Automating Kantian Ethics

Lavanya Singh

December 6, 2021

Contents

1 Introduction

As artifical reasoners become increasingly powerful, computers become capable of performing complex ethical reasoning. Moreover, as regulators and academics call for "ethical AI," the field of machine ethics becomes increasingly popular. Much prior work on machine ethics implements relatively simple philosophical theories, partially as an artifact of this work emerging out of Computer Science or Mathematics departments (Tolmeijer et al., 2021). Such work rarely capitalizes on the centuries of philosophical debate discussing the ethical theories in question. In this thesis, I attempt to automate Kantian ethics, a sophisticated ethical theory, while staying faithful to philosophical literature on the subject.

Automating a sophisticated ethical theory is interesting for two reasons. First, the proliferation of artifically autonomous agents is creating and will continue to create a demand for automated ethics. These agents must be able to reason about complex ethical theories that withstand philosophical scrutiny. Second, just as automated mathematical reasoning is a tool for mathematicians, automated ethical reasoning is a tool that philosophers can use when reasoning about ethics. I argue that computational ethics can serve as another tool in a philosopher's toolbox, like a thought experiment or counterexample.

Modelling ethics without sacrificing the intracacies of an ethical theory is a challenging computational and philosophical problem. Simple and intuitive computational approaches, such as encoding ethics as a constraint satisfaction problem, fail to capture the complexity of most philosophically plausible systems. Constraint satisfaction systems often default to some version of utiliarianism, the principle of doing the most good for the most people. Alternatively, they model basic moral principles such as "do not kill," without modelling the theory that these principles originated from. Modelling a more complex ethical theory will not only enable

smarter philosophical machines, it will also empower philosophers to study more complex ethical issues with the computer's help. The entire field of philosophy is devoted to developing and testing robust ethical theories. Plausible machine ethics must draw on plausible moral philosophy. Despite the importance of formalizing complex moral theories, it is not immediately clear how to formalize complex theories like virtue ethics.

Kantian ethics, often described as "formal" due to its rigid, rule-based structure, has been often floated as an attractive theory to automate (Powers, 2006; Bentzen and Lindner, 2018; Lin et al., 2012). This project's objective is to automate Kantian ethics. Roughly, my approach is to represent Kantian ethics as an axiom in Carmo and Jones' Dyadic Deontic Logic (DDL), a modal logic designed to reason about concepts like obligation (Carmo and Jones, 2013). Such a representation of Kantian ethics in a logic is called a "formalization." I then embed this DDL formalization in higher-order logic (HOL) and implement it in the Isabelle/HOL theorem prover (Nipkow et al., 2002). I can then use Isabelle to automatically prove or disprove theorems (such as, "murder is wrong") in my custom formalization, generating results derived from the categorical imperative. Essentially, the computer is performing ethical reasoning using the framework of Kantian ethics.

First, I recreate prior work implementing Dyadic Deontic Logic in Isabelle/HOL. Next, I present two prior attempts to formalize Kantian ethics in modal logic. The first is a naive interpretation of Kant's categorical imperative that collapses to the base logic itself. The second is Moshe Kroy's partial formalization of the categorical imperative. As part of this presentation, I also create a testing framework that can be used to evaluate these formalizations. In effect, my tests formalize expected properties of the categorical imperative (such as the fact that it prohibits murder) and test whether or not these properties hold in a given formalization. My testing framework offers a method for evaluating different formalizations against

moral intuitions and philosophical literature. Second, based on lessons learned from these prior formalization attempts, I contribute my own custom formalization of the categorical imperative. My testing framework demonstrates how my custom formalization improves upon the prior work. Third, I apply my system to an ethical dilemma to demonstrate its potential use. I also consider the philosophical implications and value of automated ethics, studying questions about the possibility and responsibility of automated ethics.

I contribute implementations of two different interpretations of the categorical imperative, examples of how each implementation can be used to model and solve ethical scenarios, and tests that examine ethical and logical properties of the system, including logical consistency, consistency of obligation, and possibility of permissibility. I also contribute a logical formalization of the categorical imperative that improves on previous work, an implementation of this formalization, and evidence of its improvement. Lastly, I demonstrate how such a system could be applied. The implementations themselves are usable models of ethical principles and the tests represent the kind of philosophical work that formalized ethics can contribute.

The rest of this project is structured as follows. In Chapter 1, I introduce the logical, computational, and philosophical frameworks underlying my project. In Chapter 2, I implement and test prior formalizations of the categorical imperative. In Chapter 3, I contribute my own custom formalization and test it. Finally, in Chapter 4, I apply my custom formalization to an example ethical dilemma.

In the rest of this chapter, I present the background necessary to undersand my work. The goal of this project is to automate sophisticated ethical reasoning. This requires three components. First, the choice of an ethical theory that is both intuitively attractive and lends itself to formalization. Second, the choice of formal logic to model the theory in. Third, the choice of automation engine to implement

the formal model in. Section **??** introduces Kantian ethics, Section 1.2 explains Carmo and Jones's Dyadic Deontic Logic as a base logic, and Section 1.3 presents the Isabelle/HOL implementation of the logic.

1.1 Choice of Ethical Theory

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 deontology offers one such procedure for the problem of making ethical judgements. I will make this intuition precise by arguing that deontological ethics is natural to formalize because rules generally require little additional data about the world and are usually easy to represent to a computer. 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 deontological ethics than for consequentialism and virtue ethics.

I do not aim to show that deontology 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. Future work could and should address the challenges I outline in this section. The more ethical theories that computational tools can handle, the more valuable computational philosophy becomes both for philosophers and for AI agents. Insofar as my project serves as an early proof-of-concept for computational ethics, I choose to automate an ethical theory that poses fewer challenges than others.

I first present deontological ethics, then consequentialism, and finally virtue 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 deontology. Finally, I will present the specific deontological theory I am automating (Kantian ethics) and will argue that it is comparatively easier to formalize. I will also outline the specific debates in the literature that my formalization takes a stance on and potential challenges for formalizing deontology.

1.1.1 Deontological Ethics

Crash Course on Deontology

Deontological ethical theories evaluate actions as permissible, obligatory, or prohibited. The deontological tradition argues that an action should not be judged on its consequences, but rather on "its confirmity with a moral norm" (Alexander and Moore, 2021). In other words, deontological theories define a set of moral norms or rules and evaluate actions based on their confirmity or lackthereof to these rules. Deontologists do not believe that an agent should maximize the number of times that they conform to such rules; rather they argue that an agent should never violate any of the moral laws. A wrong choice is wrong, regardless of its consequences.

Formalizing Deontology

Deontology is immediately an attractive candidate for formalization because computers tend to understand rules; programming languages are designed to teach computers algorithms. Deontological ethical theories give inviolable rules that an automated agent can apply, without having to adjust the rules in changing situations or contexts. Moreover, because deontological theories focus on the action itself, they require relatively little data. A deontological moral judgement does not require as much information about context, consequences, or moral character as the other theories presented later in this section. All that matters is the action and some limited set of circumstances in which it is performed. I will later argue that, in the case of the specific deontological ethic that I implement (Kantian ethics), the action has

a very "thin" representation that is space and memory efficient and requires little data about the external world.

Like all ethical traditions, deontology has debates that any implementation of automated deontological ethics would need to resolve. Deontologists disagree about whether ethics should focus on agents and prescribe them action or should focus on the patients or potential victims of actions and their rights. Different deontological theories have different conceptions of what an action is, from the actual physical act to the agent's mental state at the time of acting to the principle upon which the agent acted.

While these debates are certainly open, "if any philosopher is regarded as central to deontological moral theories, it is surely Immanuel Kant" (Alexander and Moore, 2021). Out of the three ethical traditions considered in this section, deontology has the most central representative in the form of Kant. Many modern deontologists claim to interpret Kant in a particular way, and thus agree on his ethic as the foundation of their theory but disagree in how to interpret it. In this paper, I will formalize Kantian ethics. Deontology's comparatively greater focus on Kant means that the choice of Kant as a guiding figure will be less controversial to deontologists than, for example, the choice of Bentham as the guiding figure of consequentialism. Moreover, at the end of this section, I also argue that internal debates in the part of Kantian ethics that I focus on tend to be less controversial than those in the consequentialist or virtue ethical traditions.

I do not argue that deontology is the only tractable theory to formalize, but instead that it is easier to formalize because it requires less data, is easily representable to a computer, and has fewer and less controversial open debates than consequentialism or virtue ethics. That being said, any ethical tradition has debates that an automated agent will need to take a stance on and deontological ethics is no exception. Those who disagree with my stances (e.g. non-Kantians) will not trust my

system's judgements. This project is not irrelevant for them because it still serves as a case study in the power of computational ethics, but Kantians will trust my system's moral judgements the most.

1.1.2 Consequentialism

A consequentialist ethical theory is, broadly speaking, any ethical theory that evaluates an action by evaluating its consequences.[1] For example, utilitarianism is a form of consequentialism in which the moral action is the action that results in the best consequences or produces the most good (Driver, 2014). The focus on the consequences of action distinguishes consequentialists from deontologists, who derive the moral worth of an action from the action itself. Some debates in the consequentialist tradition include which consequences of an action matter, what exactly 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 deontology. Under a naive version of consequentialism, evaluating an action requires perfect knowledge of all consequences following an action. This requires that an automated ethical system somehow collect all of the infinite consquences following an action, a difficult, if not impossible, task. Moreover, compiling this database of consequences requires answering difficult questions about which consequences were actually caused by an action.[2]

These challenges also apply to human reasoners, so most consequentialists do not

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

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

actually adopt the naive view that agents need to calculate all the consequences of their actions. Plausible strategies to avoid this problem include stopping calculation early because constant calculation paralyzes action or only evaluating consequences that the agent could reasonably forsee before acting. Another solution is the "proximate cause" approach, which only holds the agent responsible for the immediate consequences of their acts, but not for consequences resulting from others' voluntary responses to the agent's original act (Sinnott-Armstrong, 2021).

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

The fact that consequentialism requires more knowledge about the world makes it more difficult to formalize. Automated consequentialist ethical systems would need to represent complex states of the world and causal chains in an efficient manner and reason about them. This both presents a difficult technical challenge and impedes the usability of such a system. Such a system would need to come

equipped with a large enough database of knowledge about the world to extrapolate the consequences of an actions and up-to-date information about the state of the world at the moment of action. Not only does collecting and representing this data pose a technical challenge, it also creates a larger "trusted code base" for the automated system. Trusting my deontological ethical reasoner merely requires trusting the logical implementation of the categorical imperative and the formulation of a maxim. Trusting a consequentialist ethical reasoner, on the other hand, requires trusting both the logical machinery that actually evaluates the act and the background/situational knowledge that serves as an input to this machinery.

The challenge of understanding and representing the circumstances of action is not unique to consequentialism, but is particularly acute for consequentialism. Deontologists robustly debate which circumstances of an action are "morally relevant" and should be included in the formulation of the action.[3] Those using my system will need to use common-sense reasoning to determine which aspects of the circumstances in which the action is performed are morally relevant and should thus be represented to the computer.[4] However, because deontology merely evaluates the action itself, the surface of this debate is much smaller than the debate about circumstances and consequences in a consequentialist system. An automated consequentialist system must make judgements about the act itself, the circumstances in which it is performed, and the circumstances following the act. The "trusted code base" is smaller for deontology than for consequentialism. All ethical theories will require some set of circumstances and common-sense knowledge as part of the trusted code base, but this set is larger for consequentialism than for deontology.

Theory of the Good

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

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

Another debate that an automated consequentialist reasoner would need to take a stance on is the question of what qualifies as a "good consequence," or what the theory of the good is. Hedonists associate good with the presence of pleasure and the absence of pain. Preference utiliarians believe that good is the satisfaction of desires and is thus derived from individuals' preferences, as opposed to some sensation of pleasure or pain. Other consequentialists adopt a pluralistic theory of value, under which many different kinds of things are good for different reasons. For example, Moore values beauty and truth and other pluralists value justice, love, and freedom (Moore, 1903). Welfare utilitarians value a person's welfare and utilitarians of right value states of affairs in which respect for some set of rights is maximized (Sinnott-Armstrong, 2021).

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

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

To solve this problem, some consequentialists reject aggregation entirely and instead prefer wholistic evaluations of a state of affairs. While this approach no longer requires that a reasoner define an aggregation algorithm, the reasoner still needs to calculate a goodness measurement for a state of affairs. Whereas before the reasoner could restrict analysis to a single person, the algorithm must now evaluate an entire state wholistically. Evaluating the goodness of an entire state of affairs is more complicated than evaluating the goodness of a single person. As consequentialists modulate between aggregation and wholistic evaluation,

they face a tradeoff between the difficulty of aggregation and the complexity of goodness measurements for large states of affairs. This tradeoff also holds for an automated consequentialist moral agent. Such an agent either needs to define an aggregation function, thus opening the door to critique from those who disagree with this definition, or needs to evaluate the goodness of entire states of affairs, which is a complex and data-intensive philosophical and technical challenge.

Prior Attempts to Formalize Consequentialism

None of the challenges described above are intractable or capture the full literature of all variations of consequentialism. Instead, the challenges above require that the developer "plant certain flags" and take a stance on certain philosophical debates. Such debates are present in any ethical theory, but consequentialism has more such points of difficulty than deontology and is thus more difficult to automate.

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.[5] It leaves open the question of what the reward function is and thus leaves the theory of the good, arguably the defining trait of a particular consequentialist view, undefined. Similarly, Anderson and Anderson's proposal of a hedonistic act utilitarian automated reasoner chooses hedonism[6] 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

[5]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 $\sum_{i=0}^{\infty} \gamma^i r_i$.

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

question. Other attempts to automate consequentialist ethics will suffer similar problems because, at some point, a useful automated consequentialist moral agent will need to resolve the above debates.

1.1.3 Virtue Ethics

What Is Virtue

The virtue ethical tradition places the virtues, or those traits that constitute a good moral character, at the center. Virtue ethicists evaluate actions based on the character traits that such actions would help cultivate. A virtue is commonly accepted as a character trait that "makes its possessor good" (Hursthouse and Pettigrove, 2018). For example, under Aristotlean virtue ethics, virtues are the traits that enable human flourishing or fulfill the purpose of a human being. Many modern virtue ethicists abandon Aristotle's notion of a "purpose" of human beings, and instead define virtue in terms of the characteristic activity of human beings (in ethical terms, not telelological terms) (Snow, 2017). Just as consequentialists must offer a view of which consequences are good, virtue ethicists must offer some theory of the virtues which presents and justifies a list of the virtues. Such theories vary from Aristotle's virtues of courage and temperance to the Buddhist virtue of equanimity (Aristotle, 1951; McRae, 2013). Another theory is Sen's conception of the virtues as capabilites that create "effective opportunities to undertake the actions and activities" an agent wants to engage in (Robeyns, 2005). An automated virtue ethical agent will need to commit to a particular theory of the virtues, opening itself up to criticism from those who disagree with this theory of the virtues. Any automated virtue ethical agent will need to justify its choice of virtues.

Evaluating Moral Character

Another difficulty with automating virtue ethics is that the unit of evaluation for

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

Machine Learning and Virtue Ethics

One potential appeal of virtue ethics is that many virtue ethical theories involve some form of moral habit, which seems to 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). Under one view of virtue ethics, the virtuous act is what the virtuous person would do. Both of these ideas imply that ethical behavior can be learned from some dataset of virtuous acts, either those prescribed by a moral teacher or those that a virtuous ideal agent would undertake. Indeed, these theories seem to point towards a machine learning approach to computational 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 also may suffer explanability problems that my logical, theorem-prover based approach does not experience. Many machine learning algorithms cannot sufficiently explain their decisions to a human being, and often find patterns or correlations in datasets that don't actually cohere with the trends and causes that a human being would identify (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. Indeed, 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.

Explanability is particularly important in the case of ethics because ethical judgements are often controversial and ethics generally requires reflection. Often, the most interesting and important ethical judgements result from ethical dilemmas. These judgements are usually controversial because people's intuitions differ and different theories generate different answers. In these cases, explainability is particularly important to convince human beings of the correctness of an ethical judgement. If a machine tells us to kill one person to save five without justifying this decision, acting on this judgement becomes difficult. Second, ethics is a reflective subject. Practical reason is the exercise of using reason to decide what to do. Someone who believes an automated reasoner's judgements without examining or understanding the reasons for these judgements doesn't seem to be doing ethics correctly.[7] This does not preclude other uses of automated ethics, such as automated moral agents or hypothesis generation for philosophy, but it does make

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

computer-assisted ethical judgement difficult.

My arguments about theories of virtues and explanability are in the context of virtue ethics and machine learning. Such arguments also apply to a broader class of projects in automated ethics that use "bottom-up" approaches, in which a system learns moral judgements from prior judgements, as opposed to a top-down ethical theory. I will extend this argument to bottom-up approaches more generally in Section Related Work.

1.1.4 Kantian Ethics

As mentioned above, in this paper I focus on Kantian ethics, a specific branch of deontology. Kant is widely seen as the most popular representative of deontology, so this choice is not surprising. In this section, I will present a crash course on Kant's ethical theory and then explain why his particular theory is more amenable to formalization than consequentialist or virtue ethical theories.

Crash Course 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. In *The Groundwork of the Metaphysics of Morals*, Kant's most influential text on ethics, he explains that rational beings are unique because we can act "in accordance with the representations of laws" (Kant, 1785, 4:412). 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. A rational being's behavior is purposive and their actions are guided by practical reason. They have reasons for acting, even when these reasons

may be opaque to them. This 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). There is debate about what exactly must be included in a maxim, but many philosophers agree that a maxim consists of some combination of circumstances, act, and goal.[8] 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 certain maxims have a form or logical structure that requires any rational agent to will them, and these maxims are obligatory.

The form of an obligatory maxim is given by the categorical imperative. 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 under all circumstances. Kant argues that the moral law must be a categorical imperative, for otherwise it would not have the force that makes it a moral law (Kant, 1785, 5). In order for an imperative to be categorical, it must be derived from the will's authority over itself. Our wills are autonomous, so the only thing that can have unconditional authority over a rational will is the rational will itself. In Velleman's version of this argument, he claims that 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

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

reason. To question this authority is to demand a reason for acting for reasons, which concedes the authority of reason itself (Velleman, 2005, 23). Therefore, the only possible candidates for the categorical imperative are those rules that are required of the will because it is a will. The categorical imperative must be a property of practical reason itself.

Armed with this understanding of practical reason, Kant presents the categorical imperative. He presents three "formulations" or versions of the categorical imperative and goes on to argue that all three formulations are equivalent. In this project, I focus on the first formulation, the Formula of Universal Law, but will briefly present the other two as well.[9]

The first formulation of the categorical imperative is the Formula of Universal Law (FUL), which reads, "act only according to that maxim through which you can at the same time will that it become a universal law" (Kant, 1785, 34). This formulation generates the universalizability test, which tests the moral value of a maxim by imagining a world in which it becomes a universal law and attempting to will the maxim in that world. If there is a contradiction in willing the maxim in a world in which everyone universally wills the maxim, the maxim is prohibited. Velleman presents a concise argument for the FUL. He argues that reason is universally shared among reasoners. For example, all reasoners have equal access to the arithmetic logic that shows that "2+2=4" (Velleman, 2005, 29). The chain of reasoning that makes this statement true is not specific to any person, but is universal across people. Therefore, if I have sufficient reason to will a maxim, so does every other rational agent. There is nothing special about the operation of my practical reason that other reasoners don't have access to. Practical reason is shared, so in adopting a maxim, I implicitly state that all reasoners across time also have reason to adopt that maxim. Therefore, because I act on reasons, I must obey the FUL. Notice

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

that this fulfills the above criterion for a categorical imperative: the FUL is derived from a property of practical reason itself and thus derives authority from the will's authority over itself, as opposed to some external authority.

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

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

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, or a formal principle of practical reason. It does not require any situational knowledge or contingent beyond the circumstances included in the maxim itself and thus requires far less contingent facts than other ethical theories. Instead, it is purely a property of the proposed principle for action. This formalism

makes Kantian ethics an attractive candidate for formalization. While other ethical theories often rely on many facts about the world or the actor, Kantian ethics simply relies on the form of a given maxim. A computer evaluating a maxim doesn't require any knowledge about the world beyond what is contained in a maxim. A maxim is the only input that the computer needs to make a moral judgement. Automating Kantian ethics merely requires making the notion of a maxim precise and representing it to the computer. This distinguishes Kantian ethics from consequentialism and virtue ethics, which, as I argued above, require far more knowledge about the world or the agent to reach a moral decision.

Not only does evaluating Kantian ethics focus on a maxim, a maxim itself is an object with a thin representation for a computer, as compares 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 representation of a causal chain or a state of affairs or moral character. A maxim is a principle with a well-defined form, so representing a maxim to the computer merely requires capturing this form. This property not only reduces the computational complexity (in terms of time and space) of representing a maxim, it also make the system easier for human reasoners to interact with. A person crafting an input to a Kantian automated agent needs to reason about relatively simple units of evaluation, as opposed to the more complex features that consequentialism and virtue ethics require. I will make the comparision to consequentialism and virtue ethics explicit below.

Difficulties in Automation

My choices to interpret maxims and the Formula of Universal Law in a particular way represent debates in Kantian ethics over the meanings of these terms that I take a stance on. Another debate in Kantian ethics is the role of "common-sense"

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

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. My system does not need to answer this question because I assume a well-formed maxim as input and apply the categorical imperative to this input, but if my system were ever to be used in a faully automated agent, answering this question would require significant computational and philosophical work. For more, see Section AI Ethics.

Common-sense reasoning is also relevant in applying the universalizability test itself. Consider an example maxim tested using the Formula of Universal Law: "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 this kind of common sense to determine the consequences of an action and virtue ethical agents need this kind of common sense to determine which virtues an action reflects. Making any ethical judgement requires relatively robust conceptions of the action or situation 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. All moral theories evaluating falsely promising will a robust definition of the convention of promising, but consequentialism and virtue ethics will also require additional information about consequences or character that Kantian ethics will not. Thus, although the need for common sense poses

a challenge to automated Kantian ethics, this challenge is more acute for consequentialism or virtue ethics so Kantian ethics remains within the closest reach of automation.

1.2 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 base logic.

1.2.1 Deontic Logic

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 Cresswell and Hughes (1996). These logics usually also contain the possibility operator $\diamond$, where $\diamond p \iff \sim \Box \sim p$. Additionally, modal logics include operators of propositional logic like $\sim, \wedge, \vee, \rightarrow$.

A deontic logic is a special kind of modal logic designed to reason about obligation. Standard deontic logic (Cresswell and Hughes, 1996; McNamara and Van De Putte, 2021) replaces $\Box$ with the obligation operator O, and $\diamond$ with the permissibility operator P. Using the Kripke semantics for O, Op is true at w if p is true at all ideal deontic alternatives to w. 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 [11]. In situations where duty is

[11] 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

violated, the logic breaks down and produces paradoxical results. Thus, I use an improved deontic logic instead of SDL for this work.

1.2.2 Dyadic Deontic Logic

I use as my base logic Carmo and Jones's dyadic deontic logic, or 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". This gracefully handles contrary-to-duty conditionals. The obligation operator uses a neighborhood semantics Scott (1970); MONTAGUE (1970), instead of the Kripke semantics. Carmo and Jones define a function ob that maps from worlds to sets of sets of worlds. Intuitively, each world is mapped to the set of propositions obligated at that world, where a proposition p is defined as the worlds at which the p is true.

DDL also includes other 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.

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 impacts. DDL has a large set of axioms involving quantification over complex higher-order logical expressions. Proofs involving these axioms will

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)

be computationally expensive. Benzmueller and Parent warned me that this may become a problem if Isabelle's automated proof tools begin to time out.

1.3 Isabelle/HOL Implementation

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.

I started my project by reimplementing Benzmueller, Farjami, and Parent's implementation of DDL in Isabelle/HOL Benzmüller et al. (2021); Benzmüller et al. (2019). This helped me learn how to use Isabelle/HOL, and the implementation showcased in the next few sections demonstrates the power of Isabelle.

Benzmueller, Farjami, and Parent use a shallow semantic embedding. This kind of embedding models the semantics of DDL as constants in HOL and axioms as constraints on DDL models. This document will contain a subset of my implementation that is particularly interesting and relevant to understanding the rest of the project. For the complete implementation, see the source code in `paper22.thy`.

1.3.1 System Definition

The first step in embedding a logic in Isabelle is defining the relevant terms and types.

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

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 $\{\text{True}\}$ would be true at any set of worlds.

The main accessibility relation that I will use is the *ob* relation:

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

1.3.2 Axiomatization

For 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, so it suffices to place conditions on the accessibility relations. These axioms can be quite unweildy, so luckily I was able to lift BFP's implementation of Carmo and Jones's original axioms directly (Benzmüller et al., 2021). Here's an example of an axiom:

and *ax-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-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.

1.3.3 Syntax

The syntax that I will work with is defined as abbreviations. 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), abbreviations may hurt performance. In some complicated proofs, we want to control

definition unfolding. Benzmueller and Parent told me that the performance cost of abbreviations can be mitigated using a definition instead.

Modal operators will be useful for my purposes, but the implementation is pretty standard.

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

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

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

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

The most important operator for our purposes is the obligation operator.

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 we need a monadic obligation operator. Below is some syntactic sugar for a monadic obligation operator.

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

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

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

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

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

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

— Intuitively, the context `True` is the widest context possible because `True` holds at all worlds.

Validity will be useful when discussing metalogical/ethical properties.

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

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

1.3.4 Syntactic Properties

One way to show that a semantic embedding is complete is to show that the syntactic specification of the theory (axioms) are valid for this semantics - so to show that every axiom holds at every world. Benzmueller, Farjami, and Parent provide a complete treatment of the completeness of their embedding, but I will include selected axioms that are particularly interesting here. This section also demonstrates many of the relevant features of Isabelle/HOL for my project.

Consistency

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

— Isabelle has built-in support for Nitpick, a model checker. Nitpick successfully found a model satisfying these axioms so the system is consistent.

— Nitpick found a model for card i = 1:

Empty assignment

Nitpick Blanchette and Nipkow (2010) can generate models or countermodels, so it's useful to falsify potential theorems, as well as to show consistency. by simp indicates the proof method. In this case, simp indicates 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.

Modus Ponens

lemma *modus-ponens*: **assumes** $\models A$ **assumes** $\models (A \rightarrow B)$

shows $\models B$

using *assms*(*1*) *assms*(*2*) **by** *blast*

— Because I have not defined a "derivable" operator, inference rules are written using assumptions.

— The rule blast is a classical reasoning method that comes with Isabelle out of the box. Nipkow et al. (2002)

— This is an example of a metalogical proof in this system using the validity operator.

Another relevant operator for our purposes is $\Box$, the modal necessity operator. In this system, $\Box$ behaves as an S5 Cresswell and Hughes (1996) modal necessity operator.

lemma *K*:

shows $\models ((\Box(A \rightarrow B)) \rightarrow ((\Box A) \rightarrow (\Box B)))$ **by** *blast*

lemma *T*:

shows $\models ((\Box A) \rightarrow A)$ **by** *blast*

lemma *5*:

shows $\models ((\Diamond A) \rightarrow (\Box(\Diamond A)))$ **by** *blast*

As mentioned earlier, the obligation operator is most interesting for my purposes. Here are some of its properties.

lemma *O-diamond*:

shows $\models (O\{A|B\} \rightarrow (\Diamond(B \wedge A)))$

using *ax-5b ax-5a*

by *metis*

— A is only obligatory in a context if it can possibly be true in that context. This is meant to prevent impossible obligations.

lemma *O-nec*:

shows $\models (O\{B|A\} \rightarrow (\Box O\{B|A\}))$

by *simp*

— Obligations are necessarily obligated. This axiom is faithful to Kant's interpretation of ethics and is evidence of DDL's power in representing Kant's theory. Kant claimed that

the categorical imperative was not contingent on any facts about the world, but instead a property of the concept of morality itself Kant (1785). Under this view, obligation should not be world-specific.

Below is an example of a more involved proof in Isabelle. This proof was almost completely automatically generated. The property itself here is not very interesting for my purposes because I will rarely mix the dyadic and monadic obligation operators.

lemma *O-to-O*:

shows $\models(O\{B|A\}\rightarrow O\{(A\rightarrow B)|\top\})$

proof−

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

— I had to manually specify this subgoal, but once I did Isabelle was able to prove it automatically.

by (*smt ax-5d ax-5b ax-5b''*)

— Isabelle's proof-finding tool, Sledgehammer Paulson and Blanchette (2015), comes with out-of-the-box support for smt solving Blanchette et al. (2011).

thus *?thesis*

proof −

have *f1*: $\forall p\ pa\ pb.\ ((\neg\ (ob\ p\ pa)) \vee (\exists i.\ (p \wedge (\neg\ pb))\ i)) \vee (ob\ pb\ ((pb \wedge (\neg\ p)) \vee\ pa))$

using ‹$\forall X\ Y\ Z.\ ob\ X\ Y \wedge (\models(X\rightarrow Z)) \longrightarrow ob\ Z\ (\ (Z \wedge (\neg\ X)) \vee\ Y)$› **by** *force*

obtain *ii* :: $(i \Rightarrow bool) \Rightarrow (i \Rightarrow bool) \Rightarrow i$ **where**

$\forall x0\ x2.\ (\exists v3.\ (x2 \wedge (\neg\ x0))\ v3) = (x2 \wedge (\neg\ x0))\ (ii\ x0\ x2)$

by *moura*

then have $\forall p\ pa\ pb.\ ((\neg\ ob\ p\ pa) \vee (p \wedge (\neg\ pb))\ (ii\ pb\ p)) \vee ob\ pb\ (\ (pb \wedge (\neg\ p)) \vee\ pa)$

using *f1* **by** *presburger*

then show *?thesis*

by *fastforce*

qed

— This entire Isar style proof was automatically generated using Sledgehammer.

qed

The implementation of DDL showcases some of the useful features of Isabelle. Abbreviations allow us to embed the syntax of DDL into HOL without defining an entire abstract sytax tree. Automated support for proof-finding using Sledgehammer makes proving lemmas trivial, and proving more complex theorems far easier. Nitpick's model finding ability is useful to check for consistency and create countermodels.

2 Prior Formalizations of The Categorical Imperative

In this section, I present two formalizations of the categorical imperative and a testing framework to evaluate them. In Section 2.1, I will consider an intuitive but naive formalization of the formula of universal law. This formalization is equivalent to a theorem in my base logic (DDL), so thus does not actually increase the power of my base logic. In effect, this formalization serves as a control group that I use to present the testing architecture used to evaluate following formalizations. In Section 2.2, I will explore Moshe Kroy's partial formalization of the categorical imperative.

2.1 Naive Formalization of the Formula of Universal Law

This section presents a simple and intuitive formalization of the Formula of Universal Law (FUL). This naive formalization will hold in the base logic itself, so this formalization does not actually improve upon an ordinary deontic logic at all. This section serves two purposes. First, the naive formalization is a toy example that demonstrates the implementation and testing process that will be used for the more complex formalizations presented later in Chapters 2 and 3. Second, this formalization is effectively a control group used to determine which properties of

obligaition hold in the base logic. Future formalizations will improve on the base logic by passing more tests, or equivalently, proving more properties of obligation than the base logic can.

The FUL roughly states that, if a maxim cannot be willed in a world where it is universalized, it is prohibited. One reading of this rule is that a maxim is only permissible if it is necessarily permissible. To formalize a reading of the FUL like this naive one, I will first represent the reading as a sentence in my logic and then add this sentence as an axiom in my logic.

2.1.1 Formalization

Many of the formalizations of the categorical imperative that I present in this thesis require some logical background. This naive formalization requires that I define the notion of permissibility, where an action is permissible if and only if it is not prohibited.

abbreviation *ddlpermissable*::$t \Rightarrow t$ (P-)

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

— An act A is permissible if its negation is not obligated. For example, buying a red folder is permissible because I am not required to refrain from buying a red folder.

This naive formalization requires no additional logical machinery, but more complex formalizations may require additional logical concepts.

Let's now consider a naive reading of the Formula of Universal Law (FUL): "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant, 1785). An immediate translation to DDL is that if A is not necessary permissible then it is prohibited. In other words, if we cannot universalize PA (where universalizing is represented by the modal necessity operator), then A is prohibited. This sentence is formalized in the abbreviation below:

abbreviation *FUL-naive* **where** $\mathit{FUL\text{-}naive} \equiv \lambda A.\ ((\neg(\Box\ (P\ A))) \rightarrow (O\ \{(\neg A)\}))$

— For a given maxim 'A', the FUL states that if A is not necessarily permissible, it is prohibited.

This naive formalization holds as a theorem of DDL. I show this using Isabelle below:

lemma $\forall A.\ \models (\mathit{FUL\text{-}naive}\ A)$

by *simp*

— In this short and simple proof, the statement "by simp" demonstrates that the proof is completed using the "simp" tool, which is Isabelle's term rewriting engine. In this case, the result follows from the definitions of the modal operators in DDL, so term rewriting suffices to complete the proof.

The general process of implementing a formalization of the FUL will be to represent the formalization as a sentence in my logic, as above, and then to add the formalization as an axiom to the logic. Kant's ethical theory is rule based, so it involves applying the categorical imperative to solve ethical dilemmas. In logic, this is equivalent to adopting the categorical imperative as an axiom and then reasoning in the newly formed logic to come to ethical conclusions. Adding the categorical imperative as an axiom makes it impossible to violate it and thus represents the categorical imperative as the supreme, unviolable law of morality.

Note that under this approach, reasoning about violations of obligation is difficult. Any violation of the categorical imperative immediately results in a contradiction. Developing a Kantian account of contrary- to-duty obligations is a much larger philosophical project that is still open Korsgaard (1986). This paper will focus on the classical Kantian notion of an ideal moral world, and thus does not reason about violations of the moral law O'Neill (2009).

Because my naive formalization holds in the base logic, adding it as an axiom does not make the logic any more powerful. No new theorems can be derived using the

naive formalization that could not already be derived in the base logic. Thus, this section serves as a "control group." Tests performed in this section establish which properties of obligation don't hold in the base logic. The fact that these tests will pass for the later, more sophisticated formalizations will serve as evidence for the superiority of these formalizations over the base logic.

axiomatization where

FUL-1: $\models ((\neg(\Box\,(P\,A))) \rightarrow (O\,\{(\neg A)\}))$

Once I add a formalization of the FUL as an axiom to my system, I will test the formalization. Each test will take the form of a lemma which I expect to either hold or be disproven by the categorical imperative. For example, one test might be the lemma "murder is wrong." I will evaluate formalizations based on their ability to prove expected properties of the categorical imperative, as determined by philosophical literature. These tests fall into two categories: metaethical tests, which focus on abstract properties of the ethical system, and application tests, which simulate the kind of practical reasoning that an agent would actually perform by specifying a simple model.

One way to understand computational ethics is as translational work that seeks to translate an ethical theory presented by a philosopher to something that a computer can reason about. My testing architecture evaluates how faithful a particular formalization is to the ethical theory that it translate. This testing approach is not specific to my ethical theory and could be used to evaluate other formalizations of other theories as well.

2.1.2 Metaethical Tests

First, I present metaethical tests for the naive formalization (or equivalently the base logic). These tests evaluate abstract properies of the system, independent of a particular agent, situation, or act. For example, one metaethical test may be

to determine if the system is capable of generating models in which actions are obligated. If the system can never obligate anything, this indicates that it is not a good ethical system.

Preliminary Tests

The immediate test for any formalization is consistency, or the property of being free of contradictions. An inconsistent formalization is immediately useless, because all sentences are true in an inconsistent logic. Nitpick, Isabelle's model checker, offers a handy way of checking consistency. Specifically, if Nitpick can find a model that satisfies all the axioms of the logic, then the logic is consistent.

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

— Nitpick found a model for card i = 1:

Empty assignment

— Nitpick tells us that the FUL is consistent[12]

An initial property that we might be interested in is the possibility of permissibility, or whether or not the system can generate models in which certain acts are permissible. In modern ethics, permissibility is a well-accepted phenomenon. An ethical theory that doesn't allow for permissibility would require that every action is either obligatory or prohibited. If that is the case, many counterintuitive theorems follow, including that all permissible actions are obligatory.[13] Therefore, I will include the possibility of permissibility as one test for my formalizations.

lemma *permissible*:

shows $\exists A.\ ((\neg\ (O\ \{A\}))\ \wedge\ (\neg\ (O\ \{\neg\ A\})))\ w$

[12] "oops" at the end of a lemma indicates that the proof is left unfinished. It does not indicate that an error occurred. In this case, we aren't interested in proving True (the proof is trivial and automatic), hence the oops.

[13] Proof is in the appendix.

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

— Nitpick found a model for card i = 1:

Skolem constant: A =($\lambda x\_.\_$)(i_1 := False)

— I want to show that there exists a model where there is some formula A that is permissible, or, in English, that permissibility is possible. Nitpick finds a model where the above formula holds, so permissibility is indeed possible.

— Quick note on how to read Nitpick results. Nitpick is Isabelle's model checker, and it can either time out, find a model that satisfies the given theorem, or find a counterexample that disproves the given theorem. It will then provide the corresponsing model by specifying model components. For readability, all terms except for the free variables are hidden. This model has cardinality 1 for the world (i) type. The term 'A' is defined as false at world i_1.

— These details will be elided for most Nitpick examples, but this provides guidance on how to interpret the output.

Another similar property is that for any arbitrary action A, there is a model that makes it permissible. This property is actually not desirable, because if A is "murder" then the CI should require that it be prohibited in every world. Therefore, in order for this test to pass, Nitpick should *not* be able to find a satisfying model for this formula.

lemma *fixed-formula-is-permissible*:

fixes *A*

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 := False)

— Because Nitpick finds a satisfying model for this formula, this test fails for the naive interpretation.

Another initial property is that arbitary actions should not be obligated. No sensible ethical theory would require that any arbitrary action A is obligated, because A may

be something obviously wrong, like murder. In order for this test to pass, Nitpick must disprove the formula below by finding a counterexample.

lemma *arbitrary-obligations*:

fixes A::t

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

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

— Nitpick found a counterexample for card i = 1:

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

— Nitpick finds a counterexample disproving the statement that any arbitrary action is obligatory, so this test passes.

Conflicting Obligations

The next set of tests will focus on conflicting obligations. There is some debate about Kant's personal stance on conflicting obligations, but neo-Kantians agree that the FUL itself cannot obligate conflicting actions. For more complete discussion of conflicting obligations in Kantian literature, see Section 2.3.1. I will first test whether or not, for some arbitrary action, Nitpick can find a model in which that action is both obligated and prohibited.

lemma *conflicting-obligations*:

fixes A

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

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

— Nitpick found a model for card i = 2:

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

— Nitpick found a model with conflicting obligations, so this tests fails.

The above is a rather weak notion of contradictory obligations. Korsgaard additionally argues that Kantian ethics also has the stronger property that if two maxims imply a contradiction, they must not be willed (Korsgaard, 1985). I test this

property below. Because this property is stronger than the previous test, and the previous test failed, this test will also fail.

lemma *implied-contradiction*:

fixes A::t

fixes B::t

assumes $\models(\neg\ (A \wedge B))$

shows $\models(\neg\ (O\ \{A\} \wedge O\ \{B\}))$

nitpick [*user-axioms*]

proof −

have $\models(\neg(\Diamond(A \wedge B)))$

by (*simp add*: *assms*)

then have $\models(\neg\ (O\ \{A \wedge B\}))$ **by** (*smt O-diamond*)

— Notice that this is **almost** the property we are interested in. In fact, if $O\{A \wedge B\}$ is equivalent to $O\{A\} \wedge O\{B\}$, then the proof is complete.

thus *?thesis* **oops**

— Nitpick found a counterexample for card i = 2:

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

— Sadly the property I'm actually interested in doesn't follow.

The above proof yields an interesting observation. $O\{A \wedge B\}$ is not equivalent to $O\{A\} \wedge O\{B\}$. 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, so they should mean the same thing in our logic as well. This "distributive" property of obligation is another test.

lemma *distributive-property-for-obligation*:

shows $\models (O\ \{A\} \wedge O\ \{B\}) \equiv \models O\ \{A \wedge B\}$

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

— Nitpick found a counterexample for card i = 2:

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

again, this tests fails in the control group.

Miscellaneous Properties

The last set of metaethical tests involve miscellaneous properties of the categorical imperative. First, I show that the naive formalization is equivalent to the below theorem, which clearly fails to track intuition about ethics.

lemma *FUL-alternate*:

shows $\models ((\Diamond\ (O\ \{\neg\ A\})) \rightarrow (O\ \{\neg\ A\}))$

by *simp*

— This means that if something is possibly prohibited, it is in fact prohibited.

This is a direct consequence[14] of the naive formalization, but it's not clear to me that this is actually how we think about ethics. For example, imagine an alternate universe where smiling at someone is considered an incredibly rude and disrespectful gesture. In this universe, I am probably prohibited from smiling at people, but this doesn't mean that in this current universe, smiling is morally wrong.

The "ought implies can" principle is attributed to Kant[15] and is rather intuitive: you can't be obligated to do the impossible. Deontic logics evolved specifically from this principle, so this should hold in the base logic (Cresswell and Hughes, 1996).

lemma *ought-implies-can*:

shows $\forall A.\ \models (O\ \{A\} \rightarrow (\Diamond A))$

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

This test passes in the base logic, and will thus hold in all future formalizations as well. Therefore, it's an interesting property but not actually useful in evaluating different formalizations of the FUL.

[14] For a manual proof, see the Appendix.

[15] The exact philosophical credence of this view is disputed, but the rough idea holds nonetheless. See Kohl (2015) for more.

2.1.3 Application Tests

The second category of tests I will consider is Application tests, which involve specifying models to encode certain facts into the system, and then asking questions about obligations. Metaethical tests focus on properties that apply to all acts, circumstances, and actors, but application tests focus on specific acts. Let's start with analyzing an obvious example - that murder is wrong.

First, I will define murder as a constant below. Notice that right now, this constant is just a term. I haven't specified any properties of murder, so as of now, it's interchangeable with any other term. Application tests generally define an act and then define properties of the act (e.g. if X is murdered, X dies). The tests aim to show that acts with certain properties are either obligated or prohibited.

consts M::t

abbreviation *murder-wrong*::*bool* **where** *murder-wrong* $\equiv \models(O\ \{\neg\ M\})$

— This abbreviation merely represents the statement that murder is prohibited.

I will now define properties of murder and see if they achieve the desired result that murder is prohibited. First, I start with the rather basic property that murder is prohibited in some world, or that murder is possibly wrong. This is quite a strong assumption because it gives the system a moral fact about a kind of prohibition against murder. Ideally, an ethical theory can take nonmoral facts about murder (like murder kills) and use these to generate a moral judgement about the wrongness of murder. This property is much stronger than the assumptions that we make in ordinary moral reasoning and thus should be more than enough to show that murder is wrong.

abbreviation *possibly-murder-wrong*::*bool* **where** *possibly-murder-wrong* $\equiv (\Diamond\ (O\ \{\neg\ M\}))\ cw$

lemma *wrong-if-possibly-wrong*:

shows *possibly-murder-wrong* $\longrightarrow$ *murder-wrong*

by *simp*

— This lemma gets to the "heart" of this naive interpretation. If something isn't necessarily obligated, it's not obligated anywhere.

The above example does exactly what I expect it to: it shows that if something is wrong somewhere it's wrong everywhere. That being said, it seems like quite a weak claim. I assumed a very strong, moral fact about murder (that it is wrong somewhere), so it's not surprise that I was able to show the wrongness of murder.

Let's try a different example using a much weaker, nonmoral assumption. Kant argues that the FUL prohibits lying.[16] In this example, I will define lying as a term such that not everyone can lie simultaneously. This is one of Kant's canonical examples of the universalizability test. Lying fails the universalizability test because, in a world where everyone lied, no one would believe each other anymore, so the very system or truth-telling would break down, making lying impossible. I can represent this reasoning in my logic as the assumption that not everyone can lie simultaneously.

To fully capture this idea, I need some notion of a person, so that I can argue that not all people can lie simultaneously.

typedecl *person*

consts *lie*::*person*$\Rightarrow$*t*

consts *me*::*person*

Again, this machinery is quite empty because it doesn't specify any axioms about what a person can or cannot do. In future formalizations, I will define a more robust notion of a person, but the naive formalization has no conception of a person.

abbreviation *lying-not-universal*::*bool* **where** *lying-not-universal* $\equiv \forall w.\ \neg\ ((\forall x.\ lie(x)$

[16] Specifically, he prohibits making a false promise in order to get some cash (Kant, 1785, idk page no).

$w) \wedge (lie(me)\ w))$

This is a rough translation of failure of the universalizability test: I test the maxim universally, as represented by the universal quantifier in the first conjunct, and simultaneously, as represented by the second conjunct (Kleingeld, 2017). The FUL says that if this sentence is true, then lying should be prohibited. Therefore, the above sentence should imply that lying is prohibited.

lemma *breaking-promises*:

assumes *lying-not-universal*

shows $(O\ \{\neg\ (lie(me))\})\ cw$

nitpick [*user-axioms*]

oops

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

Empty assignment

This test fails. The FUL should say that lying is prohibited and the fact that it doesn't demonstrates the weakness of this naive formulation of the categorical imperative. Kant's version of the FUL universalizes across people, as in the definition of *lying-not-universal* $\equiv \forall\, w.\ \neg\ ((\forall\, x.\ lie\ x\ w) \wedge lie\ me\ w)$. When universalizing an act, Kant imagines a world in which all *people* perform the act. The naive formalization, on the other hand, universalizes an act across *worlds* because it uses the $\Box$ operator to represent universalization. This is the philosophical error that makes the naive formalization so naive, and future formalizations will need to remedy this error. This serves as an example of the kind of reasoning that Isabelle empowers us to do. Even this simple argument has philosophical consequences. It tells us that reading the FUL as a claim about consistency across possible worlds, instead of consistency across agents, leads to counterintuitive conclusions.

Additionally, Kant argued that obligations are not person-specific but instead apply

equally to all rational agents.[17] Thus, any formalization of the categorical should generate obligations that are consistent across people. This next step analyzes this property.

lemma *equal-obligations*:

assumes $\models O\ \{(lie(me))\}$

shows $\forall x.\ \models (O\ \{(lie(x))\})$

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

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

Free variable: lie = $(\lambda x.\_)(p_1 := (\lambda x.\_)(i_1 := \text{False}, i_2 := \text{True}), p_2 := (\lambda x.\_)(i_1 := \text{False}, i_2 := \text{False}))$ Skolem constant: x = p_2

In this section, I presented the framework I will use to implement and test different interpretations of the categorical imperative. An implementation consists of some necessary logical background, a representation of the FUL using that logical background, and a logical system that adds that representation as an axiom. To test such an implementation, I design a "test suite" that consists of properties of the categorical imperative verified by philosophical literature. I demonstrated the performance of these tests in my base logic, which serves as a control group.

I will evaluate more sophisticated formalizations of the FUL using this testing framework. The properties I test will remain more or less consisten across different formalizations, but the exact logical representation of the tests will depend on the specifics of a particular implementation. In the next section, I will implement Moshe Kroy's formalization of the FUL and evaluate it using this testing framework. Finally, I will use the results of these tests to define clear goals for a custom formalization of the categorical imperative. These goals represent areas of improvement over previous formalizations, and I will justify them using philosophical literature.

[17] For a philosophical analysis of this idea, see Section 2.3.1

2.2 Kroy's Formalization of the Categorical Imperative

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.

2.2.1 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[18]. $P(d/e)$ 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

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

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.

abbreviation *os-neg*::$os \Rightarrow os$ ($\neg$-)
  where $(\neg A) \equiv \lambda x.\ \neg(A(x))$
abbreviation *os-and*::$os \Rightarrow os \Rightarrow os$ (-$\wedge$-)
  where $(A \wedge B) \equiv \lambda x.\ ((A(x)) \wedge (B(x)))$
abbreviation *os-or*::$os \Rightarrow os \Rightarrow os$ (-$\vee$-)
  where $(A \vee B) \equiv \lambda x.\ ((A(x)) \vee (B(x)))$
abbreviation *os-ob*::$os \Rightarrow os$ (O{-})
  where $\mathrm{O}\{A\} \equiv \lambda x.\ (O\ \{A(x)\})$

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

abbreviation *ddl-permissible*::$t \Rightarrow t$ (P {-})
  where $P\ \{A\} \equiv \neg\ (O\ \{\neg\ A\})$
abbreviation *os-permissible*::$os \Rightarrow os$ (P {-})
  where 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 = $(\lambda x.\_)(i_1 := \text{False})$

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

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

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 [19]. 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[20]. 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**

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

[20] 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.

— Nitpick found a counterexample for card i = 2:

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

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

2.2.2 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. 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 = $(\lambda x\_.)(s_1 := (\lambda x\_.)(i_1 := \text{True}, i_2 := \text{True}), s_2 := (\lambda x\_.)(i_1 := \text{False}, i_2 := \text{False}))$ 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!

2.2.3 Application Tests

In the following sections, I will use the application and metaethical tests presenting in Sections 2.1.3 and 2.1.2 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 2.2.1, 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 2.1.3, 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 2.1.3—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* $\langle \exists p. \models \mathrm{O}\{\neg M\text{-}kroy\}\ p \rangle$ **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[21].

Lying

[21] 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.

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-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

$p = s_1$

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 = $(\lambda x.\_)(a_1 := (\lambda x.\_)(i_1 :=$ 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[22] that this lemma holds in his logic.

[22] See footnote 19 on p. 199

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 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.

2.2.4 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 = (λx._)(i_1 := 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 = (λx._)(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 = (λx._)(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 2.1.2 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 = $(\lambda x\_.)(i_1 := \text{False}, i_2 := \text{True})$

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

2.2.5 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** $FUL\text{-}strong \equiv \forall w\ A.\ ((\exists p{::}s.\ ((P\ \{A\ p\})\ w))$ $\longrightarrow(\ ((\ P\ \{\ \lambda x.\ \forall p.\ A\ p\ x\})\ w)))$

lemma *strong-implies-weak*:
shows $FUL\text{-}Strong \longrightarrow FUL$
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 $FUL\text{-}strong \longrightarrow \models(\neg\ P\ \{lie(p)\})$

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* ≡ *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[23] 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* ⟶ *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* ⟶ *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 what to do about this.

2.3 Lessons Learned and Goals for Chapter 3

In this chapter, I tested two prior attempts to formalize the Formula of Universal Law and found that these attempts didn't faithfully interpret the FUL. Specifically, certain properties that we expect to hold of the FUL didn't hold in these implemen-

[23] 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.

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 indicating which goals are met by the naive formalization, Kroy's formalization, and the custom formalization respectively.

tations. In an attempt to remedy these shortcomings, in the next chapter, I present my own custom implementation of the FUL that satisfies these properties and is thus more faithful to the literature. Before presenting my implementation, I define some goals for my custom formalization based on learnings from prior attempts and from philosophical literature on the FUL.

The results presented below are summarized in Figure 1. For each goal, I indicate which interpretations successfully meet that goal. The fact that my custom formalization meets all the goals indicates that it improves on prior formalization attempts. Below I will explain and justify the goals.

2.3.1 Goals From Prior Attempts

FUL Stronger than DDL One simple objective that the naive formalization failed to meet is the fact that the FUL should not hold in the base logic (DDL). Recall that the naive formalization of the FUL[24] held in the base logic, so adding it as an axiom didn't make the logic any stronger. This is troubling because the base logic does not come equipped with the categorical imperative built-in. It defines basic properties of obligation, such as ought implies can, but contains no axioms that

[24]This formalization reads $\vDash ((\neg(\Box P\{A\})) \longrightarrow O\{\neg A\})$.

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. The naive interpretation holds in DDL but Kroy's formalization does not. Because the naive interpretation is no stronger than DDL, it is acts a control group equivalent to DDL itself.

Obligation Universalizes Across People Another property of the Formula of Universal Law that any implementation should satisfy is that obligation generalizes across people. In other words, if a maxim is obligated for one person, 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 equually (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 latches onto this property and makes it the content of his formalization, which essentially says that if an act is permissible for someone, it is permissible for everyone.[25] While Kroy's interpretation clearly satisfies this property, the naive interpretation does not.

Contradictory Obligations Another problem with prior formalizations was that they didn't prohibit contradictory obligations, partially because DDL itself allows contradictory obligations. Kant subscribes to the general, popular view that morality is supposed to guide action, so ought implies can.[26] Kohl reconstructs his argument for the 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).

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

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

This defeats the purpose of Kant's theory—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). [27]. Both the naive formalization and Kroy's formalization allow contradictory obligations.

During testing, I also realized that contradictory obligations are closely related to two other properties that also fail in both of these systems. 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. Moreover, in a system with contradictory obligations, this property fails because there is some A that is obligated but also prohibited and therefore not permisible. Indeed, formalizing this property below shows that this follows from the definition of implication in propositional logic.

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

by *simp*

Distributive Property Another property related to contradictory obligations is the

[27]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.

distributive property for the obligation operator.[28] 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, so they should mean the same thing in logic as well. Moreover, if that (rather intuitive) property holds, then contradictory obligations are impossible, as shown in the below proof.

lemma *distributive-implies-no-contradictions*:
assumes $\forall A\ B.\ \models ((O\ \{A\} \wedge O\ \{B\}) \equiv O\ \{A \wedge B\})$
shows $\forall A.\ \models(\ \neg(O\ \{A\} \wedge O\ \{\neg\ A\}))$
using *O-diamond assms* **by** *blast*

Thus, while testing contradictory obligations, I also test the distributive property for the obligation operator. Again, this property fails in the naive formalization and for Kroy's formalization.

Un-universalizable Actions This goal is inspired by a test performed for Kroy's formalization. Under a naive reading of the Formula of Universal Law, it 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. More precisely, 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 the naive formalization and Kroy's formalization and is a goal for my custom formalization.

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

2.3.2 Goals From Philosophical Literature

The goals above come from moral intuition, properties of Kantian ethics, and logical requirements. In order to stay faithful to centuries of philosophical debate about the meaning of the Formula of Universl Law, I also present some goals inpsired by this literature.

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 neither of the prior attempts do.

Conventional Acts Kantians debate over the most philosophically sound interpretation of the Formula of Universal Law. One litmus test that Korsgaard introduces makes a distinction between conventional and natural acts (Korsgaard, 1985). A conventional act is one like promising, which relies on the convention of promising, in which we all implicitly understand a promise as a commitment. Conventional acts are generally easier to show the wrongness of 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 anymore, so falsely promising is prohibited. Despite the relative ease of this property, it fails in both the naive and Kroy's interpretations, demonstrating the weakness of these formalizations. This property will hold for my custom formalization.

Natural Acts The more difficult kind of act to show the wrongness of is 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. This property of natural acts makes it difficult for interpretations of the FUL to show the

wrongness of violence. Both the naive and Kroy's interpretations fail to show the wrongness of natural acts (in fact, they fail to show the weaker wrongness of conventional acts). This property will hold for the custom formalization. I will show the wrongness of both natural and conventional acts by formalizing Korsgaard's practical contradiction interpretation of the FUL, which is widely accepted as the canonical interpretation of the FUL (Korsgaard, 1985). I will explain this decision in greater detail in the next chapter, where I present my custom formalization.

3 Novel Formalization of the Categorical Imperative

In this section, I present a custom formalization of the categorical imperative, as inspired by the goals from the previous chapter.

3.1 Logical Background

The previous attempts to model the categorical imperative in Chapter 2 partially failed due to an inability to fully represent the complexity of a maxim. Specifically, they treated actions as a single, monolithic unit of evaluation, whereas most Kantians consider the unit of evaluation for the FUL to the more complex notion of a maxim. In this section, I will present some logical background necessarily to fully capture the spirit of a maxim. I will begin by borrowing some machinery to handle "subjects" who perform actions from Chapter 2.

typedecl s — s is the type for a "subject," i.e. the subject of a sentence. In this interpretation, a subject is merely defined as "that which can act." It does not include any other properties, such as rationality or dignity. As I will show, for the purposes of defining the universalizability test, this "thin" representation of a subject suffices.

type-synonym $os = (s \Rightarrow t)$ — Recall that an open sentence maps a subject to a term to model the substitution operator.

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

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

I define a maxim as a circumstance, act, goal tuple (C, A, G), read as "In circumstances C, act A for goal G." Isabelle's strict typing rules mean that the choice of the type of each member of this tuple is significant. A circumstance is represented as a set of worlds t where that circumstance holds. A goal is also a term because it can be true or false at a world if it is realized or not. An act is an open sentence because an act itself is not the kind of thing that can be true or false (as in, an act is not truth-apt), but the combination of a subject performing an act can be true or false at a world depending on whether or not the act is indeed performed by that subject. For example, "running" is not truth-apt, but "Sara runs" is truth-apt.

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

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

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

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

O'Neill's inclusion of circumstances is potentially controversial because it leaves open the question of what qualifies as a relevant circumstance for a particular maxim. This is gives rise to "the tailoring objection" (Kitcher, 2003, 217) [29], under which maxims are arbitrarily specified to pass the FUL. For example, the maxim "When my name is Lavanya Singh, I will lie to get some easy money," is universalizable, but is clearly a false positive. One solution to this problem is to argue that the circumstance "When my name is Lavanya Singh" is not morally relevant to the act and goal. This solution requires some discussion of what qualifies as a relevant circumstance.

O'Neill seems to acknowledge the difficulty of determining relevant circumstances when she concedes that a maxim cannot include all of the infinitely many circumstances in which the agent may perform the action(O'Neill, 2013, 4:428). She argues that this is an artifact of the fact that maxims are rules of practical reason, the kind of reason that helps us decide what to do and how to do it Bok (1998). Like any practical rule, maxims require the exercise of practical judgement to de-

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

termine in which circumstances they should be applied. This judgement, applied in both choosing when to exercise the maxim and in the formulation of the maxim itself, is what determines the "morally relevant circumstances."

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

While this may appear to be a weakness of my system, I believe that it actually allows my system to retain some of the human complexity that many philosophers agree cannot be automated away.[30] Ethics is a fundamentally human activity. Kant argues that the categorical imperative is a statement about the properties of rational wills. In fact, Korsgaard argues that morality derives its authority over us, or nor-

[30] Powers presents the determination of morally relevant circumstances as an obstacle to the automation of Kantian ethics Powers (2006).

mativity, only because is it a property of a rational will, and we, as human beings, are rational wills. If ethics is meant to guide human behavior, the role of the computer becomes clear as not a replacement for our will, but instead as a tool to help guide our wills and reason more efficiently and more effectively. Just as calculators don't render mathematicians obsolete, computational ethics does not render human judgement or philosophy obsolete. Chapter 4 Section ?? will be devoted to a more complete discussion of this issue.

Exclusion of Motive

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

While Kitcher's conception of a maxim captures Kant's idea of acting for duty's own sake, I will not implement it because it is not necessary for putting maxims through the FUL. Indeed, Kitcher acknowledges that O'Neill's formulation 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.

abbreviation *will* :: $maxim \Rightarrow s \Rightarrow\ t\ (W\ \text{-}\ \text{-})$

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

Korsgaard claims that "to will an end, rather than just wishing for it or wanting it, is to set yourself to be its cause" (Korsgaard and O'Neill, 1996, 38). To will a maxim is to set yourself to be the cause of its goal by taking the means specified in the maxim in the relevant circumstances. This coheres with Kitcher's and Korsgaard's understanding of a maxim as a principle or rule to live by.

At worlds where the circumstances do not hold, a maxim is vacuously willed. If you decide to act on the rule "I will do X in these cirumstances", then you are vacuously obeying it when the circumstances don't hold.

The above discussion implies that willing a maxim is particular to the agent, justifying my choice to require that a particular subject will a maxim. O'Neill argues for this interpretation when she distinguishes between the evaluation of a principle, which is generic, and a maxim, which she views as "individuated only by referring to a person"(O'Neill, 2013, 13). I adopt the spirit of this interpretation but modify it slightly by representing the general maxim as a principle that anyone could adopt, and the act of willing the maxim as a person-particular instantiation of the maxim.

I additionally represent a subject as willing a maxim because I use the word 'will' as a verb, to mean committing oneself to living by the principle of a maxim. This coheres with the FUL, which tests the act willing of a maxim by determining if the maxim could be a universal law that everyone committed to. Formalizing this idea, the type of a willed maxim is a term, allowing me to use DDL's obligation operator on the notion of willing a maxim. Concretely, my system will prove or disprove statements of the form "Lavanya is obligated to will the maxim M."

Worlds where the circumstances do not hold are not relevant for determining obligation. Recall that in Benzmueller et. al's definition of the obligation operator, $O\{B|A\}$ is true at all worlds iff ob(B)(A), or if the obligation function maps A to obligatory in context B (where the context is a set of worlds) Benzmüller et al. (2021). This definition implies that worlds outside of B have no bearing on the moral status of A in context B, which coheres with intuitions about contextual obligation. Thus, the dyadic obligation operator disqualifies worlds where the context does not hold, so the vacuous truth of the will statement in these worlds does not matter.

Given that the will abbreviation already excludes worlds where the circumstances fail (by rendering the statement vacuously true at them), one may conclude that the dyadic obligation operator is now useless. Using the dyadic obligation operator allows me to take advantage of the power of DDL to represent the bearing that circumstances have on obligation. DDL has powerful axioms expressing the relationship between circumstances and obligation, such as the fact that obligations are monotonically increasing with respect to broader circumstances. Using the monadic obligation operator would require me to either operate with an empty notion of context or to redefine these axioms. The dyadic obligation operator lets me take advantage of the full power of DDL in expressing contrary-to-duty obligations. This is particularly important for Kantian ethics and the FUL specifically

because many critiques of the FUL rely on attention to circumstances (tailoring objection) or lack thereof (ideal theory). This is also an innovation that my custom formalization presents over the prior work. By formally including the notion of a circumstance or context, I am able to represent these objections that Kantian scholars study. Formalizing Kantian ethics in a dyadic deontic logic instead of a monadic deontic logic is a key contribution of this thesis.

abbreviation *effective* :: $maxim \Rightarrow s \Rightarrow t\ (E\ -\ -)$

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

print-theorems

A maxim is effective for a subject when, if the subject wills it then the goal is achieved, and when the subject does not act on it, the goal is not achieved[31] Menzies and Beebee (2020). The former direction of the implication is intuitive: if the act results in the goal, it was effective in causing the goal. This represents 'necessary' causality.

The latter direction represents 'sufficient' causality, or the idea that, counterfactually, if the maxim were not willed, then the goal is not achieved Lewis (1973a). 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 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.[32]

If the circumstances do not hold and the goal is achieved, then the maxim is vac-

[31] Thank you to Jeremy D. Zucker for helping me think through this.

[32] Should I wave a hand at critiques of counterfactual causality?

uously effective, since it is vacuously willed (as described above). While this scenario is counterintuitive, it is not very interesting for my purposes because, when the circumstances do not hold, a maxim is not applicable. It doesn't really make sense to evaluate a maxim when it's not supposed to be applied. The maxim "When on Jupiter, read a book to one-up your nemesis" is vacuously effective because it can never be disproven.

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

abbreviation *holds*::*maxim*⇒*t* **where**
$holds \equiv \lambda(c, a, g).\ c$

abbreviation *not-universalizable* :: *maxim*⇒*s*⇒*bool* **where**
$not\text{-}universalizable \equiv \lambda M\ s.\ \forall w.\ ((universalized\ M)\ \rightarrow\ (\neg\ (E\ M\ s)))\ w$
— The formula above reads "at world w, if M is universalized and M is acted on (i.e. the circumstances of M hold), then M is not effective."
Notice that the antecedent specifies that the circumstances hold at the given world. When evaluating if a maxim is universalizable or not, we want to ignore worlds where the circumstance do not hold. At these worlds, the maxim is trivially effective and thus trivially universalizable. If we didn't exclude such worlds from consideration, a maxim with circumstances that ever fail to hold would be universalizable. Clearly this is not a desirable conclusion, since maxims like "When you need money, lie to get easy money" would be universalizable.

As before, the concepts of prohibition and permissibility will be helpful here. The unit of evaluation for my formalization of the FUL is the act of willing a maxim, which entails performing the maxim's act in the relevant circumstances. Therefore, I will say that, just as the act of willing a maxim can be obligatory for a subject, it can be prohibited or permissible for a subject.[33]

[33]In the rest of this section, for convenience, I will use the phrase "subject s willing maxim M

abbreviation *prohibited*::*maxim*⇒*s*⇒*t* **where**

prohibited $\equiv \lambda(c, a, g)\ s.\ O\{\neg\ (will\ (c,a, g)\ s) \mid c\}$

abbreviation *permissible*::*maxim*⇒*s*⇒*t*

where *permissible* $\equiv \lambda M\ s.\ \neg\ (prohibited\ M\ s)$

— I will say that a maxim is permissible for a subject if it is not prohibited for that subject to will that maxim.

When analyzing the naive formalization and Kroy's formalization, I learned that DDL and the prior formalizations allow contradictory obligations. This is a major weakness of these systems, and my formalization should fix this. To do so, I will add as an axiom the idea that obligations cannot contradict each other or their internal circumstances. Formally, conflicting obligations are defined below.

abbreviation *non-contradictory* **where**

non-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-contradictions*:$\forall A::t.\ \forall B::t.\ \forall c::t.\ \forall w::i.$ *non-contradictory A B c w*

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

3.2 Formalizing the FUL

Below is my first attempt at formalizing Korgsaard's definition of the practical contradiction interpretation: a maxim is not universalizable if, in the world where the maxim becomes the standard practice (i.e. everyone acts on the maxim), the

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 I will show to be equivalent to M being obligatory for a specific subject.

agent's attempt to use the maxim's act to achieve the maxim's goal is frustrated. In other words, if the maxim is universally willed (captured by applying a universal qunatifier and the will function to the maxim on the LHS), then it is no longer effective for the subject s (RHS above).

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

— 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 (obligated not to) from willing the maxim.

lemma *FUL0* $\longrightarrow$ *False* **using** *O-diamond*

using *case-prod-conv no-contradictions old.prod.case old.prod.case* **by** *fastforce*

FUL0 is not consistent, and sledgehammer is able to prove this by showing that it implies a contradiction usig axiom O_diamond, which is $\models\lambda w.$ *ob ?B ?A* $\longrightarrow \neg \models\neg$ *?B*$\wedge$*?A*. This axiom captures the idea that an obligation can't contradict its context. This is particularly problematic if the goal or action of a maxim are equivalent to its circumstances. In other words, if the maxim has already been acted on or the goal has already been achieved, then prohibiting it is impossible. In any model that has at least one term, it is possible to construct a maxim where the circumstances, goal, and act (once a subject acts on it) are all that same term, resulting in a contradiction.

To get around this, I will exclude 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. To understand why, I return to Korsgaard's and O'Neill's interpretations of a maxim as a practical guide to action. 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 not contradictory or prohibited, but it is the wrong kind of question to be asking. It is not a well-formed maxim, so the categorical imperative does not apply to it. (more explanation in philosophical writing collection)

abbreviation *well-formed*::*maxim*⇒*s*⇒*i*⇒*bool* **where**

well-formed $\equiv \lambda(c, a, g).\ \lambda s.\ \lambda w.\ (\neg\ (c \rightarrow g)\ w) \wedge (\neg\ (c \rightarrow a\ s)\ w)$

— This abbreviation formalizes the well-formedness of a maxim for a subject. The goal cannot be already achieved in the circumstances and the subject cannot have already performed the act.

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

— Let's try the exact same formalization of the FUL as above, except that it only applies to maxims that are well-formed at every world.

lemma *FUL*

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

— The FUL does not hold in DDL, because nitpick is able to find a model for my system in which it is false. If the FUL were already a theorem of the system, adding it wouldn't make the system any more powerful, so this is the desired result.

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

Skolem constants: M = $((\lambda x.\_)(i_1 :=$ True), $(\lambda x.\_)(s_1 := (\lambda x.\_)(i_1 :=$ False)), $(\lambda x.\_)\ (i_1 :=$ False)) λ w. p = $(\lambda x.\_)(i_1 := s_1)$ s = s_1

axiomatization where *FUL*:*FUL*

lemma *True*

nitpick[*user-axioms*, *falsify*=*false*] **by** *simp*

— Nitpick is able to find a model in which all axioms are satisfied, so this version of the

FUL is consistent.

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

Empty assignment

During the process of making FUL0 consistent, I used Isabelle to gain philosophical insights about vacuous maxims. This process is an example of the power of computational tools to aid philosophical progress. I used Nitpick and Sledgehammer to quickly test if a small tweak to FUL0 fixed the inconsistency or if I was still able to derive a contradiction. I then realized that if I defined the circumstances, act, and goal as constants, then FUL0 was indeed consistent. After some experimentation, Prof. Amin correctly pointed out that as constants, these three entities were distinct. However, when merely quantifying over (c, a, g), all members of a tuple could be equivalent. Within a minute, I could formalize this notion, add it to FUL0, and test if it solved the problem. The fact that it did spurred my philosophical insight about vacuous maxims.

The logic confirmed 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. The syntax and typing of Isabelle/HOL forced me to bind the free-variable M in the FUL in different ways and allowed me to quickly test many bindings. The discovery of this logical inconsistency then enabled a philosophical insight about which kinds of maxims make sense as practical principles. This is one way to do computational ethics: model a system in a logic, use computational tools to refine and debug the logic, and then use insights about the logic to derive insights about the ethical phenonema it is modelling. This procedure parallels the use of proofs in theoretical math to understand the mathematical objects they model.

One potential problem with my formalization is that it does not use the modal

nature of the system. All of the properties that the FUL investigates hold at all worlds, in effect removing the modal nature of the system. This approach simplifies logical and therefore computational complexity, improving performance. On the other hand, it doesn't use the full expressivity of DDL. If I run into problems later on, one option is to tweak the FUL to use this expressivity.

end

theory *paper412* **imports** *paper41*

begin

3.3 Application Tests

As with the naive formalization and Kroy's formalization, I will apply my testing framework to my custom formalization of the FUL. I will begin with some basic application tests. In these tests, I specify particular maxims as constants with no properties and gradually add properties to understand how the system handles different kinds of maxims.

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**

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

— 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 $\forall s. \models (\textit{prohibited false-promising } s)$

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[34]. 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)

— 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*:

[34]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.

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 $\forall s. \models (\textit{prohibited false-promising } s)$

proof–

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

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

thus *?thesis*

using *FUL assms*(*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[35]. 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 $\textit{everyone-can't-lie} \equiv \forall\, w.\ \neg\ (\forall\, s.\ \textit{falsely-promise}\ s\ w)$, 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[36]

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,

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

[36]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.

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 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

This completes the application tests for my formalization. I showed that my formalization correctly handles an example from Korsgaard with two possible interpretations and also sufficiently handles both conventional and natural acts.

3.4 Metaethical Tests

Recall that metaethical tests test formal properties of the system that apply to any maxim, not just those specified in the application tests. In this section I adapt the metaethical tests developed in previous sections to my formalization of the categorical imperative. I preserved the philosophical goal of each test but modified them to test the stronger, richer notion of a maxim.

The first set of tests consider how obligation generalizes, first across worlds and then across people. As expected, the tests below show that both wrongness (prohibition) and rightness (obligation) generalize across both worlds and people. In other words, if something is obligated at some world, it is obligated at every world

and if something is obligated for some person, then it is obligated for every person.

Generalization across worlds is a consequence of the fact that my interpretation does not make use of the modal nature of DDL. In particular, I do not use any property of the world when prescribing obligations at that world.

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

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

by *blast*

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

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

by *blast*

lemma *wrong-if-wrong-somewhere*:

shows $\forall c\ g.\ \exists w1.\ O\{\neg\ (W\ (c, M, g)\ s)\ |\ c\}\ w1 \longrightarrow (\forall w2.\ \ O\{\neg\ (W\ (c, M, g)\ s)\ |\ c\}\ w2)$

by *blast*

lemma *right-if-right-somewhere*:

shows $\forall c\ g.\ \exists w1.\ O\{W\ (c, M, g)\ s\ |\ c\}\ w1 \longrightarrow (\forall w2.\ \ O\{W\ (c, M, g)\ s\ |\ c\}\ w2)$

by *blast*

As expected, obligation generalizes across people and worlds. In the next set of tests, I will analyze basic properties of permissibility, obligation, and prohibition.

First, I verify that the logic allows for permissible maxims, as this is a problem that prior iterations ran into. Below, I use Nitpick to find a model in which there is a circumstance, act, goal tuple that is permissible but not obligated at some world.

lemma *permissible*:

shows $((\neg\ (O\{(W\ (c, a, g)\ s)\ |\ c\}))\ \wedge\ (\neg\ (O\{\neg\ (W\ (c, a, g)\ s)\ |\ c\})))\ w$

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

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

Free variables: a = $(\lambda x.\_)(s_1 := (\lambda x.\_)(i_1 := \text{False}))$ c = $(\lambda x.\_)(i_1 := \text{False})$ g = $(\lambda x.\_)(i_1 := \text{False})$ s = s_1Recall that Nitpick is a model checker that finds models making certain formulae true or false. In this case, Nitpick finds a model satisfying the given formula (which simply requires that the sentence "s wills (c, a, g)" is permissible but not obligator). This model consists of the above specifications of a, c, g, and s.

I also expect that any arbitrary maxim should be either permissible or prohibited, since all acts are either permissible or prohibited.

lemma *perm-or-prohibited*:

shows $((permissible\ (c, a, g)\ s) \vee (prohibited\ (c, a, g)\ s))\ w$

by *blast*

— This simple test passes immediately by the definitions of permissible and prohibited.

Obligation should be strictly stronger than permissibility. In other words, if a maxim is obligated at a world, it should be permissible at that world. Below I test this property.

lemma *obligated-then-permissible*:

shows $(O\{W(c, a, g)\ s|c\} \rightarrow ((permissible\ (c, a, g)\ )\ s))\ w$

using *no-contradictions* **by** *auto*

— This test passes and Isabelle is able to find a proof for the fact that all obligatory maxims are also permissible.

The above test failed under Kroy's formalization of the categorical imperative and is thus evidence that my formalization improves upon Kroy's. Interestingly, this new test passes because of the additional added axiom that prohibits contradictory obligations (recall that Kroy's formalization allowed contradictory obligations).

Next, I will test if the formalization allows for vacuous obligations or modal collapse. These tests are sanity checks confirmed that the obligation operator is

doesn't collapse. First, I will check that any arbitrary term isn't obligated.

lemma *arbitrary-obligations*:

fixes c A::t

shows $O\{A|c\}$ w

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

— This test passes—Nitpick finds a model where A isn't obligated in circumstances c. Nitpick found a counterexample for card i = 1 and card s = 2:

Free variables: A = $(\lambda x.\_)(i_1 :=$ True) c = $(\lambda x.\_)(i_1 :=$ False) Previous iterations of this test used the monadic obligation operator, which tests the term in the context "True" (equivalently the set of all worlds since True holds everywhere). In this iteration, I test the term in a context c, because my formalization uses the dyadic obligation operator and must thus specify circumstances.

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'. The proposition below should be falsifiable.

lemma *modal-collapse*:

shows $((W\ (c, a, g)\ s)\ w) \longrightarrow O\{W\ (c, a, g)\ s|c\}\ w$

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

— Nitpick finds a counterexample, so willing doesn't imply obligation, so this test passes. Nitpick found a counterexample for card i = 1 and card s = 2:

Free variables: a = $(\lambda x.\_)(s_1 := (\lambda x.\_)(i_1 :=$ False), $s_2 := (\lambda x.\_)(i_1 :=$ False)) c = $(\lambda x.\_)(i_1$:= False) g = $(\lambda x.\_)(i_1 :=$ False) s = s_2 w = i_1Once again, I modify this test to use the dyadic obligation operator instead of the monadic operator.

The final set of tests deal with ought implies can and conflicting obligations. Recall that I specifically added an axiom in my formalization to disallow contradictory obligations, so I expect these tests to pass. Kroy's formalization fails these tests, so this is another area of improvement over Kroy's formalization.

lemma *ought-implies-can*:

shows $O\{W\ (c, a, g)\ s|c\}\ w \longrightarrow (\Diamond\ W\ (c, a, g)\ s)\ w$

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

— This test is a lemma of DDL itself, so it's no surprise that this test passes.

lemma *conflicting-obligations*:

shows $\neg\ (O\{W\ (c, a, g)\ s|c\} \wedge O\{\neg(W\ (c, a, g)\ s)|\ c\})\ w$

using *no-contradictions* **by** *blast*

— This test passes immediately by the new axiom prohibited 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*

— Recall that the we also expect the stronger property that the combination of obligatory maxims can't imply a contradiction. The added axiom also makes this test pass.

lemma *distribution*:

assumes $\models (O\ \{A\} \wedge O\ \{B\})$

shows $\models O\ \{A \wedge B\}$

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

The metaethical test suite ran on both Kroy's formalization and my formalizaion show two clear improvements. First, my formalization shows that obligatory maxims are permissible, whereas Kroy's paradoxically does not. Second, my formalization doesn't allow contradictory maxims, but Kroy's does. Both of these improvements are derived from the new axiom I added in my formalization that disallows contradictory obligations. Additionally, my formalization also improves on Kroy's by staying faithful to the strongest interpretation of the FUL, Korsgaard's practical contradiction interpretation. (maybe stick philosophical writing here or above?)

3.5 Formalization Specific Tests

In this section, I explore tests specific to my formalization of the categorical imperative. First, in my previous (buggy) implementation of DDL, prohibiting contradictory obligation led to the strange result that all permissible actions are obligatory. I will test if this bug appears in this implementation as well.

lemma *bug*:

shows *permissible* $(c, a, g)\ s\ w \longrightarrow O\{W(c, a, g)\ s \mid c\}\ w$

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

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

Free variables: a = $(\lambda x.\_)(s_1 := (\lambda x.\_)(i_1 := \text{False}), s_2 := (\lambda x.\_)(i_1 := \text{False}))$ c = $(\lambda x.\_)(i_1 := \text{False})$ g = $(\lambda x.\_)(i_1 := \text{False})$ s = s_2 w = undefined This strange result does not hold; good!

4 Applications

In this chapter, I demonstrate my system's ability to formalize longer, more complicated ethical arguments and present the additional capabilities necessary to use my system in practice. In Chapter 2 and Chapter 3, I performed metaethical reasoning while testing different formalizations of the FUL. Metaethical reasoning analyzes properties of moral thought itself and involves questions about the nature of ethical truth. This kind of reasoning contrasts with "applied ethical reasoning," which is the use of ethics to resolve dilemmas and make judgements about what an agent should or should not do. Applied ethical reasoning is relevant with respect to an agent's particular situation, not just to an ethical theory as an abstract entity. In previous chapters' application tests, I performed some toy examples of applied ethical reasoning, but this chapter is an extended exploration of my system's ability to perform applied ethical reasoning. Metaethical reasoning is most interesting

to philosophers who are trying to formulate the "best" theory of ethics. Applied ethical reasoning, on the other hand, is useful to ordinary people who are trying to decide how to live their lives. In order for automated ethics to guide AI agents, it must perform applied ethical reasoning.

One challenge of applied ethical reasoning is that it requires more common sense knowledge than metaethical reasoning. Because metaethics is about ethics itself, and not about the dilemmas that ethics is supposed to help us resolve, this kind of reasoning requires very little knowledge about the world. In previous chapters, I perform metaethical reasoning using my system (which formalizes an ethical theory) and basic logical facts and objects. Applied ethical reasoning, on the other hand, focuses on a particular ethical dilemma and thus requires enough "common sense" to understand the dilemma and options at hand. For example, an applied ethicist evaluating the permissibility of lying needs some robust definition of the term lying and likely some understanding about the activities of communication and truth telling. Kantians specifically describe this common sense as "postulates of rationality" that are nontrivial and nonnormative, but still part of the process of practical reasoning itself (Silber, 1974). Powers (2006) notes that common sense is a hurdle for an automated Kantian ethical agent. In this chapter, I attempt to tackle this challenge and endow my system with this kind of common sense in the specific case of lying. In order for my system to be used to perform applied ethical reasoning, it will need to be equipped with a database of common sense facts and definitions as I present in this chapter. My system will contribute the core reasoning about the Formula of Universal Law, but this reasoning must be applied to objects that are defined in this common sense database. In this chapter, I try to understand what this common sense reasoning looks like for the specific example of lying.

Because these common sense facts can determine my system's judgements, they

are part of the trusted code base for my system. In order for someone to trust my system's judgements, they must trust that the common sense database is correct becase changing these common sense facts will change the judgements that my system makes. For example, if we define truth telling as an act that is self-contradictory (perhaps by defining it as $p \wedge \neg p$), then my system will output that truth telling is prohibited. Malicious common sense facts and definitions will result in bad ethical judgements. In other words, garbage in, garbage out. This common sense reasoning is also not a part of the system I contribute here—in order for my system to be used, future work must develop a common sense database for the particular situation at hand. The challenge of endowing automated ethical reasoners with common sense reasoning is not unique to my system, and virtually all prior attempts in machine ethics face similar challenges.[37] Most prior attempts sidestep this question, whereas I contribute an prototype implementation of one kind of common sense reasoning.

The specific kind of common sense reasoning required appears to be a challenge for automating Kantian ethics, and may imply that consequentialist or virtue ethical automated agents are within closer reach for automation. Ultimately, Kantian ethics is still easiest to automate because it will require fewer, less controversial common sense facts than other ethical systems. As the examples in this section demonstrate, Kantian ethics requires a definition of lying (which any other theory would also requires) and the knowledge that if everyone lies in a given context, no one will believe each other in that particular context. This latter fact may not be required for every ethical theory, but is relatively uncontroversial. It is a kind of intuition about human behavior that is generally accepted. Neither a definition of lying nor this property of lying seem like unreasonable prerequisites for ethical reasoning.

[37] See Section Related Work for a survey of the common sense required in prior work.

Consequentialism, on the other hand, would require much more specific data about the consequences of a lie, perhaps for the specific person's credibility, for the victim of the lie, for the third-parties watching the lie unfold. Consequentialism requires more numerous and specific judgements, all of which are likely to be more controversial than the two outlined above for Kantian ethics. Similarly, virtue ethics would likely require information about the actor's entire moral character, including their attitude towards the lie and their moral history. Virtue ethics would also require robust definitions of the relevant virtues, in addition to a definition of lying. Thus, while Kantian ethics requires some common sense reasoning, it requires fewer and less controversial background facts than other ethical theories.

The extended examples presented in this section also demonstrate the difficulty of formulating a maxim to pass as input to my implementation of the FUL. A large part of the challenge of applied Kantian ethics is formulating a maxim that accurately captures an agent's principle of action, so a totally automated agent using my system will need some way to formulate these maxims well. In this section, I manually implement Korsgaard's formulation of certain maxims, and I will later argue that manual formulation of maxims is, at present, the way forward.

This chapter will provide additional examples of the kinds of common sense facts and maxim formulation required to get my system off the ground. I will aim to use a lean and uncontroversial common sense database to achieve robust and powerful results. This serves as evidence for the ease of automating Kantian ethics, an example of what additional work my system requires, and a demonstration of the contributions that I make. These examples demonstrate that nuanced common sense facts and maxims can cause my system to contribute nuanced judgements faithful to philosophical literature. On one hand, this means that my system can perform sophisticated ethical reasoning, but on the other, the quality of this reasoning is heavily reliant on the common sense database and the input maxim. Thus,

my system *could* make smart ethical judgements in practice, but getting to that point will require a robust, trusted common sense database and maxim formulator. In this paper, I contribute an implementation of the Formula of Universal Law that, when equipped with relatively thin common sense facts, can performed nuanced, sophisticated ethical reasoning. When given as input a maxim in the format specified in Chapter 3, my system can show that the maxim is permissible, obligatory, or prohibited. It can provide a verifiable proof of its judgement and a human-readable list of the facts or axioms it used to reach its conclusion. My system is faithful to philosophical literatire on Kantian ethics and serves as the first step towards a fully or partially automated ethical AI agent. It also establishes the power and potential of computational ethics, or the use of computational tools for ethical investigation, as explored in this section's Applied Ethical reasoning.

Stick a bit here about the philosophical work that will go in this section

4.1 Simple Lying Examples

This chapter focuses on the example of lying because this case is hotly debated in the Kantian literature. I draw on examples of ethical reasoning as presented in Korsgaard's "Right to Lie," which examines exactly how strict Kant's prohibition on lying is. She picks up a long-running debate in the literature through the example of someone who shows up at your door and asks, "Is Sara home?" Unbeknownst to them, you know that they want to know Sara's location in order to murder her. Ordinary moral intuition prescribes that, if Sara is home, you should lie and say that she is not in order to protect her from the murderer, but the categorical imperative seems to prohibit lying in all circumstances. In this section, I will formalize Korsgaard's treatment of lying and joking under the categorical imperative, focusing on the common sense assumptions necessary to achieve her conclusions. In the next section, I will formalize the core of Korsgaard's argument that the categorical im-

perative coheres with ordinary intuition and does not prohibit lying to the murderer to protect Sara.

First, Korsgaard argues that the categorical imperative appears to prohibit all lies because, when universalized, lies will no longer be believed. Thus, lying could never be an effective way of achieving any goal when universalized. Korsgaard points out that "we believe what is said to us in a given context because most of the time people in that context say what they really think" (Korsgaard, 1986, 4). In order to formalize this argument, I first need to define the relevant terms.

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

— Person s::subject believes sentence t::term. The concept of belief will play a crucial role both in the arguments for lying being prohibited and for the maxim of lying to the murderer being permissible. Logicians and epistemologists have developed robust, complex logics of belief and knowledge (Baltag and Renne, 2016). For the sake of this project, I avoid this complexity by merely defining the concept of belief as an empty constant that maps a subject, term pair to a term. For the examples in this section, this choice suffices, as my common sense beliefs encode enough properties of belief for my purposes. Future work could integrate a much more complex logic of belief into my system.

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

— Person s::subject utters sentence t::term.

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

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

— Person s utters falsehood t if and only if s utters t and t is false.

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

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

— Person s::subject knowingly utters a falsehood t, if they both utter t as a falsehood and don't believe t. This and the above abbreviations are the core of my formalization of Korsgaard's definition of lying. They are also relatively uncontroversial and have little normative content.

abbreviation *lie*::*maxim*⇒*bool* **where**

lie ≡ λ (*c*, *a*, *g*). ∃*t*. (*a* ⟶ (λ*s*. *knowingly-utter-falsehood s t*)) ∧ (∃*p*. ∀*w*. (*g* → (*believe p t*)) *w*)

— Using the above definitions, I can characterize a maxim as a lie if (a) the act requires knowingly uttering a falsehood and, (b) the end requires that some person p believe the false statement t.

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

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

consts *me*::*s*

— I am trying to reason about *my* obligations so I will define myself as a specific subject.

consts *m*::*maxim*

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

consts *c*::*t* *a*::*os* *g*::*t*

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

In the following lemma, I use my system to show that lying is prohibited. The assumptions of this lemma represent the logical and common sense background necessary to reach this conclusion. This common sense background is a direct for-

malization of the premises of Korsgaard's argument. Using these relatively "thin" premises about individual behavior, my system can derive a prohibition against lying.

lemma *lying-prohibited*:

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

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

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

assumes *lie m*

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

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

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

assumes $\forall w.\ c\ w$

— Restrict our focus to worlds in which the circumstances hold, as these are the morally interesting worlds for this example. A technical detail.

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

proof −

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

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

— Unlike many of the other proofs in this project, this proof is a little heavier and requires some manual work to produce. After I divide the proof into the intermediate steps shown here, and Isabelle is able to do the rest. This step says that if m is universalized, then the

circumstances won't lead to the goal, which is quite close to the idea of the maxim not being universalizable.

have *not-universalizable m me*

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

thus *?thesis*

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

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

qed

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

abbreviation *joke*::*maxim*⇒*bool* **where**

joke ≡ λ (*c*, *a*, *g*). ∃ *t*. (*a* ⟶ (λ*s*. *knowingly-utter-falsehood s t*)) ∧ ¬ (∃ *p*. ∀ *w*. (*g* → (*believe p t*)) *w*)

This definition of a joke merely defines a joke as a falsehood uttered for some purpose that doesn't require deception, where deception involves someone believing the uttered falsehood. Notice that this is quite a thin definition of a joke; it doesn't require any conception of humor but merely distinguishes jokes from lies. As far as common sense reasoning goes, this is a relatively tame proposition. I will now demonstrate that this conception of a joke is sufficient to show that joking is permissible.

Korsgaard argues that her above argument for a prohibition against lying also implies that joking is permissible, because its purpose is *not* to decieve, but something

else entirely. This means that, even armed with the same core convention of lying assumption as above, joking must be permissible. The lemma below shows exactly that.

lemma *joking-not-prohibited*:

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

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

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

assumes *joke m*

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

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

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

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

— Restrict our focus to worlds in which the circumstances hold, as these are the morally interesting worlds for this example. An irrelevant technical detail.

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

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

One potential worry with the above argument is the fact that my definition of joke requires that achieving the goal of a joke does not rely on anyone believing the falsehood told in the joke. On its face, this is not a complete representation of a joke; the ordinary conception of a joke would be something like a false statement uttered with the goal of amusing. I define the goal of the joke as not requiring that anyone believe the false statement to distinguish it from lying, which has deception as the goal. I intentionally avoid providing robust formalizations of both deception and amusement; instead I present a thin conception that relies entirely on the concept of belief. This conception does not require any notion of humor, satire, intention, or malice, all of which are concepts that may be necessary to de-

fine amusement and deception completely.

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

This account not only describes the kind of maxims that fail or pass the universalizability test, it also provides a guide to constructing permissible maxims about uttering falsehoods. As an example, consider the idea of throwing a surprise birthday party. At first glance, the maxim of action is something like, "When it is my friend's birthday, I will secretly plan a party so that I can surprise them." The goal "so that I can surprise them" clearly requires that your friend believe the falsehood that you are not planning a party, else the surprise would be ruined. The analysis above seems to imply that Kantian ethics would prohibit surprise parties, which is a sad conclusion for birthday-lovers everywhere. Noticing that the problem with this maxim is that the goal requires belief in a falsehood provides a way to rescue the

beloved concept of surprise parties. When throwing a surprise party, the ultimate objective is *not* to surprise your friend, but is to celebrate your friend and help them have a fun birthday. If someone ruins the surprise, but the party is still fun and the birthday person still feels loved, then we would consider such a party a success! Someone who called this party a failure clearly would be missing the point of a surprise party. Thus, the goal of a surprise party is not the surprise itself, but rather to celebrate the birthday person. Thus modified, the goal no longer requires belief in the falsehood and thus passes the universalizability test.

The implications of this section are twofold. First, my system is capable of performing ethical reasoning sophisticated enough to show that lying is prohibited but joking is not. The sophistication necessary to distinguish between lying and joking was a direct consequence of my system's use of a robust conception of a maxim, which encoded the goal of an act as part of the maxim being evaluated. Second, in the process of making this argument precise, my system isolated a necessary and sufficient condition of a maxim about uttering a falsehood being prohibited: that the goal require that someone believe the falsehood. This condition both made an long-standing argument in Kantian ethics more precise and can guide the correct formulation of future maxims. In other words, an insight generated by the computer provides value to ethicists.

Moreover, all of the reasoning in this chapter required relatively few and uncontroversial common sense facts. The deepest assumption required was that, if everyone lies about a given statement, no one will believe that statement. This assumption is not merely definitional; it does encode some synthetic knowledge about the world, but it is relatively uncontroversial. Indeed, it is so well-accepted that Korsgaard does not bother to justify it in her argument. These examples showed that, while common sense reasoning is an obstacle that must be overcome for my system to be used in practice, it is surmountable.

4.2 Lying to a Liar

Once Korsgaard completes her preliminary work differentiating between lies and jokes, she begins her main argument, which examines the controversial case of the murderer at the door. Recall that the murderer appears at your door and asks if his intended victim is at home. Ordinary intuition requires that it is at the very least permissible (if not obligatory) to lie to the murderer in order to protect the victim. Korsgaard notes that a murderer who wishes to find his victim cannot simply announce his intentions to murder; instead, he must "must suppose that you do not know who he is and what he has in mind" (Korsgaard, 1986, 5).[38] Thus, she can modify the maxim in question to specify that when someone lies to you, you are allowed to lie to them. The maxim of lying to the murderer is actually the maxim of lying to a liar, which she argues is permissible. Notice that her argument hinges on this clever, but ultimately sensical formulation of your maxim. She notes that there is something relevant and significant about the fact that the person demanding to know Sara's location is a murderer and that he is trying to take advantage of your honesty. This claim is not unfounded or wildly controversial, but it does demonstrate the importance of correctly formulating the maxim to test. In this section, I will formalize her argument for lying to a liar.

As usual, I first define my terms.

consts *murderer*::*s*

— This example involves one more subject: the murderer.

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

— This statement represents the lie that the murderer tells you. By not announcing his intention, he is implicitly telling you that he is not a murderer, as people normally assume

[38] Korsgaard assumes that the murderer will lie about his identity in order to take advantage of your honesty. In footnote 5, she accepts that her arguments will not apply in the case of the honest murderer who announces his intentions, so she restricts her focus to the case of lying to a liar. She claims that in the case of the honest murderer, the correct act is to refuse to respond, but she does not argue for this in this paper.

that those knocking on their door are not murderers.

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

— These are the circumstances that the murderer is in.

consts *find-victim*::*t*

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

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

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

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

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

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

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

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

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

consts *protect-victim*::*t*

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

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

my-maxim ≡ (*murderer-at-door*, λs. *knowingly-utter-falsehood s victim-not-home*, *protect-victim*)

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

Now that I have defined the maxims at hand, I can begin reasoning about them. First, I will show that, using the same convention of lying common sense fact as above, the murderer's maxim is prohibited. Effectively, this tests that the assumption is indeed strong enough to prohibit lying.

lemma *murderers-maxim-prohibited*:

assumes $\forall w.$ *well-formed murderers-maxim murderer w*

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

assumes $\models$ (*find-victim* $\rightarrow$ (*believe me not-a-murderer*))

— Assumption that, in order for the murderer to find their victim, you must not believe that he is a murderer. This is an example of the kind of situation-specific common sense reasoning necessary to use my system. Again, this is Korsgaard's uncontroversial assumption; the murderer assumes that if you knew he was a murderer, you would not disclose the victim's location to him.

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

— The convention of lying common sense assumption from above.

assumes $\forall w.$ *when-at-my-door w*

— Restrict our focus to worlds in which the circumstance of the murderer being at my door holds, as these are the morally interesting worlds for this example. An irrelevant technical detail.

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

proof −

— Again, this proof is too heavy for Isabelle to finish on its own, so I needed to specify some intermediate steps. The same intermediate steps as above sufficed, effectively providing a pattern for the proof. Isabelle does allow users to define custom 'proof methods,' so a more robust version of my system could define this proof pattern as a method and apply it in cases involving lies.

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

using *assms*(*2*) *assms*(*4*) **by** *auto*

have *not-universalizable murderers-maxim murderer*

using *assms*(*2*) *assms*(*4*) **by** *auto*

thus *?thesis*

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

qed

I will now formalize Korsgaard's argument for the permissibility of lying to a liar.

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

lemma *lying-to-liar-permissible*:

assumes $\models$ (*well-formed murderers-maxim murderer*)

assumes $\models$ (*well-formed my-maxim me*)

— Assume that we're working with well-formed maxims.

assumes $\models$ (*protect-victim* $\rightarrow$ (*murderer believes victim-not-home*))

— In order for you to successfully protect the victim, the murderer must believe that the victim is not home. This is a noncontroversial assumption about the specific act at hand.

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

— This is one of two assumptions that encode Korsgaard's core argument. If person1 believes that person2 utters a sentence as a falsehood, then person1 won't believe that sentence. This is a modification of the convention of lying assumption from above, and I will refer to it as the "convention of belief" assumption. Again, like the convention of lying assumption, this assumption is uncontroversial: if I think you are saying a false sentence, then I won't believe that sentence.

assumes $\forall$ *c a g w*. (*universalized* (*c*, *a*, *g*) *w*) $\longrightarrow$ ((*person1 believes* (*person2 believes* *c*)) $\rightarrow$ (*person1 believes* (*a person2*))) *w*

— This is the second major common sense assumption. If the maxim (c, a, g) is universalized, then if person1 believes person2 believes they are in the given circumstances,

then person1 believes person2 performs the act. In other words, person1 will believe that person2 wills the maxim. I will refer to this as the "convention of willing" assumption. This follows directly from Korsgaard's conception of universalizability: when a maxim is universalized, everyone wills it and thus notices the pattern of everyone willing it. If you observe that many do X in circumstances C, you will assume that everyone does X in circumstance C.

assumes $\forall w.$ *murderer-at-door w*

— Restrict our focus to worlds in which the circumstance of the murderer being at my door holds, as these are the morally interesting worlds for this example. An irrelevant technical detail.

shows $\models$ (*permissible my-maxim me*)

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

— Notice the use of the first and sixth assumption in this automatically generated proof. Essentially, the common sense assumptions given are not strong enough to generate a prohibition against lying to a liar, and are thus unused in this proof.

The above lemma shows that, with a more nuanced set of common sense facts, my system can show that lying to a liar is permissible. Moreover, I know that this set of assumptions is correct because it can also show that the murderer's maxim is prohibited. I show this in the lemma below.

lemma *murderers-maxim-prohibited2*:

assumes $\forall w.$ *well-formed murderers-maxim murderer w*

— The murderer's maxim is a well-formed maxim composed of some circumstances, act, and goal.

assumes $\models$ (*find-victim* $\rightarrow$ (*believe me not-a-murderer*))

— Assumption that, in order for the murderer to find their victim, you must not believe that they are a murderer.

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

— The convention of belief assumption from above.

assumes $\forall$ *c a g w*. (*universalized* (*c*, *a*, *g*) *w*) $\longrightarrow$ ((*person1 believes* (*person2 believes*

$c)) \rightarrow (person1\ believes\ (a\ person2))))\ w$

— The convention of willing assumption from above.

assumes $\forall\, w.\ when\text{-}at\text{-}my\text{-}door\ w$

— Restrict our focus to worlds in which the circumstance of the murderer being at my door holds, as these are the morally interesting worlds for this example. An irrelevant technical detail.

shows $\models (prohibited\ murderers\text{-}maxim\ murderer)$

proof −

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

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

have *not-universalizable murderers-maxim murderer*

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

thus *?thesis*

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

qed

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

While it is true that lying to the murderer should be permissible, Korsgaard notes that many will want to say something stronger, like the fact that lying to the murderer is obligatory in order to protect the intended victim (Korsgaard, 1986, 15). It seems like we would be doing something wrong if I revealed the victim's location,

knowing that this revelation would cost them their life. Korsgaard solves this problem by noting that, while the FUL shows that lying to the murderer permissible, other parts of Kant's ethics show that it is obligatory. Recall that Kant presents perfect and imperfect duties, where the former are strict, inviolable, and specific and the latter are broader prescriptions for action. Perfect duties always supersede imperfect duties when the two conflict. For example, the duty to not murder is a perfect duty and the duty to give to charity is an imperfect duty. The FUL generates perfect duties and Kant's extended theory of virtues generates imperfect duties. The details of this theory and these distinctions are outside the scope of this paper, but the crucial note is that other parts of Kant's ethical theory generate the obligation to lie to the murderer. I chose to formalize the FUL because it is, in some sense, the strongest of version of the categorical imperative. An even more sophisticated Kantian reasoner could formalize his theory of virtue and his other formulatations of the categorical imperative in order to generate the obligation to lie to the murderer, but the FUL is the strongest and most foundational of these principles. The fact that my system merely shows that lying to the murderer is permissible, but not obligatory is consistent with the part of Kant's ethical theory that I formalize and demonstrates that I have faithfully implemented the FUL.

While this example demonstrates the power of my system (when equipped with some common sense), it also shows how vital the role of the common sense reasoning is. Slight, intuitive changes in the common sense facts achieved totally different conclusions about lying. This represents an obstacle to fully automated ethical reasoning; such an agent would need a trusted database of common sense facts, which is still an unsolved problem. My work is one step towards such an agent, but the importance of common sense means that much progress must be made in order to completely automate ethics.

The reasoning of this section also demonstrated one additional place where a Kan-

tian must make vital judgements: the formulation of the maxim itself. Korsgaard's argument for the permissibility of lying to a murderer hinged on a clever formulation of the maxim highlighting a particular facet of the circumstances, namely that the murderer is lying to you. Indeed, there is robust debate in the literature on what circumstances should be considered when formulating a maxim. Some critics of Kant raise the "tailoring objection," which is the worry that arbitrarily specific circumstances render any maxim universalizable. For example, the maxim "When my name is Lavanya Singh and I am wearing a purple shirt and it is November 26th, I will lie in order to get some easy cash" passes the universalizability test. Even if this maxim is willed universally, the circumstances are so specific that lying will not become the general mechanism for getting easy cash, so the lender will believe my lie and the maxim will remain effective. By tailoring the circumstances, any maxim can evade universalization.

The Kantian response to this criticism is to require that the circumstances included in the formulation of the maxim be "morally relevant." In the example above, my purple shirt and the date clearly have no bearing on the moral status of lying. On the other hand, consider the maxim, "When I am unemployed, I will murder someone in order to take their job." The circumstances of being unemployed clearly have some bearing on the moral relevance of the murder in question; they speak to the motivation for the murder. While this view seems to track how we actually perform moral reasoning, it leaves open the question of how to determine which circumstances are morally relevant. Here, O'Niell reminds us that the Formula of Universal Law is a "test of moral worth rather than of outward rightness" (O'Neill, 1990, 98). The FUL is a way for an agent to decide how they should behave, not for a third-party to judge their behavior. Ethics is a personal process for Kant, so the FUL is designed to help agents internally make decisions, not to judge others' decisions. Because agents use the FUL to evaluate their own behavior, the test is at

its best when they make a good faith effort to isolate the *principle* of their action, rather than some "surface intent" (O'Neill, 1990, 87). The FUL is supposed to determine if an agent's principle of action is universally consistent, so it is at its most effective when an agent accurately formulates the principle they act on. Circumstances are morally relevant if they accurately reflect the way that the agent is thinking about their own action. In the example above, the circumstance of wearing a purple shirt doesn't reflect the principle of the liar's action. Its inclusion is clearly a disingenous attempt to evade the universalizability test, but because the FUL is a test of personal integrity, it cannot withstand this kind of mental gymnastics.

While this account of the formulation of a maxim prescribes how a well-intentioned agent should decide how to live their life, it poses a challenge for automated ethics. In order for an automated ethical agent to use the categorical imperative to its fullest extent, the input maxim fed into my system or any automation of the FUL must be a good-faith attempt to capture the agent's principle of action. However an action is turned into a maxim for my system to process, whether manually as I do during these tests or automatically using some kind of input parser, this transformation from action to maxim has huge bearing on the outcome of the test. The formulation of a maxim must be a good-faith attempt to capture the principle of action, and must therefore include only the morally relevant circumstances, and nothing more. This is a significant judgement that my system does not make, and is thus another hurdle that must be overcome in order for my system to be used in practice. I will argue in Section ?? that this kind of input parsing work should be left to human beings for now, and that major technical and philosophical progress must be made to automate this portion of the system.

The formulation of a maxim and the common sense database pose the two greatest challenges to the adoption of my system in practice. In this chapter, I argued that using manual, human involvement, these challenges can be overcome in relatively

uncontroversial ways. They are also ripe areas for future work.

4.3 Philosophical Analysis: Is Automated Ethics a Good Idea?

5 Related Work

In 1685, Leibniz dreamed of a universal calculator that could be used to resolve philosophical and theological disputes. At the time, the logical and computational resources necessary to make his dream a reality did not exist. Today, automated ethics is a growing field, spurred in part by the need for ethically intelligent AI agents.

Tolmeijer et al. Tolmeijer et al. (2021) developed a taxonomy of works in implementing machine ethics. An implementation is characterized by (1) the choice of ethical theory, (2) implementation design decisions (e.g. testing), and (3) implementation details (e.g. choice of logic).

In this paper, I formalize Kantian ethics. There is a long line of work implementing other kinds of ethical theories, like consequentialism Abel et al. (2016); Anderson et al. (2004) or particularism Ashley and McLaren (1994); Guarini (2006). Kantian ethics is a deontological, or rule based ethic, and there is also prior work implementing other deontological theories Govindarajulu and Bringsjord (2017); Anderson and Anderson, 2014). Kantian ethics specifically appears to be an intuitive candidate for formalization and implementation Powers (2006); Lin et al. (2012). In 2006, Powers Powers (2006) argued that an implementation of of Kantian ethics presented technical challenges, such as automation of a non-monotonic logic, and philosophical challenges, like a definition of the categorical imperative. There has also been prior work in formalizing Kantian metaphysics using I/O logic Stephenson et al. (2019). Deontic logic itself is inspired by Kant's "ought implies can" principle, but it does not include a robust formalization of the entire categori-

cal imperative Cresswell and Hughes (1996).

Lindner and Bentzen Bentzen and Lindner (2018) have presented a formalization and implementation of Kant's second formulation of the categorical imperative using a custom logic. They present their goal as "not to get close to a correct interpretation of Kant, but to show that our interpretation of Kant's ideas can contribute to the development of machine ethics." My work aims to formalize Kant's ethic as faithfully as possible. I draw on the centuries of work in moral philosophy, as opposed to developing my own ethical theory. I also hope to formalize the first and third formulations of the categorical imperative, in addition to the first.

The implementation of this paper builds on Benzmueller, Parent, and Farjami's work with the LogiKey framework for machine ethics Benzmüller et al. (2021); Benzmüller et al. (2019). The LogiKey project has been used to implement metaphysics Benzmüller and Paleo (2013); Kirchner et al. (2019). Fuenmayor and Benzmueller Fuenmayor and Benzmüller (2018) have implemented Gewirth's principle of generic consistency, which is similar to Kant's formula of universal law.

6 Future Work

I intend to continue this research for the next year as part of my senior thesis. To make that process easier, I will sketch some goals for the rest of the project. In Section 3.2, I present a young and unfinished implementation of Kroy's formalization of the categorical imperative. The finished version of my project will ideally include an implementation of Kroy's formalization of the second formulation of the categorical imperative as well. I also hope to write robust tests for both of these implementations to explore their limitations. These tests will help inform my eventual formalization of the categorical imperative.

The ultimate goal of the project is to present my own formalization of the categor-

ical imperative that escapes the limitations of the naive formalization and Kroy's formalization. This formalization will likely require some additional logical machinery to handle the complete notion of a maxim, including an agent, action, and end. My formalization will also patch up some of the holes in DDL itself that have been problematic for my project so far, such as the existence of contradictory obligations. I intend to formalize and implement all three formulations of the categorical imperative.

I will then test my formalization of the categorical imperative. I will create two kinds of tests. First, I will create metaethical tests that show logical properties independent of any model specification, as I did for the first two formalizations. Second, I will create tests that specify models and apply my formalization to real, concrete ethical dilemmas. This part of the project will seek to demonstrate the power and limitations of automated ethical reasoning. Questions to be explored here include: How much model specification is necessary to achieve ethical results? How should models be represented and specified? Does the automation of ethical reasoning provide anything, or is all the ethical work hidden in the model specification itself?

This final question is both technical and philosophical, and will be interesting to explore in the written component of my thesis. This question is related to Kant's distinction between analytic and synthetic reasoning Kant (1785). Analytic statements are true simply by virtue of their meaning, such as "All bachelors are unmarried." Synthetic reasoning involves some contribution by the reasoner, in the form of new insight or facts about the world. Kant presents the statements "All bachelors are alone" and "7+5=12" as examples of synthetic propositions. The analytic/synthetic distinction is hotly debated and has been refined significantly since Kant, and this area will require further research.

Kant believes that ethics is synthetic a priori reasoning, but it is unclear if auto-

mated theorem provers like Isabelle are capable of anything more than analytic reasoning. Many of the basic proof solving tools like `simp` or `blast` simply unfold definitions and apply axioms, and they appear to perform analytic reasoning. SMT solvers like `Nitpick` and `z3` (bundled with Isabelle) are candidates for synthetic reasoning. Model finding seems more sophisticated than the simple unfolding of definitions, but this requires further exploration.

Lastly, I hope to explore Kant's argument that the three formulations of the categorical imperative are equivalent. This hypothesis has been the subject of controversy, but many neo-Kantians believe that his claim is plausible, if not true. Armed with formalizations of each formulation, I will have all the tools necessary to test this hypothesis. I would like to either prove or disprove this hypothesis for my formalization, and analyze the philosophical implications of my result.

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. D. Ashley and B. M. McLaren. A cbr knowledge representation for practical ethics. In *Selected Papers from the Second European Workshop on Advances in Case-Based Reasoning*, EWCBR '94, page 181–197, Berlin, Heidelberg, 1994. Springer-Verlag. ISBN 3540603646.

A. Baltag and B. Renne. Dynamic Epistemic Logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2016 edition, 2016.

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

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

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.

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.

J. C. Blanchette, S. Böhme, and L. C. Paulson. Extending sledgehammer with smt solvers. In *Proceedings of the 23rd International Conference on Automated Deduction*, CADE'11, page 116–130, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642224379.

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.

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

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

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.

D. Fuenmayor and C. Benzmüller. Formalisation and evaluation of alan gewirth's proof for the principle of generic consistency in isabelle/hol. *Archive of Formal Proofs*, Oct. 2018. ISSN 2150-914x. https://isa-afp.org/entries/GewirthPGCProof.html, Formal proof development.

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

M. Guarini. Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, 21(4):22–28, 2006. doi: 10.1109/MIS.2006.76.

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

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.

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.

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

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

P. Kleingeld. Contradiction and kant's formula of universal law. *Kant-Studien*, 108 (1):89–115, 2017. doi: doi:10.1515/kant-2017-0006. URL https://doi.org/10.1515/kant-2017-0006.

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. Lin, K. Abney, and G. A. Bekey. *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, pages 35–52. 2012.

P. McNamara and F. Van De Putte. Deontic Logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2021 edition, 2021.

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.

P. Menzies and H. Beebee. Counterfactual Theories of Causation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.

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. *Bounds of Justice*. Cambridge University Press, December 2009.

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.

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.

I. Robeyns. The capability approach: a theoretical survey. *Journal of Human Development*, 6(1):93–117, 2005. doi: 10.1080/146498805200034266. URL https://doi.org/10.1080/146498805200034266.

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.

N. E. Snow. *The Oxford Handbook of Virtue*. Oxford University Press, 2017.

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.

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

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.

S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein. Implementations in machine ethics. *ACM Computing Surveys*, 53(6):1–38, Feb 2021. ISSN 1557-7341. doi: 10.1145/3419633. URL http://dx.doi.org/10.1145/3419633.

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

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