macchine', *Sistemi intelligenti, Rivista quadrimestrale di scienze cognitive e di intelligenza artificiale* pp. 377–392.
URL: *https://www.rivisteweb.it/doi/10.1422/101195*

Timmermann, J. (2013), 'Kantian dilemmas? moral conflict in kant's ethical theory', *Archiv für Geschichte der Philosophie* **95**(1), 36–64.
URL: *https://doi.org/10.1515/agph-2013-0002*

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M. and Bernstein, A. (2021), 'Implementations in machine ethics', *ACM Computing Surveys* **53**(6), 1–38.
URL: *http://dx.doi.org/10.1145/3419633*

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

Vickers, J. (2000), 'I believe it, but soon i'll not believe it any more: Scepticism, empiricism, and reflection', *Synthese* **124**, 155–174.

Vincent, J. (2021), 'The ai oracle of delphi uses the problems of reddit to offer dubious moral advice'.

Wallach, W. and Allen, C. (2008), *Moral Machines: Teaching Robots Right From Wrong*, Oxford University Press.

Winfield, A., Blum, C. and Liu, W. (2014), 'Towards an ethical robot: Internal models, consequences and ethical action selection', *Conference Towards Autonomous Robotic Systems* **8717**.

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

Yamazaki, T., Igarashi, J. and Yamaura, H. (2021), 'Human-scale brain simulation via supercomputer: A case study on the cerebellum', *Neuroscience* **462**, 235–246. In Memoriam: Masao Ito—A Visionary Neuroscientist with a Passion for the Cerebellum.
URL: *https://www.sciencedirect.com/science/article/pii/S030645222100021X*

Zahoransky, V. and Benzmüller, C. (2020), 'Modelling the us constitution to establish constitutional dictatorship'.

A Alternate Definitions of a Maxim

A.1 Korsgaard's Act-Goal View

I adopt O'Niell's definition of a maxim, which builds on Korsgaard's weaker interpretation of a maxim as an act, goal pair. She interprets Kant's example meanings as having the form "to-do-this-act-for- the-sake-of-this-end," which could be formalized as a pair of an act and goal (Korsgaard, 2005). For example, under this view, one example maxim might be, "Falsely promise to repay a loan in order to get some easy cash."

O'Niell's view only differs from this view in the inclusion of the circumstances on which the agent acts. This inclusion creates a representation of a maxim that is strictly more expressive than Korsgaard's interpretation; every (circumstance, act, goal) tuple can be represented as an (act, goal) pair by simply dropping the circumstances, but the same (act, goal) pair could correspond to many different (circumstance, act, goal) tuples, all with varying moral statuses. Because my representation of a maxim is more expressive than Korsgaard's, my results are stronger than those that would be achieved with Korsgaard's view. Thus, proponents of Korsgaard's view could simply ignore the circumstances in my representation of a maxim and still achieve their desired results.

One issue with Korsgaard's view is that an actionable maxim will necessarily require some circumstances built-in because the agent will need to know when to act on the maxim. For example, the falsely promising maxim bakes in the circumstances that the actor has access to lender, needs money, and that the lender will expect their money back. At an even more granular level, this maxim implicitly includes a definition of a lender and of falsely promising, both of which are circumstantial. Given that all maxims necessarily include some circumstances, O'Niell's view makes these implicit circumstances explicit. This precision is a benefit; so long as my circumstances are not so finely grained that they are uninterpretable, they render O'Niell's maxims more precise than Korsgaard's maxims.

A.2 Kitcher's View on Motives

Kitcher begins with O'Niell's circumstance, act, goal view and expands it to include the motive for a 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. Kitcher acknowledges that O'Niell's formulation suffices for the universalizability test, but merely argues that it is not the most general form 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 a circumstance, act, goal tuple. My project takes this input and returns the motivation that the dutiful, moral agent would adopt, which is "because this maxim is morally worthy." 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.

B Kroy's Formalization

In this appendix, I implement a formalization of the categorical imperative introduced by Moshe Kroy in 1976 (Kroy, 1976). Kroy used Hinktikka's deontic logic to formalize the Formula of Universal Law and the Formula of Humanity. I will first import the additional logical tools that Hintikka's logic contains, then examine the differences between his logic and DDL, and finally implement and test Kroy's formalization of the FUL

B.1 Implementing Kroy's Formalization

In this section, I present necessary logical background, working my way up to implementing Kroy's formalization by the end of the section. First, Kroy's logic requires the notion of a subject, which I define as a new type, just as I did for my implementation.

typedecl s — s is the type for a "subject," i.e. the subject of a sentence

Kroy also defines a substitution operator.[37] $P(d/e)$ is read in his logic as "P with e substituted for d." DDL has no such notion of substitution, so I will use the more generalized notion of an "open sentence," as I did for my formalization. An open sentence takes as input a subject and returns a complete or "closed" DDL formula by binding the free variable in the sentence to the input. For example, "does action" is an open sentence that can be instantiated with a subject.

type-synonym $os = (s \Rightarrow t)$

— "P sub (d/e)" can be written as "$S(e)$", where $S(d) = P$.

— The terms that we substitute into are instantiations of an open sentence, and substitution re-instantiates the open sentence with a different subject.

New Operators

Because Isabelle is strongly typed, I define new operators to handle open sentences. These operators are similar to DDL's original operators and will simplify notation.

[37] See page 196 in Kroy (1976).

abbreviation *os-neg*::*os* $\Rightarrow$ *os* ($\neg$-)

where $(\neg A) \equiv \lambda x.\ \neg(A(x))$

abbreviation *os-and*::*os* $\Rightarrow$ *os* $\Rightarrow$ *os* (-$\wedge$-)

where $(A \wedge B) \equiv \lambda x.\ ((A(x)) \wedge (B(x)))$

abbreviation *os-or*::*os* $\Rightarrow$ *os* $\Rightarrow$ *os* (-$\vee$-)

where $(A \vee B) \equiv \lambda x.\ ((A(x)) \vee (B(x)))$

abbreviation *os-ob*::*os* $\Rightarrow$ *os* (O{-})

where $\mathrm{O}\{A\} \equiv \lambda x.\ (O\ \{A(x)\})$

Once again, the notion of permissibility will be useful here. Recall that an action can either be obligated, permissible, or prohibited. A permissible action is acceptable (there is no specific prohibition against it), but not required (there is no specific obligation requiring it).

abbreviation *ddl-permissible*::*t*$\Rightarrow$*t* (P {-})

where $P\ \{A\} \equiv \neg\ (O\ \{\neg\ A\})$

abbreviation *os-permissible*::*os*$\Rightarrow$*os* (P {-})

where $\mathrm{P}\ \{A\} \equiv \lambda x.\ P\ \{A(x)\}$

Differences Between Kroy's Logic (Kr) and DDL

There is potential for complication because Kroy's original paper uses a different logic than DDL. His custom logic is a slight modification of Hintikka's deontic logic (Hintikka, 1962). In this brief interlude, I examine if the semantic properties that Kroy's logic (which I abbreviate to Kr) requires hold in DDL. These differences may explain limitations of Kroy's formalization (including failed tests), but I argue that the alternative properties of DDL cohere better with moral intuition. Thus, even if Kroy's formalization would pass more tests if it were implemented using Hintikka's logic, the logic itself would be less morally plausible than DDL, and would thus remain a worse implementation of automated Kantian ethics.

Many of the differences between Kr and DDL can be explained by a difference in their semantics. The most faithful interpretation of Kr is that if A is permissible in a context, then it must be true at some world in that context. Kr operates under the "deontic alternatives" or Kripke semantics, summarized by Solt as follows: "A proposition of the sort OA is true at the actual world w if and only if A is true at every deontic alternative world to w" (Solt, 1984).

Under this view, permissible propositions are obligated at some deontic alternatives, or other worlds in the system, but not at all of them. This property does not hold in DDL.

lemma *permissible-semantics*:

fixes A w

shows $(P\ \{A\})\ w \longrightarrow (\exists x.\ A(x))$

nitpick[*user-axioms*] **oops**

— Nitpick found a counterexample for card i = 1:

Free variable: A = $(\lambda x.\_)(i_1$:= False)

DDL uses neighborhood semantics, not the deontic alternatives view, which is why this proposition fails in DDL. In DDL, the *ob* function abstracts away the notion of deontic alternatives. Even if one believes that permissible statements should be true at some deontic alternative, it's not clear that permissible statements must be realized at some world. This also coheres with our understanding of obligation. There are permissible actions like "Lavanya buys a red folder" that might not happen in any universe.

An even stricter version of the semantics that Kr requires is that if something is permissible at a world, then it is obligatory at some world. This is a straightforward application of the Kripke semantics. This also fails in DDL.

lemma *permissible-semantics-strong*:

fixes A w

shows $P\ \{A\}\ w \longrightarrow (\exists x.\ O\ \{A\}\ x)$

nitpick[*user-axioms*] **oops**

— Nitpick found a counterexample for card i = 1:

Free variable: A = $(\lambda x.\_)(i_1$:= False)

This also doesn't hold in DDL because DDL uses neighborhood semantics instead of the deontic alternatives or Kripke semantics. This also seems to cohere with our moral intuitions. The statement "Lavanya buys a red folder" is permissible in the current world, but it's hard to see why it would be oblgiatory in any world.

Another implication of the Kripke semantics is that Kr disallows "vacuously permissible statements." In other words, if something is permissible it has to be obligated at some deontically perfect alternative. If we translate this to the language of DDL, we expect that if A is permissible, it is obligated in some context.

lemma *permissible-semantic-vacuous*:

fixes A w

shows $P\ \{A\}\ w \longrightarrow (\exists x.\ ob(x)(A))$

nitpick[*user-axioms*] **oops**

— Nitpick found a counterexample for card i = 1:

Free variable: A = $(\lambda x.\_)(i_1$:= False)

In order to make this true, we'd have to require that everything is either obligatory or prohibited somewhere, but this makes permissibility impossible, which is clearly undesirable.

Kroy's formalization of the FUL

I now implement Kroy's formalization of the Formula of Universal Law. Recall that the FUL reads "act only in accordance with that maxim which you can at the same time will a universal law" (Kant, 1785, 34). Kroy interprets this to mean that if an action is permissible for a specific agent, then it must be permissible for everyone. This formalizes the moral intuition prohibiting free-riding. According to the categorical imperative, no one is a moral exception.

abbreviation *FUL*::*bool* **where** $FUL \equiv \forall\, w\, A.\ ((\exists p{::}s.\ ((P\ \{A\ p\})\ w)) \longrightarrow ((\forall p.(\ P\ \{A\ p\})\ w)))$

— If action A is permissible for some person, then, for any person p, action A must be permissible for p. The notion of "permissible for" is captured by the substitution of x for p.

This formalization does not hold in DDL, the base logic. This means that Kroy's formalization already passes one test, and that adding it as an axiom will strengthen the logic.

lemma *FUL*:

shows *FUL*

nitpick[*user-axioms*] **oops**

— Nitpick found a counterexample for card s = 2 and card i = 2:

Skolem constants: A = $(\lambda x.\_)(s_1$:= $(\lambda x.\_)(i_1$:= True, i_2 := True), s_2 := $(\lambda x.\_)(i_1$:= False, i_2 := False))

p = s_1

x = s_2

axiomatization where *FUL*: *FUL*

Now that I have added Kroy's formalization of the FUL as an axiom, I will check that it is consistent by looking for a model that satisfies it and all the other axioms of DDL.

lemma *True* **nitpick**[*user-axioms*, *satisfy*, *card=1*] **oops**

— Nitpick found a model for card i = 1:

Empty assignment

This completes my implementation of Kroy's formalization of the first formulation of the categorical imperative. I defined new logical constructs to handle Kroy's logic, studied the differences between DDL and Kr, implemented Kroy's formalization of the Formula of Universal Law, and showed that it is both non-trivial and consistent.

B.2 Testing Kroy's Formalization

In this section, I use my testing framework to evaluate Kroy's formalization. I find that, while the formalization is considerably stronger than the naive formalization, it still fails many of these tests. Some of these failures are due to the differences between Kroy's logic and my logic mentioned in Section B.1, but some reveal philosophical problems with Kroy's interpretation of what the formula of universal law means.

Obligations Universalize Across People I already showed above that Kroy's formalization is stronger than DDL. Next, I test whether or not obligations universalize across people. This test passes, perhaps trivially, due to the fact that this property is definitionally the basis of Kroy's formalization; his formalization states, intuitively, that obligations must hold across all people.

lemma *obligation-universalizes*:

shows $\forall w.\ (\exists p.\ O\ \{A\ p\}\ w) \longrightarrow (\forall p.\ O\ \{A\ p\}\ w)$

nitpick[*user-axioms*, *falsify=true*] **oops**

Obligations Universalize Across People The next test verifies that obligations cannot contradict. Kroy's formalization fails this test because Nitpick can find a model in which A and

$\neg A$ are both obligated.

lemma *conflicting-obligations*:

fixes A w

shows $(O\,\{A\} \wedge O\,\{\neg\, A\})\; w$

nitpick [*user-axioms*, *falsify=false*] **oops**

— Nitpick found a model for card i = 2 and card s = 1:

Free variable: A = $(\lambda x.\_)(i_1$:= False, i_2 := True)

Recall the stronger version of this property: if two maxims imply a contradiction, they may not be simultaneously obligated. This test also fails for Kroy's formalization.

lemma *implied-contradiction*:

fixes A B w

assumes $((A \wedge B) \rightarrow \bot)\; w$

shows $\neg\,(O\,\{A\} \wedge O\,\{B\})\; w$

nitpick [*user-axioms*, *falsify*] **oops**

— Nitpick found a counterexample for card i = 2 and card s = 1:

Free variables: A = $(\lambda x.\_)(i_1$:= True, i_2 := False) B = $(\lambda x.\_)(i_1$:= True, i_2 := False) w = i_2

Distributive Property for Obligations Next, I test the closely related distributive property for obligations. As expected, this property also fails, since it is a derivative of contradictory obligations.

lemma *distributive-property*:

fixes A B w

shows $O\,\{A \wedge B\}\; w \equiv O\,\{A\} \wedge O\,\{B\}\; w$

nitpick [*user-axioms*, *falsify*] **oops**

— Nitpick found a counterexample for card i = 2 and card s = 1:

Free variables: A = $(\lambda x.\_)(i_1$:= False, i_2 := True) B = $(\lambda x.\_)(i_1$:= True, i_2 := False)

Prohibits Actions That Are Impossible to Universalize Next, I test if Kroy's formalization is strong enough to prohibit actions that are impossible to universalize. As when running this test for my formalization, I need to define some logical background to run this test. I use lying as a toy example of an action that is impossible to universalize.

To run this test, I epresent the sentence "At all worlds, it is not possible that everyone lies simultaneously," in Kroy's logic. This requires the following abbreviations.

consts *lie*::*os*

— This is an empty constant to represent the act of lying, which is an open sentence. Unlike Chapter 4, I do not specify any properties of lying, so this could be replaced with any action that is impossible to universalize.

abbreviation *everyone-lies*::*t* **where** *everyone-lies* $\equiv \lambda w.\ (\forall\, p.\ (lie(p)\ w))$

— This represents the term "all people lie".

— The term above is true for a set of worlds i such that, at all the worlds w in i, all people at w lie.

abbreviation *lying-not-possibly-universal*::*bool* **where** *lying-not-possibly-universal* $\equiv \models(\neg\,(\Diamond\ everyone\text{-}lies))$

— Armed with *everyone-lies* $\equiv \lambda w.\ \forall\, p.\ lie\ p\ w$, it's easy to represent the desired sentence. The abbreviation above reads, "At all worlds, it is not possible that everyone lies."

With this logical background, I can test if lying not being possible to universalize implies that it is prohibited. Surprisingly, Kroy's formalization fails this test.

lemma *lying-prohibited*:

shows *lying-not-possibly-universal* $\longrightarrow$ $(\ \models(\neg\ P\ \{lie(p)\}))$

nitpick[*user-axioms*] **oops**

— Nitpick found a counterexample for card i = 1 and card s = 2:

Free variables:

lying_not_possibly_universal = True

$p = s_1$

Kroy's formalization is not able to show that if lying is not possible to universalize, it is prohibited. This is unexpected—Kroy's formalization seemingly hinges on universalizability, so it seems as though it should pass this test. To understand this, I outline the syllogism that one might *expect* to prove that lying is prohibited and test each component of this syllogism in Isabelle.

1. At all worlds, it is not possible for everyone to lie. *(This is the assumed sentence.)*

2. At all worlds, there is necessarily someone who doesn't lie. *(Modal dual of (1))*

3. If A is permissible for subject p at world w, A is possible for subject p at world w. *(Modified Ought Implies Can)*

4. If A is permissible at world w for any person p, it must be possible for everyone to A at w. *(FUL and (3))*

5. Lying is impermissible. *(Follows from (4) and (1))*

I now test each step of this syllogism to determine where Kroy's formalize deviates from the expected results. Step 1 holds by assumption, and Step 2 holds as shown below, but the syllogism breaks down at Step 3.

lemma *step2*:
 shows *lying-not-possibly-universal* $\longrightarrow \models((\Box (\lambda w.\ \exists p.\ (\neg (lie(p))\ w))))$
 by *simp*
— Step 2 holds.

lemma *step3*:
 fixes *A p w*
 shows $P\ \{A(p)\}\ w \longrightarrow (\Diamond\ (A(p))\ w)$
 nitpick [*user-axioms*, *falsify*] **oops**
— Nitpick found a counterexample for card 'a = 1, card i = 1, and card s = 1:
Free variables: A = $(\lambda x.\_)(a_1$:= $(\lambda x.\_)(i_1$:= False)) p = a_1

The above lemma shows that the syllogism fails at Step 3, explaining why the lemma doesn't hold as expected. Kroy explicitly states[38] that Step 3 holds in his logic, so this failure may be explained by this difference in Kr and DDL. However, upon reflection, it is not clear that the statement expressed in Step 3 should actually hold. Step 3 states that all permissible actions must be possible, but this implies that impossible actions are not permissible, so they must be prohibited, which seems silly. For example, imagine I make a trip to Target to pur-

[38] See footnote 19 on p. 199 of Kroy (1976)

chase a folder, and they offer blue and black folders. Even though it is impossible for me to purchase a red folder, it doesn't seem impermissible for me to purchase a red folder.

A deeper issue inspired by this test is that Kroy's interpretation of the FUL is empty in a circular way. His formalization interprets the FUL as prohibiting A if there is someone for whom A'ing is not permissible. This requires some preexisting notion of the permissibility of A, and is thus circuar. The categorical imperative is supposed to be the complete, self-contained rule of morality, but Kroy's version of the FUL prescribes obligations in a self-referencing manner. The FUL is supposed to define what is permissible and what isn't, but Kroy defines permissibility in terms of itself.

Neither of these errors are obvious from Kroy's presentation of his formalization of the categorical imperative. This is another example of the power of computational ethics. Making Kroy's interpretation of the categorical imperative precise demonstrated philosophical problems with that interpretation.

Remaining Tests It is clear that Kroy's formalization does not encode a robust conception of a maxim, as it simply evaluates actions. Moreover, the emptiness discussed above implies that Kroy's formalization cannot actually generate *any* obligations from scratch, and so the formalization automatically fails to prohibit conventional or natural acts.

Thus, this completes my testing of Kroy's formalization. While Kroy's formalization represents some progress over the control group (it passes the first two tests), it is clear that many limitations remain. My implementation passes all of the tests that Kroy's formalization fails, and thus represents significant progress.

Miscellaneous Tests I conclude my examination of Kroy's formalization by presenting one more test specific to Kroy's formalization. In addition to his formalization of the FUL, Kroy also presents a formalization of a stronger version of the FUL and argues that his formalization is implied by the stronger version. I can test that claim by formalizing this stronger formalization as well.

abbreviation *FUL-strong*::*bool* **where** *FUL-strong* $\equiv\ \forall\, w\, A.\ ((\exists p{::}s.\ ((P\ \{A\ p\})\ w))\ \longrightarrow(\ ((\ P\ \{\ \lambda x.\ \forall\, p.\ A\ p\ x\})\ w)))$

lemma *strong-implies-weak*:

shows *FUL-Strong* $\longrightarrow$ *FUL*

using *FUL* **by** *blast*

— This lemma holds, showing that Kroy is correct in stating that this version of the FUL is stronger than his original version.

The difference between the strong and weak versions of the FUL is subtle. The consequent of FUL is "for all people p, it is permissible that they A." The consequent of this stronger statement is "it is permissible that everyone A." This stronger statement requires that it is permissible for everyone to A simultaneously. Kroy immediately rejects this version of the categorical imperative, arguing that it's impossible for everyone to be the US president simultaneously, so this version of the FUL prohibits running for president.

Most Kantians would disagree with this interpretation. Consider the classical example of lying, as presented in Kemp (1958) and in Korsgaard (1985). Lying fails the universalizability test because in a world where everyone lies simultaneously, the practice of lying would break down. If we adopt Kroy's version, lying is only prohibited if, no matter who lies, lying is impermissible. As argued above, this rule circularly relies on some existing prohibition against lying for a particular person, and thus fails to show the wrongness of lying.

This misunderstanding is actually related to another weakness of Kroy's formalization: it lacks a robust conception of a maxim. Consider Kroy's example of the maxim of running for president. If the action being evaluated is, "I will be president of the United States," as Kroy interprets it, then it is clearly not universalizable for the reason he argued. However, using the most robust circumstance, act, tuple representation of a maxim, the maxim of action would be something like, "When I believe that I would make a good president, I will launch a presidential campaign to become president." This more nuanced maxim is universalizable: it is clearly possible for all people who believe they would make good presidents to run for president. Under this more sophisticated representation of a maxim, Kroy's stronger version of the FUL is actually correct.

It is tempting to claim that this issue explains why the tests above failed. To test this hypothesis, I check if the stronger version of the FUL implies that lying is impermissible. Sadly, even the stronger version of the FUL fails this test.

lemma *strongFUL-implies-lying-is-wrong*:

fixes p

shows *FUL-strong* $\wedge$ *lying-not-possibly-universal* $\longrightarrow \models(\neg\, P\, \{lie(p)\})$

nitpick[*user-axioms*, *falsify*] **oops**

— Nitpick found a counterexample for card i = 1 and card s = 1:

Free variable: p = s_1 Skolem constant: λy. p = $(\lambda x.\_)(i_1 := s_2)$

The failure of this test implies that not even the stronger version of Kroy's formalization of the FUL can show the wrongness of lying. As mentioned earlier, there are two independent errors. The first is the the assumption that impossible actions are impermissible and the second is the circularity of the formalization. The stronger FUL addresses the second error, but the first remains, and so the stonger formalization of the FUL still fails this test.

C Additional Tests

In this section, I show that my system can correctly show prohibitions against actions that are impossible to universalize, against conventional acts, and against natural acts. In the process of running these tests, I discover and resolve an ambiguity in Korsgaard's canonical example of a prohibited maxim. I show that her maxim actually has two readings, one reading under which it is a natural act, and another under which it is a conventional act. My formalization can correctly handle both readings. The recognition of this ambiguity is another example of the power of computational ethics, and demonstrates that the process of making a philosophical argument precise enough to represent to a machine can generate philosophical insights.

In this section, I show that the maxim, "When strapped for cash, falsely promise to pay your friend back to get some easy money," is prohibited. Korsgaard uses this example when arguing for the practical contradiction interpretation of the FUL (Korsgaard, 1985). She argues that this maxim describes a conventional act, or an act that is possible due to some pre-existing social system, and is thus within reach for the logical contradiction interpretation. Natural acts, on the other hand, are acts that are possible simply due to the laws of nature, such as murder, and the logical contradiction interpretation cannot correctly handle such acts.

I argue that, in addition to Korsgaard's reading of this maxim as a conventional act, there is also another reading of the maxim as a natural act. Under Korsgaard's reading, the act of falsely promising is read as entering a pre-existing, implicit, social system of promising with no intention of upholding your promise. Under the second reading, the act of falsely promising is equivalent to uttering the worlds "I promise X" without intending to do X. There is a difference between promising as an act with meaning in a larger social structure and merely uttering the words "I promise," so these two readings are distinct.

Under Korsgaard's reading, the maxim fails because falsely promising is no longer possible in a world where everyone everyone does so, or because the action of falsely promising is literally impossible to universalize. Recall that this is how the logical contradiction interpretation prohibits this maxim—falsely promising is no longer possible when universalized because the institution of promising breaks down. First, I formalize this argument and show that my system can show the wrongness of the false promising maxim under Korsgaard's reading. This also shows that my system can show the wrongness of a maxim that is impossible to universalize.

To formalize this argument, I first define the relevant maxim.

consts *when-strapped-for-cash*::*t*

— This constant represents the circumstances "when strapped for cash."

consts *falsely-promise*::*os*

— This constant represents the act "make a false promise to pay a loan back."

consts *to-get-easy-cash*::*t*

— This constant represents the goal "to get some money."

abbreviation *false-promising*::*maxim* **where**

false-promising ≡ (*when-strapped-for-cash*, *falsely-promise*, *to-get-easy-cash*)

— Armed with the circumstances, act, and goal above, I can define the example false promising maxim as a tuple.

The logical objects above are empty or thin, in the sense that I haven't specified any of their relevant properties, such as a robust definition of promising or any system of currency. I define only the properties absolutely necessary for my argument as assumptions and show that, if the maxim above satisfies the assumed properties, it is prohibited. Specifically, I am interested in Korsgaard's reading of this maxim, under which promising is a social convention that breaks down when abused. Instead of formally defining a conventional act, which requires wading into complex debates about trust and social contracts, I merely focus on the fact that, unders this reading, not everyone can falsely promise universally. Whatever kind of social convention promising is, my argument merely relies on the impossibility of breaking it.

abbreviation *everyone-can't-lie* **where**

everyone-can't-lie $\equiv \forall\, w.\ \neg\ (\forall\, s.\ \textit{falsely-promise}(s)\ w)$

— The above formula reads, "At all worlds, it is not the case that everyone falsely promises."

With this abbreviation, I show that if not everyone can falsely promise simultaneously, then the constructed maxim about falsely promising is prohibited.

lemma *falsely-promising-korsgaard-interpretation*:

assumes $\forall$ *w*. *when-strapped-for-cash w*

— Restrict our focus to worlds in which the circumstance of being strapped for cash holds. A technical detail.

assumes $\forall$ *s*. $\models$ (*well-formed false-promising s*)

— Initial set-up: the falsely promising maxim is well-formed.

assumes *everyone-can't-lie*

— The assumption that this is Korsgaard's reading of the maxim, in which everyone cannot falsely promise simultaneously.

shows $\forall$ *s*. $\models$ (*prohibited false-promising s*)

proof –

have $\forall$ *s*. *not-universalizable false-promising s*

by (*simp add*: *assms*(3) *assms*(1))

— As in the proofs in Chapter 4, once I split this proof into this intermediate lemma, Isabelle can automatically complete the proof.

thus *?thesis*

using *FUL assms*(2) **by** *blast*

qed

The above lemma shows that, under Korsgaard's reading of promising as a conventional act, my system can show that falsely promising is prohibited. This means that my system passes both the conventional act test and the test that requires showing the wrongness of actions that are impossible to universalize. Next, I show that my system can show a prohibition against this maxim even under the second reading, which understands it as a natural act.

Under the second reading of this maxim, the act "falsely promising" refers to uttering the sentence "I promise to do X" with no intention of actually doing X. This is a natural act because the act of uttering a sentence does not rely on any conventions, merely the laws of nature governing how your mouth and vocal cords behave.[39]

The logical contradiction interpretation cannot prohibit this version of the maxim because making an utterance is always logically possible, even if everyone else makes the same utterance. However, under this reading, the practical contradiction interpretation prohibits this maxim because, in a world where false promising is universalized, no one believes promises anymore, so the utterance is no longer an effective way to get money. Because my system implements the stronger practical contradiction interpretation of the FUL, it can show the wrongness of this maxim even under this reading. First, I formalize this reading of the maxim.

consts *believed*::*os*

abbreviation *false-promising-not-believed* **where**

$\textit{false-promising-not-believed} \equiv \forall\, w\, s.\ (\textit{falsely-promise}(s)\ w \longrightarrow \neg\ \textit{believed}(s)\ w)$

— This abbreviation formalizes the idea that if everyone falsely promises, then no one is believed when promising.

abbreviation *need-to-be-believed* **where**

$\textit{need-to-be-believed} \equiv \forall\, w\, s.\ (\neg\ \textit{believed}(s)\ w \longrightarrow \neg((\textit{falsely-promise}\ s) \rightarrow \textit{to-get-easy-cash})w)$

— This abbreviation formalizes the idea that if a promise is not believed, then it is not an effective way of getting easy cash.

Once again, I avoid giving robust definitions of hotly debates concepts like belief. Instead, I represent the bare minimum logical background: false promises won't be believed when universalized, and promises must be believed to be effective.

lemma *falsely-promising-bad-natural-act*:

assumes $\forall\, w.\ \textit{when-strapped-for-cash}\ w$

— Restrict our focus to worlds in which the circumstance of being strapped for cash holds. A technical detail.

assumes $\forall\, s.\ \models (\textit{well-formed false-promising}\ s)$

— Initial set-up: the falsely promising maxim is well-formed.

[39]Linguistic relativists may take issue with this claim and may argue that if the English language had never developed, then making this utterance would be impossible. Even if this is true, the laws of nature itself would not prohibit making the sounds corresponding to the English pronounciation of this phrase, so the act would still not be impossible in the way that a conventional act can be.

assumes *false-promising-not-believed*

assumes *need-to-be-believed*

— The two assumptions above.

shows $\forall s. \models$ (*prohibited false-promising s*)

proof –

have $\forall s.$ *not-universalizable false-promising s*

using *assms*(*1*) *assms*(*2*) *assms*(*3*) **by** *auto*

thus *?thesis*

using *FUL assms*(*2*) **by** *blast*

qed

— With some help, Isabelle is able to show that the maxim is prohibited under this reading as well.

These proofs demonstrate that my formalization is able to correctly prohibit this maxim, whether it is understood as a conventional act or a natural act. Korsgaard argues that the practical contradiction interpretation outperforms other interpretations of the FUL because it can show the wrongness of both conventional and natural acts. Therefore, the fact that my interpretation can correctly show the wrongness of both conventional and natural acts is evidence for its correctness as a formalization of the practical contradiction interpretation.