Automated Kantian Ethics: A Faithful Implementation and Testing Framework

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. Warnings from regulators, philosophers, and computer scientists about the dangers of unethical artificial intelligence, from science-fiction killer robots to criminal sentencing algorithms prejudiced against people of color, have spurred interest in automated ethics, or the development of machines that can perform ethical reasoning. Much prior work in automated ethics approaches the problem from a computational perspective and rarely engages with philosophical literature on ethics, despite its relevance to the development of AI agents that can responsibly navigate the world. If automated ethics draws on sophisticated philosophical literature, the ethical reasoning underlying such decisions will be more nuanced, precise, consistent, and trustworthy. However, faithfully translating complex ethical theories from natural language to the rigid syntax of a computer program poses technical and philosophical challenges.

In this thesis, I present an implementation of automated Kantian ethics that is faithful to the Kantian philosophical tradition. Of the three major ethical traditions, Kant's categorical imperative is the most natural to formalize because it is an inviolable formal rule that requires less situational awareness than other ethical theories. I formalize Kant's categorical imperative 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 reason about two ethical dilemmas used to criticize Kantian ethics: the difference between lying and joking and the example of a murderer knocking on your door asking about the location of their intended victim. Finally, I examine the philosophical implications of this system, exploring its limitations and its potential to help both AI agents and human beings better reason about ethics. 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

1 Introduction

As AI agents become more sophisticated and less dependent on humans, interest begins to mount in the development of computers that can perform ethical reasoning, also known as automated moral agents. 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 Allegheny 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 machine ethics (also called automated ethics), or the study of how to develop 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)), 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, a language model 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 and, in the extreme case, are black box algorithms. This reduces human trust in a machine's controversial ethical judgements. If a machine prescribes killing one person to save three without justifying this decision, it is difficult to trust this judgement enough to act on it or endorse a machine acting on it. The high stakes of automated ethics require explainability to build trust and catch mistakes.

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, but many philosophers reject this notion of aggregation (Sinnott-Armstrong, 2021). Approaches that begin with an ethical theory, instead of a computational tool, 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, is equally as important in determining the machine's decisions. Another complication is that philosophers do not agree that on a single choice of ethical theory. Philosophers who so agree that a specific ethical theory, like Kantian ethics, is true, still debate the theory's details.[1] Moreover, even once reasoning within a particular ethical theory is automated, those who disagree with that theory will disagree with the system's judgements.

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 Carmo and Jones' 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 ??, 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 ??, 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 definitions such that my system, when given an appropriately represented input, can

[1]For examples of these debates in the case of Kantian ethics, see Section Joking and Section Murderer.

prove that the input is permissible, obligatory, or prohibited. It can also return a list of facts used in the proof and, in some cases, an Isar-style human readable proof.

3. In Sections ?? and ??, 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.

4. In Section ??, I present a testing framework that can evaluate how faithful an implementation of automated Kantian ethics is. My framework includes meta-ethical tests and application tests inspired by philosophical literature. This testing framework shows that my formalization substantially improves on prior work and can be generalized to evaluate any implementation of automated Kantian ethics.

5. In Section ??, I present new ethical insights discovered using my system and argue that computational methods like the one presented in this paper can help philosophers address ethical problems. Not only can my system help machines reason about ethics, but it can also help philosophers make philosophical progress.

I choose to formalize Kant's moral rule in Carmo and Jones' Dyadic Deontic Logic (DDL) (Carmo and Jones, 2013). Deontic logic is a modal logic that can express obligation, or morally binding requirements. Traditional modal logics include the necessitation operator, denoted as $\Box$. In modal logic 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 $\diamond p \iff \neg(\Box(\neg p))$ and operators of propositional logic like $\neg, \wedge, \vee, \rightarrow$. I use DDL, 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. 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 to formalize, as I argue in Section WhyKant. 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 people should only act on those principles that can be acted on by all people without contradiction. For example, in a world where everyone falsely promises to repay a loan, lenders will no longer believe these promises and will stop offering loans. Therefore, not everyone can simultaneously falsely promise to repay a loan, so the FUL thus prohibits this act.

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 used to generate an ethical judgement, and, in some cases, construct human-readable proofs. In Sections Joking and Murderer, I use my system to arrive at sophisticated solutions to two ethical dilemmas often used in critiques of Kantian ethics. Because my system is faithful to philosophical literature, it is able to provide nuanced answers to these paradoxes that require a deep understanding of Kantian ethics.

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. I represent each of these properties as a sentence in my logic that my system should be able to prove or refute. I run the tests by using 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 $\neg(O\{A|B\} \wedge O\{\neg A|B\})$. I find that my system outperforms the control group of raw DDL, without any moral axioms added, and Moshe Kroy's prior attempt at formalizing Kantian ethics in deontic logic (Kroy, 1976).

As it stands, my implementation can evaluate the moral status of sentences represented in my logic. Given an appropriate input, my project returns a value indicating if the action is obligatory (its negation violates the FUL), permissible (consistent with the FUL), or prohibited (violates the FUL) by proving or refuting a theorem in my logic. Because the logic underlying these judgements is grounded in philosophical literature, my system can correctly resolve complex moral dilemmas with only relatively uncontroversial facts about the world. Moreover, because I automate an explicit ethical theory, the ethical reasoning underlying my system's judgements is interpretable by a human being. I implement this ethical theory using the Isabelle/HOL interactive theorem prover, which can list the axioms and theorems used in a proof, so my system is explainable.

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. I demonstrate the potential of computational ethics by presenting a philosophical insight about which kinds of maxims 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 two 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.

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, present the philosophical, logic, and computational background underlying my system, and explain the consequences of each of the three choices I make.

These specific components determine the features and limitations of my implementation of automated ethics, but 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 merely 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 rules or procedures for problem solving and Kantian ethics (which is one 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,[2] Kantian ethics is more amenable to automation.

[2]Technically, virtue ethics and consequentialism are broad ethical traditions, while Kantian ethics is a specific ethical theory within the deontological tradition, but"if any philosopher is regarded as central to deontological moral theories, it is surely Immanuel Kant" (Alexander and Moore, 2021). Out of the three major ethical traditions, deontology has the most central representative in the form of Kant. 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.

I do not aim to show that Kantian ethics is the only tractable theory to automate or to present a comprehensive overview of all consequentialist or virtue ethical theories. Instead, I present a sample of some approaches in each tradition and argue that 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 is an ethical theory that evaluates an action by evaluating its consequences.[3] For example, utilitarianism is a form of consequentialism in which the moral action is the action that produces the most good (Driver, 2014). The 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 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 requires that an automated ethical system somehow collect a subset of the infinite consquences of following an action, a difficult, if not impossible, task. Moreover, compiling this

[3]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.

database of consequences requires answering difficult questions about which consequences were actually caused [4] by an action and determining 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 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.[5] 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 take a stance on the debate over what qualifies as a "good consequence," or what the theory of the good is. 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 (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, mak-

[4] 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).

[5] 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 my approach to morally relevant circumstances, see Section 3.1.2.

ing 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 impossible. Either a human being assigns values to states of affairs, which doesn't scale, or the machine does, which requires massive common-sense and increases room for doubting the system's judgements. This may be a tractable problem, but it is much 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 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 debates 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.[6] 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[7] as the theory of the good (Anderson et al., 2004, 2). Again, their proposal assumes that pleasure and pain can be given numeric values and that these values can be aggregated with a simple sum, taking an implicit stance on the aggregation question. Other attempts to automate consequentialist ethics will suffer similar problems because, at some point, 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 particular theory of the 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 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

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

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

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

One potential appeal of virtue ethics is that 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 through moral education (Aristotle, 1951). This implies that ethical behavior can be learned from some dataset of virtuous acts, such as those that an ideal virtuous agent would undertake. These theories seem to point towards a machine learning approach to automated ethics, in which ethics is learned from a dataset of acts tagged as virtuous or not virtuous.

Just as prior work in consequentialism takes implicit or explicit stances on debates in consequentialist literature, so does 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 virtues that virtue ethicists accept, this implicit view will likely be controversial.

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.[8]

2.1.3 Kantian Ethics

In this paper I focus on Kantian ethics. Kant's theory is centered on practical reason, which is the kind of reason that we use to decide what to do and the source of our agency. 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, 4:412). In contrast, 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.[9] 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 the form of certain maxims requires any rational agent to will them, and these maxims are

[8]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 "bottom-up" approaches, in which a system learns moral judgements from prior judgements. I will extend this argument to general bottom-up approaches in Section Related Work.

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

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 explain 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, which we can use to 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 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. Later in my project, I argue that a maxim can be represented simply as a tuple of circumstances, act, and goal.[10] This representation is simple and efficient, especially when compared to the 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

[10] For more, see Section 3.1.2.

One debate in Kantian ethics is the role of "common-sense" reasoning. Kantian ethics requires common-sense reasoning to determine which circumstances are "morally relevant" in the formulation of a maxim. Many misunderstandings in Kantian ethics are due to badly formulated maxims, so this question is important for an ethical reasoner to answer.[11] 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 in a fully automated moral agent will require answering this question, a challenging computational and philosophical task.

Common-sense reasoning 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. Making any ethical judgement requires robust conceptions of the action at hand, falsely promising in this case. The advantage of Kantian ethics is that this is all the common sense that it requires, whereas a consequentialist or virtue ethical agent will require much more[12]. All moral theories evaluating falsely promising will a robust definition of promising, but consequentialism and virtue ethics will also require additional information that Kantian ethics will not. Thus, although the need for common sense poses a challenge to automated Kantian ethics, this challenge is

[11]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.

[12]In Sections Lying and Murderer, I also use my system to demonstrate that Kantian ethics requires relatively thin conceptions of concepts like falsely promising.

more acute for consequentialism or virtue ethics.

2.1.4 The Formula of Universal Law

Earlier I mentioned that Kant presents three formulations, or versions, of what he calls the "supreme law of morality," but that 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[13], 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, allowing me to remain faithful to philosophical literature on Kant. Precision reduces the need to make choices to resolve debates and ambiguities. Minimizing these choices minimizes arbitrariness

[13]The FUL is often seen as emblematic of Kantian constructivism (Ebels-Duggan, 2012, 173). My project is not committed to Kantian constructivism. I believe that computational ethics is likely a valuable tool for any ethicist, and I make the case for Kantian ethics specifically.

in my formalization and puts it on solid philosophical footing.

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 formalizations. 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. 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 a standard to live towards that might not be achieved. 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, automating it creates a functional, minimum ethical theory that can serve as a foundation for implementations of other aspects of Kant's 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).

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. [14] 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 by modifying the context. The obligation operator uses a neighborhood semantics, instead of the Kripke semantics (Scott, 1970;

[14] 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)

MONTAGUE, 1970). While Kripke semantics requires that an obligated proposition hold at all worlds, the neighborhood semantics defines a different set of neighbors, 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. 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.

To demonstrate the power and usage of Isabelle and make DDL more precise, I walk through my reimplementation of Benzmueller, Farjami, and Parent's implementation of DDL in Isabelle/HOL (Benzmüller et al., 2021; Benzmüller et al., 2019).

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 a constant.

typedecl i — This is an Isabelle comment. i is the type for a set of worlds.

— This is a line of actual code used in my implementation. For the rest of the thesis, text typeset like this represents Isabelle code.

type-synonym $t = (i \Rightarrow bool)$ — t represents a set of DDL formulae.

— A set of formulae is defined by its truth value at a set of worlds. For example, the set $\{True\}$ 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 $ob{::}t \Rightarrow (t \Rightarrow bool)$ — set of propositions obligatory in this context

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

In a semantic embedding, axioms are modelled as restrictions on models of the system. In this case, a model is specificied 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

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

$\longrightarrow ob(Z)(\lambda w.(Z(w) \wedge \neg X(w)) \vee Y(w))$

— 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 for different formulas involving the axioms above. Each DDL operator is represented as a HOL formula. Isabelle automatically unfolds formulae 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 will be useful for my purposes, implemented below.

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

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

— 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 $\Diamond A \equiv \neg(\Box(\neg A))$

— 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 {-})
 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* ($\models$-)
 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 involves unfolding definitions and applying 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 requires formalizing the FUL in Dyadic Deontic Logic and then implementing this logic in Isabelle/HOL. The final Isabelle library contains enough logical background and an axiom representing the categorical imperative. Using Isabelle's automated theorem proving abilities, my system can show that appropriately represented actions are, according to Kantian ethics, 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 two other potential implementations of automated Kantian ethics.

3.1 Formalization and Implementation of the FUL

Formalizing the FUL requires defining and 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 fundamentally 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 the notions of 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. It suffices to declare a new type to represent a subject without specifying any of its complex properties, such as the idea that a subject must be rational or human. Instead of providing a complete definition of subject, which would require wading into murky philosophical debates about the nature of agency, I will 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.

In this interpretation, the defining feature of a subject is that it 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). 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 wil 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). 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).

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, 4:428). A rational will does not act randomly (else it would not be rational), but instead in the pursuit of ends which it deems valuable.[15] The inclusion a maxim's end is essential for the version of the FUL that I will implement, explained in Section 3.1.3.

[15]Some argue that a maxim should also include the agent's motive or motivation and I address this concern in Appendix A.1.

SHOULD THIS GO HERE OR IN THE LIMITATIONS SECTION OF THE DISCUSSION

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). [16] 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. For proponents of the "human-in-the-loop" model of AI ethics, in which ethical AI requires that humans guide machines, human involvement may be a feature, not a bug (Lukowicz, 2019). In this model, a human being must create a representation for the dilemma they wish to test, translating a complex situation into a flat logical structure. This parallels the challenge that programmers face when translating the complexity

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

of reality to a programming langauge or computational representation. 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.

Another solution is to develop 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. Determining morally relevant circumstances, either using heuristics or human involvement, is a ripe area for future work.

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* $\Rightarrow s \Rightarrow\ t\ (W\text{--})$

where *will* $\equiv \lambda(c, a, g)\ s.\ (c \rightarrow (a\ s))$

— 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). 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 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, an agent commits themselves 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 would 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.

Under the practical contradiction interpretation, the FUL states, "If, when universalized, 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, 1973a).[17] Note that nothing else changes about this counterfactual world—the circumstances are identical and we neither added additional theorems nor specified the model any further. This represents Lewis's idea of "comparative similarity," where a counterfactual is true if it holds at the most similar world (Lewis, 1973b). In our case, this is just the world where everything is the same except the maxim is not acted

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

on. 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 $\equiv \lambda M.\ (\lambda w.\ (\forall\, p.\ (W\, M\, p)\ w))$
— 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.[18]

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

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

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.*[19]

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

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

abbreviation *prohibited*::*maxim*⇒*s*⇒*t* **where**
prohibited ≡ $\lambda(c, a, g)\; s.\; O\{\neg\; (will\; (c,a, g)\; s) \mid c\}$

[18] The full proof for this result is the Obligation Universalizes Across People Test in Section **??**.

[19] 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 $O\{\neg will\,(C, A, G)\; s \mid C\}$.

— 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 $permissible \equiv \lambda M\, s.\ \neg\ (prohibited\ M\ s)$

— 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 Tests, Kantian ethics never prescribes contradictory obligations, so I will add this as an axiom.

abbreviation *non-contradictory* **where**

$non\text{-}contradictory\ A\ B\ c\ w \equiv ((O\{A|c\} \wedge O\{B|c\})\ w) \longrightarrow \neg((A \wedge (B \wedge c))\ w \longrightarrow False)$

— 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 $no\text{-}contradictions$:$\forall A::t.\ \forall B::t.\ \forall c::t.\ \forall w::i.\ non\text{-}contradictory\ A\ B\ c\ w$

— 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, ultimately 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 Computational Ethics???, demonstrates the power of computational tools in aiding philosophical progress.

abbreviation *FULo*::*bool* **where** *FULo* $\equiv \forall\ c\ a\ g\ s.$ *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 *FULo* $\longrightarrow$ *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`[20] (Paulson and Blanchette, 2015). This axiom roughly states that an obligation can't contradict its context. Knowing that `FUL0` contradicts this particular maxim 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. Moreover, in many models, it is possible to construct a maxim where the circumstances, goal, and act (once a subject acts on it) are all that same term and to show that this maxim is prohibited, resulting in a contradiction.

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*,*

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

[20]The full axiom reads $\models \lambda w.\ ob\ ?B\ ?A \longrightarrow \neg \models \neg\ ?B \wedge ?A$.

For example, the maxim "When I eat breakfast, I will eat breakfast to eat breakfast" is not well-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*$\Rightarrow$*s*$\Rightarrow$*i*$\Rightarrow$*bool* **where**

$\mathit{well\text{-}formed} \equiv \lambda(c, a, g).\ \lambda s.\ \lambda w.\ (\neg\ (c \to g)\ w) \wedge (\neg\ (c \to a\ s)\ w)$

— 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**

$\mathit{FUL} \equiv \forall\, M{::}\mathit{maxim}.\ \forall\, s{::}s.\ (\forall\, w.\ \mathit{well\text{-}formed}\ M\ s\ w) \longrightarrow (\mathit{not\text{-}universalizable}\ M\ s \longrightarrow \models (\mathit{prohibited}\ M\ s)\ )$

— This formalization states that if a maxim is well-formed, then if it is not universalizable, it is prohibited.

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. It is neither contradictory nor prohibited, and it is the wrong kind of question to be asking. 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 Computational Ethics.

Now that I have a consistent version of the FUL, I can complete my implementation by adding it 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 the over 100 lines of Isabelle/HOL code necessary to complete this implementation. In Section 3.2 and Chapter Applications, 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.

Goals	Naive	Kroy's	Custom
FUL Stronger than DDL	×	✓	✓
Obligation Universalizes Across People	×	✓	✓
Contradictory Obligations	×	×	✓
Distributive Property	×	×	✓
Un-universalizable Actions	×	×	✓
Maxims	×	×	✓
Conventional Acts	×	×	✓
Natural Acts	×	×	✓

Figure 1: Table showing which tests each implementation passes. The naive interpretation is raw DDL, Kroy is based on Moshe Kroy's formalization of the FUL, and the custom formalization is my novel implementation.

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) and is thus a control group. 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 Isabelle.[21] 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

[21] I present the complete implementation in Appendix A.2.

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 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.[22] 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

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

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

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`.

Contradictory Obligations 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.[23] 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 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, 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).[24] 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 $\neg\ (O\{W\ (c, a, g)\ s|c\} \wedge O\{\neg(W\ (c, a, g)\ s)|\ c\})\ w$

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

[24] 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.

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 are either permissible or prohibited and obligation contradicts prohibition. In a system with contradictory obligations, this property fails because there is some A that is obligated but also prohibited and therefore not permisible. Formalizing this property below shows that this follows from contradictory obligations.

lemma $\models ((O\,\{A\} \wedge O\,\{\neg\,A\}) \equiv (\neg\,(O\,\{A\} \rightarrow \neg\,O\,\{\neg A\})))$

by *simp*

— `simp` is the simplification tactic, which unfolds definitions to complete a proof. The left-hand side states that A is both obligated and prohibited, and is equivalent to the right-hand side, which states that A is obligated but not permissible.

Distributive Property Another property related to contradictory obligations is the distributive property for the obligation operator.[25] 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

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

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 $\models O\{A \wedge B\}$
using *assms no-contradictions* **by** *fastforce*

— The proof proceeds almost immediately using the new axiom.

Un-universalizable Actions 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 background and the full code is presented in Appendix A.3.

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 ?? and ??. Staying faithful to the philosophical literature enables my system to make more reliable judgements.

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 like promising 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. 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 (Korsgaard, 1985). I run this test in Chapter Applications, 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 A.3.

A Appendix

A.1 Maxims and Motives

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

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

A.2 Kroy's Formalization

This section contains 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 that Kroy relies on, then examine the differences between his logic and DDL, and finally implement and test both of Kroy's formalizations.

A.2.1 Logical Background

Kroy's logic relies heavily on some notion of identity or agency. The logic must be capable of expressing statements like "x does action", which I can write as "x is the subject of the sentence 'does action.'" This requires defining a subject.

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

Kroy also defines a substitution operator[26]. $P(d/e)$ is read in his logic as "P with e substituted for d." DDL has no such notion of substitution, so I will define a more generalized notion of an "open sentence." An open sentence takes as input a subject and returns a complete or "closed" DDL formula by, in effect, 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 actually instantiations of an open sentence, and substitution just requires re-instantiating the open sentence with a different subject.

New Operators

Because Isabelle is strongly typed, we need to define new operators to handle open sentences. These operators are similar to DDL's original operators. We could probably do without these abbreviations, but they will simplify the notation and make it look more similar to Kroy's original paper.

[26] See page 196 in Kroy's original paper Kroy (1976).

abbreviation *os-neg*::*os* ⇒ *os* (¬-)

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

abbreviation *os-and*::*os* ⇒ *os* ⇒ *os* (-∧-)

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

abbreviation *os-or*::*os* ⇒ *os* ⇒ *os* (-∨-)

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

abbreviation *os-ob*::*os* ⇒ *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*⇒*t* (*P* {-})

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

abbreviation *os-permissible*::*os*⇒*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 section, I will determine if some of the semantic properties that Kroy's logic (which I will now call Kr) requires hold in DDL. These differences may become important later and can explain differences in my results and Kroy's.

Deontic alternatives versus the neighborhood 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 Solt (1984) 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." Under this view, permissible propositions are obligated at some deontic alternatives, or other worlds in the system, but not at all of them. Let's see if this holds 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)

Remember that 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. In some ways, 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. Let's test this proposition.

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.

One 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. Sadly, that breaks everything and destroys the notion of permissibility everywhere [27]. If something breaks later in this section, it may be because of vacuous permissibility.

Obligatory statements should be permissible

Kr includes the intuitively appealing theorem that if a statement is obligated at a world, then it is permissible at that world[28]. Let's see if that also holds in DDL.

lemma *ob-implies-perm*:

fixes $A\ w$

shows $O\ \{A\}\ w \longrightarrow P\ \{A\}\ w$

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

— Nitpick found a counterexample for card i = 2:

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

Intuitively, it seems untenable for any ethical theory to not include this principle. My formalization should add this as an axiom.

A.2.2 The Categorical Imperative

I will now implement Kroy's formalization of the Formula of Universal Law. Recall that the FUL says "act only in accordance with that maxim which you can at the same time will a universal law" Kant (1785). 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.

[27] See Appendix for an examination of a buggy version of DDL that led to this insight.

[28] This follows straightforwardly from the Kripke semantics. If proposition A is obligated at world w, this means that at all of w's neighbors, OA holds. Therefore, $\exists w'$ such that w sees w' and OA holds at w' so A is permissible at w.

Formalizing this interpretation requires using open sentences to handle the notion of substitution.

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

— In English, this statement roughly means that, if action A is permissible for some person p, then, for any person p, action A must be permissible. The notion of "permissible for" is captured by the substitution of x for p.

Let's check if this is already an axiom of DDL. If so, then the formalization is trivial.

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

This formalization doesn't hold in DDL, so adding it as an axiom will change the logic.

axiomatization where *FUL*: *FUL*

Consistency check: is the logic still consistent with the FUL added as an axiom?

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

— Nitpicking formula... 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. Now it's time to start testing!

A.2.3 Application Tests

In the following sections, I will use the application and metaethical tests presenting in Sections **??** and **??** to tease out the strengths and weaknesses of Kroy's formalization. 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 A.2.1, but some reveal philosophical problems with Kroy's interpretation of what the formula of universal law means. I will analyze these problems in the context of philosophical scholarship explicating the content of the formula of universal law. The findings in these sections will inform milestones for my custom formalization of the categorical imperative. They also serve as an example of how formalized and automated ethics can reveal philosophical strengths and weaknesses of an ethical theory.

Murder

In Section **??**, I began by testing the naive interpretation's ability to show that murder is wrong. I started by showing the morally dubious proposition that if murder is possibly wrong, then it is actually wrong.

consts $M::t$

— Let the constant M denote murder. I have defined no features of this constant, except that it is of the type term, which can be true or false at a set of worlds. Indeed, this constant as-is has no semantic meaning and could be replaced with any symbol, like 'Q' or 'Going to Target.' This constant will begin to take on features of the act of murder when I specify its properties. In the tests below, I specify its properties as the antecedents of lemmas. For example, the test below specifies that it is possible that murder is prohibited at the current world. This pattern will hold for most constants defined in Isabelle—they have no meaning until I program a meaning.

lemma *wrong-if-possibly-wrong*:

shows $((\Diamond\,(O\,\{\neg\,M\}))\ cw) \longrightarrow (\forall\, w.\ (O\,\{\neg\,M\})\ w)$

by *simp*

— This sentence reads: "If it is possible that murder is prohibited at world cw, then murder is prohibited at all worlds.

This is the same result we got in Section **??**—if murder is possibly wrong at some world, it is wrong at every world. The result is incredible strong—the mere possibility of wrongness at some world is sufficient to imply prohibition at every world.

Kroy's formalization shouldn't actually imply this property. Recall that this property held in the naive interpretation because it universalized a proposition across worlds (using the necessity operator). Kroy, on the other hand, interprets the FUL as universalizing across people, not worlds. In other words, Kroy's formulation implies that if murder is wrong for someone, then it is wrong for everyone.

The fact that this strange lemma holds is actually a property of DDL itself, not a property of Kroy's formalization. Indeed, repeating this experiment in DDL, with no additional axioms that represent the categorical imperative shows that, in DDL, if something is possibly wrong, it is wrong at every world. This implies that this is not a useful example to test any formulation. If a lemma is true in the base logic, without any custom axioms added, then it will hold for any set of custom axioms. Testing whether or not it holds as we add axioms tells us nothing, since it held in the base logic itself. Interesting cases are ones that fail (or are indeterminate) in the base logic, but become true as we add axioms.

To adapt the murder wrong axiom to capture the spirit of Kroy's formulation, I will modify if to state that if murder is wrong for one person, it is wrong for everyone.

consts *M-kroy*::*os*

— This time, murder is an open sentence, so that I can substitute in different agents.

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

shows $(\exists p. \models O\{\neg(M\text{-}kroy\ p)\}) \longrightarrow (\forall p. \models O\{\neg(M\text{-}kroy\ p)\})$

proof

assume $(\exists p. \models O\{\neg(M\text{-}kroy\ p)\})$

show $(\forall p. \models O\{\neg(M\text{-}kroy\ p)\})$

using *FUL* ‹$\exists p. \models \mathrm{O}\{\neg M\text{-}kroy\}\ p$› **by** *blast*

qed

This lemma gets to the heart of Kroy's formulation of the categorical imperative. If murder is prohibited for a specific person p, then it must be prohibited for all people[29].

Lying

For the naive implementation, I also tested the stronger proposition that if not everyone can simultaneously lie, then lying is prohibited. This is the equivalent of claiming that if lying fails the universalizability test, it is prohibited.

I want to represent the sentence"At all worlds, it is not possible that everyone lies simultaneously." This requires the following two abbreviations.

consts *lie*::*os*

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

— This represents the term "all people lie". Naively, we might think to represent this as $\forall p.lie(p)$. In HOL, the $\forall$ operator has type $('a \rightarrow bool) \rightarrow bool$, where $'a$ is a polymorphic type of the term being bound by $\forall$. In the given example, $\forall$ has the type $(s \rightarrow bool) \rightarrow bool$, so it can only be applied to a formula of type $s \rightarrow bool$. In the abbreviation above, we're applying the quantifier to a sentence that takes in a given subject p and returns $lie(p)w$ for any arbitrary w, so the types cohere.

— The term above is true for a set of worlds i (recall that a term is true at a set of worlds) 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."

Now that I have defined a sentence stating that lying fails the universalizability test, I can test if this sentence implies that lying is impermissible.

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

[29]This test case also revealed a bug in my original implementation of Kroy's formulation of the FUL, demonstrating the power of such automated tests and precise formulations to find bugs in ethical theories.

$p = s_1$

Kroy's formulation fails this test, and is thus not able to show that if lying is not possible to universalize, it is prohibited for an arbitrary person. To understand why this is happening, I will outline the syllogism that I *expect* to prove that lying is prohibited.

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

Armed with this syllogism, I can figure out why this test failed.

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 = $(\lambda x.\_)(a_1 := (\lambda x.\_)(i_1 := \text{False}))$ p = a_1

As we see above, the syllogism fails at Step 3, explaining why the lemma doesn't hold as expected. Kroy explicitly states[30] that this lemma holds in his logic.

The success of this lemma in Kroy's logic and the emptiness of his formalization of the FUL are two errors that contribute to the failure of this test. First, the statement expressed in

[30] See footnote 19 on p. 199

Step 3 should not actually hold. Impossible actions can be permissible (do I need a citation?). For example, imagine I make a trip to Target to purchase a folder, and they offer blue and black folders. No one would claim that it's impermissible for me to purchase a red folder, or, equivalently, that I am obligated to not purchase a red folder.

The second issue is that Kroy's interpretation of the formula of universal law is circular. 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 Kant (1785), 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 example demonstrates the power of formalized ethics. Making Kroy's interpretation of the categorical imperative precise demonstrated a philosophical problem with that interpretation.

A.2.4 Metaethical Tests

In addition to testing specific applications of the theory, I am also interested in metaethical properties, as in the naive interpretation. First, I will test if permissiblility is possible under this formalization.

lemma *permissible*:
 fixes $A\ w$
 shows $((\neg\,(O\,\{A\}))\wedge(\neg\,(O\,\{\neg\,A\})))\ w$
 nitpick [*user-axioms*, *falsify=false*] **oops**
— Nitpick found a model for card i = 1:
Free variable: A = $(\lambda x.\_)(i_1 := \text{False})$

The above result shows that, for some action A and world w, Nitpick can find a model where A is permissible at w. This means that the logic allows for permissible actions. If I further specify properties of A (such as 'A is murder'), I would want this result to fail.

Next, I will test if the formalization allows arbitrary obligations.

lemma *arbitrary-obligations*:
fixes A::t
shows $O\ \{A\}\ w$
nitpick [*user-axioms=true*, *falsify*] **oops**
— Nitpick found a counterexample for card i = 1 and card s = 1:
Free variable: A = $(\lambda x.\_)(i_1$:= False)

This is exactly the expected result. Any arbitrary action A isn't obligated. A slightly stronger property is "modal collapse," or whether or not 'A happens' implies 'A is obligated'.

lemma *modal-collapse*:
fixes $A\ w$
shows $A\ w \longrightarrow O\ \{A\}\ w$
nitpick [*user-axioms=true*, *falsify*] **oops**
— Nitpick found a counterexample for card i = 1 and card s = 1:
Free variables: A = $(\lambda x.\_)(i_1$:= True) w = i_1

This test also passes. Next, I will test if not ought implies can holds. Recall that I showed in Section **??** that ought implies can is a theorem of DDL itself, so it should still hold.

lemma *ought-implies-can*:
fixes $A\ w$
shows $O\ \{A\}\ w \longrightarrow \Diamond\ A\ w$
using *O-diamond* **by** *blast*

This theorem holds. Now that I have a substitution operation, I also expect that if an action is obligated for a person, then it is possible for that person. That should follow by the axiom of substitution Cresswell and Hughes (1996) which lets me replace the 'A' in the formula above with 'A(p)'

lemma *ought-implies-can-person*:
fixes $A\ w$
shows $O\ \{\ A(p)\}\ w \longrightarrow \Diamond\ (A\ (p))\ w$

using *O-diamond* **by** *blast*

This test also passes. Next, I will explore whether or not Kroy's formalization still allows conflicting obligations.

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)

Just as with the naive formalization, Kroy's formalization allows for contradictory obligations. Testing this lemma in DDL without the FUL shows that this is a property of DDL itself. This is a good goal to have in mind when developing my custom formalization.

Next, I will test the stronger property that if two maxims imply a contradiction, they may not be simultaneously willed.

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

Just as with the naive formalization, Kroy's formalization allows implied contradictions because DDL itself allows implied contradictions and Kroy's formalization doesn't do anything to remedy this.

Next, I will test that an action is either obligatory, permissible, or prohibited.

lemma *ob-perm-or-prohibited*:

fixes *A w*

shows $(O\,\{A\} \vee (P\,\{A\} \vee O\,\{\neg A\}))\ w$

by *simp*

— This test passes.

I also expect obligation to be a strictly stronger property than permissibility. Particularly, if A is obligated, then A should also be permissible.

lemma *obligated-then-permissible*:

shows $(O\ \{A\} \rightarrow P\ \{A\})\ w$

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

— This test fails in Kroy's interpretation! Nitpick found a counterexample for card i = 2 and card s = 1:

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

These tests show that, while Kroy's formalization is more powerful and more coherent than the naive formalization, it still fails to capture most of the desired properties of the categorical imperative. Some of these problems may be remedied by the fact that Kroy's logic doesn't allow contradictory obligations, and that possibility will be interesting to explore in my own formalization.

A.2.5 Miscellaneous Tests

In this section, I explore tests of properties that Kroy presents in his original paper. These tests not only test the features of the system that Kroy intended to highlight, but they may also inspire additional tests and criteria for my own formalization in Chapter 3. These tests further underscore the circularity of Kroy's formalization and the differences between my logic and his.

First, Kroy presents a stronger version of the formula of universal law and argues that his formalization is implied by the stronger version. Let's test that claim.

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 stronger version and $FUL \equiv \forall\, w\, A.\ (\exists\, p.\ P\ \{A\ p\}\ w) \longrightarrow (\forall\, p.\ P\ \{A\ p\}\ w)$ 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." In particular, 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 lied 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. It is tempting to claim that this issue explains why the tests above failed. To test this hypothesis, I will check if the stronger version of the FUL implies that lying is impermissible.

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

fixes p

shows *FUL-strong* $\longrightarrow \models(\neg\, P\ \{lie(p)\})$

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

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

Free variable: p = s_1

The test above also fails! This means 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.

Kroy also argues that the FUL gives us recipes for deriving obligations, in addition to deriving permissible actions. Specifically, he presents the following two principles, which are equivalent in his logic. These sentences parallel FUL and strong FUL.

abbreviation *obligation-universal-weak*::*bool* **where** *obligation-universal-weak* $\equiv \forall\, w\, A.\ ((\exists p{::}s.\ ((O\ \{A\ p\})\ w))\ \longrightarrow (\ (\forall p.\ (\ O\ \{A\ p\ \})\ w)))$

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

— Just as with FUL and FUL strong, the weaker version of the above statement has the consequent, “For all people, A is obligated." The stronger consequent is “A is obligated for all people simultaneously."

lemma *weak-equiv-strong*:
 shows *obligation-universal-weak* $\equiv$ *obligation-universal-strong*
 oops

— Isabelle is neither able to find a proof nor a countermodel for the statement above, so I can’t say if it holds or not without completing a full, manual proof. This aside is not very relevant to my project, so I will defer such a proof.

These two statements are not necessarily equivalent in my logic, but are in Kroy’s[31] This difference in logics may further explain why tests are not behaving as they should. Nonetheless, Kroy argues that the FUL implies both statements above.

lemma *FUL-implies-ob-weak*:
 shows *FUL* $\longrightarrow$ *obligation-universal-weak* **oops**

— Isabelle is neither able to find a proof nor a countermodel for this statement.

lemma *FUL-implies-ob-strong*:
 shows *FUL* $\longrightarrow$ *obligation-universal-strong* **oops**

— Isabelle is neither able to find a proof nor a countermodel for this statement.

Isabelle timed out when looking for proofs or countermodels to the statements above. This may be an indication of a problem that Benzmueller warned me about—mixing quantifiers into a shallow embedding of DDL may be too expensive for Isabelle to handle. Not sure

[31] This follows from the fact that the Barcan formula holds in Kroy’s logic but not in mine, as verified with Nitpick. See Appendix for more.

what to do about this.

end

A.3 Testing Un-Universalizable Actions

I will show that the maxim, "When strapped for cash, falsely promise to pay your friend back to get some easy money." is prohibited. This example is due to Korsgaard and she uses it to highlight the strength of her preferred interpretation of the FUL, the practical contradiction interpretation Korsgaard (1985). There are two possible readings of this maxim, and I will show that my formalization can handle both. Under the first reading, the act of falsely promising is read as 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. The differences between these readings lies in the difference between promising as an act with meaning in a larger social structure and the utterance "I promise."

Under the first reading, the maxim fails because falsely promising is no longer possible in a world where everyone everyone does so. This is how the logical contradiction interpretation reads this maxim—falsely promising is no longer possible when universalized because the institution of promising breaks down. The practical contradiction view also prohibits this maxim because if falsely promising is not longer possible, then it is no longer an effective way to achieve the end of getting some money. Below I define some logical tools to formalize this reading of this maxim.

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

— Constant representing the circumstances "when strapped for cash." Recall that the type of circumstances is a term because circumstances can be true or false at a world.

consts *falsely-promise*::*os*

— Constant representing the act "make a false promise to pay a loan back." Recall that the type of an act is an open sentence because the sentence "subject s performs act a" can be true or false at a world.

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

— Constant representing the goal "to get some money." Recall that the type of a goal is a term because a goal can be true or false at a world depending on whether it is achieved or not.

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

$\textit{false-promising} \equiv (\textit{when-strapped-for-cash}, \textit{falsely-promise}, \textit{to-get-easy-cash})$

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

The logical objects above are "empty," in the sense that I haven't specified any of their relevant properties. I will define these properties as assumptions and will show that, if the maxim above satisfies the assumed properties, it is prohibited.

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

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

— Under this reading, the problem with this maxim is that everyone can't falsely promise simultaneously because the institution of promising will break down. It's probably possible to say something stronger than this (i.e. that if enough but not necessarily all people falsely promise promising is no longer possible), but for my purposes this will suffice. The above formula reads, "At all worlds, it is not the case that everyone falsely promises."

abbreviation *circumstances-hold* **where**

$\textit{circumstances-hold} \equiv \forall\, w.\ \textit{when-strapped-for-cash}\ w$

— This assumption narrows our scope of consideration to worlds where the circumstances of being strapped for cash hold. This is important because, at worlds where the circumstances do not hold, a maxim is trivially effective (since it's never acted on) and thus trivially universalizable. This assumption also makes practical sense; when evaluating a maxim, an agent would care about it specifically at worlds where the circumstances hold, since these are the worlds where the maxim actually prescribes action.

abbreviation *example-is-well-formed* **where**

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

— This assumption states that the maxim of falsely promising is well-formed. This breaks down into two individual assumptions. First, being strapped for cash can't imply falsely promising, which is plausible because many people won't falsely promise under conditions of poverty. Second, being strapped for cash can't imply getting ready cash, which is also plausible because people often fail to secure cash even when they need it.

Putting it all together, I want to show that if the three assumptions justified above hold, then the constructed maxim is prohibited. Below is the proof

lemma *lying-bad-1*:
 assumes *everyone-can't-lie*
 assumes *circumstances-hold*
 assumes *example-is-well-formed*
 shows $\forall s. \models$ (*prohibited false-promising s*)
proof –
 have $\forall s$. *not-universalizable false-promising s*
  by (*simp add*: *assms*(*1*) *assms*(*2*))
— I manually broke the proof into this intermediate lemma and the conclusion, and then Sledgehammer automatically found a proof.
 thus *?thesis*
  using *FUL assms*(*3*) **by** *blast*
qed

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[32]. 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. Below I formalize this reading of this maixm.

consts *believed*::*os*
abbreviation *false-promising-not-believed* **where**
false-promising-not-believed $\equiv \forall w\, s.$ (*falsely-promise*(*s*) *w* $\longrightarrow \neg$ *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**
need-to-be-believed $\equiv \forall w\, s.$ ($\neg$ *believed*(*s*) *w* $\longrightarrow \neg$((*falsely-promise s*) $\rightarrow$ *to-get-easy-cash*)*w*)

[32]Note that under this reading, the maxim isn't prohibited under the logical contradiction interpretation because making an utterance is still possible even if eveyrone else makes that utterance. I will discuss this in detail later in this section in the context of the difference between natural and conventional acts.

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

lemma *falsely-promising-bad-2*:
assumes *false-promising-not-believed*
assumes *need-to-be-believed*
— The above two assumptions are specific to this reading and justified above.
assumes *circumstances-hold*
assumes *example-is-well-formed*
— These two assumptions applied to the first reading as well and were justified there.
shows $\forall s. \models (prohibited\ false\text{-}promising\ s)$
proof –
have $\forall s.$ *not-universalizable false-promising s*
using *assms*(*1*) *assms*(*2*) *assms*(*3*) **by** *auto*
thus *?thesis*
using *FUL assms*(*4*) **by** *blast*
qed
— With some help, Isabelle is able to show that the maxim is prohibited under this reading as well.

This example demonstrates that my formalization is able to correctly prohibit this maxim, regardless of its reading. This is additionally important because the two readings of this maxim represent reading the act as either a conventional or natural action, so my intrepretation can correctly handle both kinds of actions. Korsgaard draws a distinction between conventional acts and natural acts. 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). For example, promising is a conventional act because it only 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).

This distinction is important because Korsgaard argues that only the practical contradiction view can satisfactorily explain the wrongness of certain natural acts like murder[33]. The practical contradiction view is thus stronger than the logical contradiction view because it can explain the wrongness of both conventional and natural acts.

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. The first reading of the example maxim reads the act "making a false promise" as entering into an agreement within a socially established system of promising. This is clearly a conventional act, and because it is a conventional act, it is not just contradictory when universalized but literally impossible because the practice breaks down. I capture this idea in the assumption *appendix-2.everyone-can't-lie* $\equiv \forall\, w.\ \neg\ (\forall\, s.$ *appendix-2.falsely-promise* $s\ w)$, which states that, at all worlds, not everyone can falsely promise since otherwise the practice of promising would break down. The second reading, on the other hand, reads the act of making a false promise as uttering the statement "I promise to pay you back," while never intending to fulfill this promise. 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[34]

I show above that my formalization shows the wrongness of this maxim under both readings. Under the first reading, promising becomes impossible, so both the logical and practical contradiction interpretations prohibit the maxim. Under the second reading, promising is still possible, but becomes ineffective because people no longer interpret the utterance as creating a commitment. Under this view, only the practical contradiction interpretation succeeds in prohibiting the maxim. Thus, not only does my formalization likely capture the practical contradiction interpretation (as opposed to the teleological or logical contradiction interpretations), it also adequately handles both natural and conventional acts.

I can also use Isabelle to confirm that the two readings are different. If they were the

[33]For more discussion of Korsgaard's argument for the practical contradiction view, see Section Philosophical Writing

[34]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.

same, we would expect the assumptions corresponing to each to be equivalent. The RHS of the lemma below represents the second reading and the LHS represents the first reading.

lemma *readings-are-equivalent*:

shows *false-promising-not-believed* $\wedge$ *need-to-be-believed* $\equiv$ *everyone-can't-lie*

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

— Nitpick finds a counterexample, showing that the two readings are different. Nitpick found a counterexample for card i = 1 and card s = 1:

Empty assignment

end

References

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

L. Alexander and M. Moore. Deontological Ethics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

M. Anderson and S. Anderson. Geneth: A general ethical dilemma analyzer. volume 1, 07 2014.

M. Anderson and S. L. Anderson. Ethel: Toward a principled ethical eldercare robot.

M. Anderson, S. Anderson, and C. Armen. Towards machine ethics. 07 2004.

Aristotle. The nicomachean ethics. *Journal of Hellenic Studies*, 77:172, 1951. doi: 10.2307/628662.

K. Arkoudas, S. Bringsjord, and P. Bello. Toward ethical robots via mechanized deontic logic. *AAAI Fall Symposium - Technical Report*, 01 2005.

E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon. 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, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1911517117. URL https://www.pnas.org/content/117/5/2332.

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

C. Benzmüller, A. Farjami, and X. Parent. Dyadic deontic logic in hol: Faithful embedding and meta-theoretical experiments. In M. Armgardt, H. C. Nordtveit Kvernenes, and S. Rahman, editors, *New Developments in Legal Reasoning and Logic: From Ancient Law to*

Modern Legal Systems, volume 23 of *Logic, Argumentation & Reasoning*. Springer Nature Switzerland AG, 2021. ISBN 978-3-030-70083-6. doi: 10.1007/978-3-030-70084-3.

N. Berberich and K. Diepold. The virtuous machine - old ethics for new technology?, 2018.

J. C. Blanchette and T. Nipkow. *Nitpick: A Counterexample Generator for Higher-Order Logic Based on a Relational Model Finder*, volume 6172, page 131–146. Springer Berlin Heidelberg, 2010. ISBN 9783642140518. doi: 10.1007/978-3-642-14052-5_11. URL http://link.springer.com/10.1007/978-3-642-14052-5_11.

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

J. Carmo and A. Jones. Completeness and decidability results for a logic of contrary-to-duty conditionals. *J. Log. Comput.*, 23:585–626, 2013.

J.-A. Cervantes, L.-F. Rodríguez, S. López, and F. Ramos. A biologically inspired computational model of moral decision making for autonomous agents. In *2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing*, pages 111–117, 2013. doi: 10.1109/ICCI-CC.2013.6622232.

R. M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis (Oxford)*, 24(2): 33–36, 1963. ISSN 0003-2638.

C. Cloos. The utilibot project: An autonomous mobile robot based on utilitarianism. *AAAI Fall Symposium - Technical Report*, 01 2005.

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

D. Davenport. Moral mechanisms. *Philosophy and Technology*, 27(1):47–60, 2014. doi: 10.1007/s13347-013-0147-2.

L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14, 2016. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2015.11.012. URL https://www.sciencedirect.com/science/article/pii/S0921889015003000.

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

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

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

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

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

I. Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, Sep 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL http://dx.doi.org/10.1007/s11023-020-09539-2.

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

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

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

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

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

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

I. Kant. *Introduction*, pages ix–xxix. Cambridge Texts in the History of Philosophy. Cambridge University Press, 2 edition, 2017. doi: 10.1017/9781316091388.002.

J. Kemp. Kant's examples of the categorical imperative. *The Philosophical Quarterly (1950-)*, 8(30):63–71, 1958. ISSN 00318094, 14679213. URL http://www.jstor.org/stable/2216857.

P. Kitcher. What is a maxim? *Philosophical Topics*, 31(1/2):215–243, 2003. doi: 10.5840/philtopics2003311/29.

M. Kohl. Kant and 'ought implies can'. *The Philosophical Quarterly (1950-)*, 65(261):690–710, 2015. ISSN 00318094, 14679213. URL http://www.jstor.org/stable/24672780.

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

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

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

C. M. Korsgaard. Acting for a reason. *Danish Yearbook of Philosophy*, 40(1):11–35, 2005. doi: 10.1163/24689300_0400103.

C. M. Korsgaard and O. O'Neill. *The Sources of Normativity*. Cambridge University Press, 1996. doi: 10.1017/CBO9780511554476.

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

D. Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973a. doi: 10.2307/2025310.

D. Lewis. *Counterfactuals*. Blackwell, 1973b.

P. Lukowicz. The challenge of human centric ai. *Digitale Welt*, 3:9–10, 10 2019. doi: 10.1007/s42354-019-0200-0.

E. McRae. Equanimity and intimacy: A buddhist-feminist approach to the elimination of bias. *Sophia*, 52(3):447–462, 2013. doi: 10.1007/s11841-013-0376-y.

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

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

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

O. O'Neill. *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge University Press, 1990. doi: 10.1017/CBO9781139173773.

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

L. Paulson and J. Blanchette. Three years of experience with sledgehammer, a practical link between automatic and interactive theorem provers. 02 2015. doi: 10.29007/tnfd.

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

T. M. Powers. Prospects for a kantian machine. *IEEE Intelligent Systems*, 21(4):46–51, 2006. doi: 10.1109/MIS.2006.77.

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

J. Rawls. Kantian constructivism in moral theory. *The Journal of Philosophy*, 77(9):515–572, 1980. ISSN 0022362X. URL http://www.jstor.org/stable/2025790.

D. Rönnedal. Contrary-to-duty paradoxes and counterfactual deontic logic. *Philosophia*, 47, 09 2019. doi: 10.1007/s11406-018-0036-0.

D. Scott. Advice on modal logic. In K. Lambert, editor, *Philosophical Problems in Logic: Some Recent Developments*, pages 143–173. D. Reidel, 1970.

J. R. Silber. Procedural formalism in kant's ethics. *The Review of Metaphysics*, 28(2):197–236, 1974. ISSN 00346632. URL http://www.jstor.org/stable/20126622.

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

K. Solt. Deontic alternative worlds and the truth-value of 'oa'. *Logique et Analyse*, 27(107): 349–351, 1984. ISSN 00245836, 22955836. URL http://www.jstor.org/stable/44084096.

N. F. Stang. Kant's Transcendental Idealism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

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

J. D. Velleman. *A Brief Introduction to Kantian Ethics*, page 16–44. Cambridge University Press, 2005. doi: 10.1017/CBO9780511498862.002.

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

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

A. Winfield, C. Blum, and W. Liu. Towards an ethical robot: Internal models, consequences and ethical action selection. volume 8717, 09 2014. ISBN 978-3-319-10400-3. doi: 10.1007/978-3-319-10401-0_8.

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