Automated Kantian Ethics: A Faithful Implementation and Testing Framework

A Senior Thesis presented

by

Lavanya Singh

to

The Departments of Computer Science and Philosophy

in partial fulfillment of the requirements

for the degree of

Bachelor of Arts with Honors

Harvard University

Cambridge, Massachussetts

May 2022

ABSTRACT

AI agents are beginning to make decisions without human supervision in increasingly consequential contexts like healthcare, policing, and driving. These decisions are inevitably ethically tinged, but most AI agents navigating the world today are not explicitly guided by ethics. Warnings from regulators, philosophers, and computer scientists about the dangers of unethical artificial intelligence, from science-fiction killer robots to criminal sentencing algorithms prejudiced against people of color, have spurred interest in automated ethics, or the development of machines that can perform ethical reasoning. Much prior work in automated ethics approaches the problem from a computational perspective and rarely engages with philosophical literature on ethics, despite its clear relevance to the development of AI agents that can responsibly navigate the world. All decisions are moral decisions, including those that AI agents are actively making today. If automated ethics draws on sophisticated philosophical literature, the ethical reasoning underlying such decisions will be more nuanced, precise, consistent, and trustworthy. However, faithfully translating complex ethical theories from natural language to the rigid syntax of a computer program poses technical and philosophical challenges.

In this thesis, I present an implementation of automated Kantian ethics that is faithful to the Kantian philosophical tradition. Of the three major ethical traditions, Kant's categorical imperative is the most natural to formalize because it is an inviolable formal rule that requires less context than other ethical theories. I formalize Kant's categorical imperative in Carmo and Jones's Dyadic Deontic Logic, implement this formalization in the Isabelle/HOL theorem prover, and develop a testing framework to evaluate how well my implementation coheres with expected properties of Kantian ethics, as established in the literature. I also use my system to reason about two ethical dilemmas used to criticize Kantian ethics: the difference between lying and joking and the example of a murderer knocking on your door asking about the location of their intended victim. Finally, I examine the philosophical implications of this system, exploring its limitations and its potential to help both AI agents and human beings better reason about ethics.

Armed with relatively uncontroversial facts about the world, my system is able to correctly resolve complex moral dilemmas because it is grounded in philosophical literature. Moreover, because I automate an explicit ethical theory, the ethical reasoning underlying my system's judgements is interpretable by a human being. I implement this ethical theory using the Isabelle/HOL interactive theorem prover, which can list the axioms and theorems used in a proof, so my system is explainable. This work serves as an early proof-of-concept for philosophically mature AI agents and is one step towards the development of responsible, trustworthy artificial intelligence.

Contents

1 Introduction

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

Machine ethicists recognize the need for automated ethics and have made both theoretical ((Awad et al., 2020), (Davenport, 2014), (Wallach and Allen, 2008), (Gabriel, 2020)) and practical progress ((Arkoudas et al., 2005), (Cervantes et al., 2013), (Jiang et al., 2021), (Winfield et al., 2014)) towards automating ethics. However, prior work in machine ethics using popular ethical theories like deontology ((Anderson and Anderson, 2014), (Anderson and Anderson)), consequentialism ((Abel et al., 2016), (Anderson et al., 2004), (Cloos, 2005)), and virtue ethics (Berberich and Diepold, 2018) rarely engages with philosophical literature and thus often misses philosophers' insights. Even the above example of the malfunctioning

self-driving car is an instance of Phillipa Foot's trolley problem, in which a bystander watching a runaway trolley can pull a lever to kill one instead of three (Foot, 1967). Decades of philosophical debate have developed ethical theories that can offer nuanced and consistent answers to the trolley problem. Like the trolley problem, the moral dilemmas that artifical agents face are not entirely new, so solutions to these problems should take advantage of philosophical progress. Philosophers are devoted to the creation of better ethical theories, so the more faithful that automated ethics is to philosophical literature, the more nuanced, precise, consistent, and therefore trustworthy it will be.

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

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

that begin with an ethical theory, instead of a computational tool, must contend with the fact that ethical theories are almost always described in natural language and must be made precise enough to represent to a computer. Even once ethics is translated from natural language to program syntax, the factual background given to the machine, such as the description of an ethical dilemma, is equally as important in determining the machine's decisions. Another complication is that philosophers do not agree that on a single choice of ethical theory. Philosophers who so agree that a specific ethical theory, like Kantian ethics, is true, still debate the theory's details.[1] Moreover, even once reasoning within a particular ethical theory is automated, those who disagree with that theory will disagree with the system's judgements.

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

1. In Section ??, I make a philosophical argument for why Kantian ethics is the most natural of the three major ethical traditions (deontology, virtue ethics, utilitarianism) to formalize.

2. In Section ??, I present a formalization of the practical contradiction interpretation of Kant's Formula of Universal Law in Dyadic Deontic Logic. I implement this formalization in the Isabelle/HOL theorem prover. My implementation includes axioms and definitions such that my system, when given an appropriately represented input, can

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

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

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

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

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

I choose to formalize Kant's moral rule in Carmo and Jones' Dyadic Deontic Logic (DDL) (Carmo and Jones, 2013). Deontic logic is a modal logic that can express obligation, or morally binding requirements. Traditional modal logics include the necessitation operator, denoted as $\Box$. In modal logic using the Kripke semantics, $\Box p$ is true at world w if p is true at all worlds that neighbor w (Cresswell and Hughes, 1996). Modal logics also contain the possibility operator $\Diamond$, where $\Diamond p \iff \neg(\Box(\neg p))$ and operators of propositional logic like $\neg, \wedge, \vee, \rightarrow$. I use DDL, in which the dyadic obligation operator $O\{A|B\}$ represents the sentence "A is obligated in the context B." The introduction of context allows DDL to express more nuanced reasoning. DDL is both deontic and modal, so sentences like $O\{A|B\}$ are terms that can be true or false at a world. For example, the sentence $O\{\text{steal}|\text{when rich}\}$ is true at a world if stealing when rich is obligated at that particular world.

I automate Kantian ethics because it is the most natural to formalize, as I argue in Section WhyKant. Kant presents three versions of a single moral rule, known as the categorical imperative, from which all moral judgements can be derived. I implement a version of this rule called the Formula of Universal Law (FUL), which states that people should only act on those principles that can be acted on by all people without contradiction. For example, in a world where everyone falsely promises to repay a loan, lenders will no longer believe these promises and will stop offering loans. Therefore, not everyone can simultaneously falsely promise to repay a loan, so the FUL thus prohibits this act.

Prior work by Benzmüller, Farjami, and Parent (Benzmüller et al., 2019; Benzmüller et al., 2021) implements DDL in Isabelle/HOL and I add the Formula of Universal Law as an axiom on top of their library. The resulting Isabelle theory can automatically or semi-automatically generate proofs in a new logic that has the categorical imperative as an axiom. Because proofs in this logic are derived from the categorical imperative, they judge actions as obligated, prohibited, or permissible. Moreover, because interactive theorem provers are designed to be interpretable, my system is explainable. Isabelle can list the axioms and facts it used to generate an ethical judgement, and, in some cases, construct human-readable proofs. In Sections Joking and Murderer, I use my system to arrive at sophisticated solutions to two ethical dilemmas often used in critiques of Kantian ethics. Because my system is faithful to philosophical literature, it is able to provide nuanced answers to these paradoxes that require a deep understanding of Kantian ethics.

In addition to presenting the above logic and implementation, I also contribute a testing framework that evaluates how well my formalization coheres with philosophical literature. I formalize expected properties of Kantian ethics as sentences in my logic, such as the property that obligations cannot contradict each other. I represent each of these properties as a sentence in my logic that my system should be able to prove or refute. I run the tests by using Isabelle to automatically find proofs or countermodels for the test statements. For example, my implementation passes the contradictory obligations test because it is able to prove the sentence $\neg(O\{A|B\} \wedge O\{\neg A|B\})$. I find that my system outperforms the control group of

raw DDL, without any moral axioms added, and Moshe Kroy's prior attempt at formalizing Kantian ethics in deontic logic (Kroy, 1976).

As it stands, my implementation can evaluate the moral status of sentences represented in my logic. Given an appropriate input, my project returns a value indicating if the action is obligatory (its negation violates the FUL), permissible (consistent with the FUL), or prohibited (violates the FUL) by proving or refuting a theorem in my logic.

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

2 System Components

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

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

2.1 Kantian Ethics

In this thesis, I automate Kantian ethics. In 2006, Powers posited that deontological theories are attractive candidates for automation because rules are generally computationally tractable (Powers, 2006, 1). Intuitively, algorithms are rules or procedures for problem solving and 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. Insofar as my project serves as an early proof-of-concept, 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.

2.1.1 Deontological Ethics

Deontological theories evaluate actions as permissible, obligatory, or prohibited and judge actions not on their consequences, but rather on "confirmity with a moral norm" (Alexander and Moore, 2021). In other words, deontological theories define a set of moral rules and evaluate actions using these rules. Deontologists believe that we should never violate any moral law. A wrong choice is wrong, regardless of its consequences.

Deontology is 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. 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.

Like all ethical traditions, deontology has debates that any implementation of automated deontological ethics must resolve. Deontologists disagree about whether ethics should focus on agents' actions or on the rights of those impacted by an action. Different deontological theories have different conceptions of what an action is, from the physical act itself to the agent's mental state at the time of acting to the principle upon which the agent acted.

While these debates are 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. 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.

2.1.2 Consequentialism

A consequentialist ethical theory is an ethical theory that evaluates an action by evaluating its consequences.[2] For example, utilitarianism is a form of consequentialism in which the moral action is the action that produces the most good (Driver, 2014). The focus on the consequences of action distinguishes consequentialists from deontologists, who derive the moral worth of an action from the action itself. Some debates in the consequentialist tradition include which consequences matter, what constitutes a "good" consequence, and how we can aggregate the consequences of an action over all the individuals involved.

Which Consequences Matter

Because consequentialism evaluates the state of affairs following an action, this kind of ethical reasoning requires more knowledge about the state of the world than deontology. Consequentialism requires knowledge about some or all consequences following an action. This requires that an automated ethical system somehow collect a subset of the infinite consequences of following an action, a difficult, if not impossible, task. Moreover, compiling this database of consequences requires answering difficult questions about which consequences

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

were actually caused [3] by an action and determining the state of the world before and after an action. As acts become more complex and affect more people, the computational time and space required to calculate and store their consequences increases. 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 challenge of representing the circumstances of action is not unique to consequentialism, but is particularly acute in this case. Kantian ethicists robustly debate which circumstances of an action are "morally relevant" when evaluating an action's moral worth.[4] Because deontology merely evaluates a single action, the surface of this debate is much smaller than debates about circumstances and consequences in a consequentialist system. An automated consequentialist system must make such judgements about the act itself, the circumstances in which it is performed, and the circumstances following the act. All ethical theories relativize their judgements to the situation in which an act is performed, but consequentialism requires far more knowledge about the world than deontology.

Theory of the Good

An automated consequentialist reasoner must also take a stance on the debate over what qualifies as a "good consequence," or what the theory of the good is. For example, hedonists associate good with the presence of pleasure and the absence of pain, while preference utiliarians believe that good is the satisfaction of desire. Other consequentialists, like Moore, adopt a pluralistic theory of value, under which many different kinds of things are good for different reasons (Moore, 1903).

Most theories of the good require that a moral reasoner understand complex features about individuals' preferences, desires, or sensations in order to evaluate a moral action, making automated consequentialist ethics difficult. Evaluating a state of affairs requires many

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

[4] Powers (2006) identifies this as a challenge for automating Kantian ethics and briefly sketches solutions from O'Neill (1990), Silber (1974), and Rawls (1980). For more on morally relevant circumstances, see Section WhatIsAMaxim.

controversial judgements about whether a state of affairs actually satisifes the relevant criteria for goodness. Perfect knowledge of tens of thousands of people's pleasure or preferences or welfare or rights is impossible. Either a human being assigns values to states of affairs, which doesn't scale, or the machine does, which requires massive common-sense and increases room for doubting the system's judgements. This may be a tractable problem, but it is much more difficult than the equivalent 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. The more complex the theory of the good, the more difficult this aggregation becomes. For example, pluralistic theories struggle to explain how different kinds of value can be compared (Sinnott-Armstrong, 2021). How do we compare one unit of beauty to one unit of pleasure? Resolving this debate requires that an automated reasoner choose one specific aggregation algorithm, but those who disagree with this choice will not trust the reasoner's moral judgements. Moreover, for complex theories of the good, this aggregation algorithm may be complex and may require a lot of data.

To solve this problem, some consequentialists reject aggregation entirely and instead prefer wholistic evaluations of a state of affairs. While this approach no longer requires that an aggregation algorithm, an automated ethical system still needs to calculate a goodness measurement for a state of affairs. Whereas before the system could restrict its analysis to a single person, the algorithm must now evaluate an entire state wholistically. As consequentialists modulate between aggregation and wholistic evaluation, they face a tradeoff between the difficulty of aggregation and the complexity of goodness measurements for large states of affairs.

Prior Attempts to Formalize Consequentialism

Because of its intuitive appeal, computer scientists have tried to formalize consequentialism in the past. These efforts cannot escape the debates outlined above. For example,

Abel et al. represent ethics as a Markov Decision Process (MDP), with reward functions customized to particular ethical dilemmas (Abel et al., 2016, 3). While this is a convenient representation, it either leaves unanswered or takes implicit stances on the debates above. It assumes that consequences can be aggregated just as reward is accumulated in an MDP.[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 consequentialism, 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 question. Other attempts to automate consequentialist ethics will suffer similar problems because, at some point, a usable automated consequentialist moral agent will need to resolve the above debates.

2.1.3 Virtue Ethics

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

Another difficulty with automating virtue ethics is that the unit of evaluation for virtue ethics is often a person's entire moral character. While deontologists evaluate the act itself, virtue ethicists evaluate the actor's moral character and their disposition towards the act. If states of affairs require complex representations, an agent's ethical character and disposition

[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.

are even more difficult to represent to a computer. This is more than just a data-collecting problem; it is a conceptual problem about the formal nature of moral character. Formalizing the concept of character appears to require significant philosophical and computational progress, whereas deontology immediately presents a formal rule to implement.

Prior Work in Machine Learning and Virtue Ethics

One potential appeal of virtue ethics is that many virtue ethical theories involve some notion of moral habit, which seems to be amenable to a machine learning approach. Artistotle, for example, argued that cultivating virtuous action requires making such action habitual through moral education (Aristotle, 1951). This implies that ethical behavior can be learned from some dataset of virtuous acts, either those prescribed by a moral teacher or those that a virtuous ideal agent would undertake. These theories seem to point towards a machine learning approach to automated ethics, in which ethics is learned from a dataset of acts tagged as virtuous or not virtuous.

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

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

ethics, but is one reason to prefer logical approaches.[7]

2.1.4 Kantian Ethics

In this paper I focus on Kantian ethics, a specific branch of deontology. Kant's theory is centered on practical reason, which is the kind of reason that we use to decide what to do and the source of our agency. In *The Groundwork of the Metaphysics of Morals*, Kant explains that rational beings are unique because we act "in accordance with the representations of laws" (Kant, 1785, 4:412). In contrast, a ball thrown into the air acts according to the laws of physics. It cannot ask itself, "Should I fall back to the ground?" It simply falls. A rational being, on the other hand, can ask, "Should I act on this reason?"" As Korsgaard describes it, when choosing which desire to act on, "it is as if there is something over and above all of your desires, something which is you, and which chooses which desire to act on" (Korsgaard and O'Neill, 1996, 100). Rational beings are set apart by this reflective capacity. We are purposive and our actions are guided by practical reason. We have reasons for acting, even when these reasons are opaque to us. This reflective choosing, or operation of practical reason, is what Kant calls the will.

The will operates by adopting, or willing, maxims, which are its perceived reasons for acting. Kant defines a maxim as the "subjective principle of willing," or the reason that the will *subjectively* gives to itself for acting (Kant, 1785, 16 footnote 1). Many philosophers agree that a maxim consists of some combination of circumstances, act, and goal.[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

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

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

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

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

Armed with this understanding of practical reason, Kant presents the categorical imperative. He presents three "formulations" or versions of the categorical imperative. In this project, I focus on the first formulation, the Formula of Universal Law, and I explain this choice in Section 2.1.5.

The Formula of Universal Law (FUL) states, "act only according to that maxim through which you can at the same time will that it become a universal law" (Kant, 1785, 34). This formulation generates the universalizability test, which we can use to test the moral worth of a maxim by imagining a world in which it becomes a universal law and attempting to will the maxim in that world. If there is a contradiction in willing the maxim in a world in which everyone universally wills the maxim, the maxim is prohibited.

Velleman presents a concise argument for the FUL. He argues that reason is universally shared among reasoners. For example, all reasoners have equal access to the arithmetic logic that shows that "2+2=4" (Velleman, 2005, 29). The reasoning that makes this statement true is not specific to any person, but is universal across people. Therefore, if I have sufficient

reason to will a maxim, so does every other rational agent. There is nothing special about the operation of my practical reason. In adopting a maxim, I implicitly state that all reasoners across time also have reason to adopt that maxim. Therefore, because I act on reasons, I must obey the FUL. Notice that this fulfills the above criterion for a categorical imperative: the FUL is derived from a property of practical reason itself and thus derives authority from the will's authority over itself.

Ease of Automation

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

Not only does evaluating Kantian ethics focus on a maxim, a maxim itself is an object with a thin representation for a computer, as compared to more complex objects like states of affairs or moral character. Later in my project, I argue that a maxim can be represented simply as a tuple of circumstances, act, and goal.[9] This representation is simple and efficient, especially when compared to the representation of a causal chain or a state of affairs or moral character. This property not only reduces the computational complexity (in terms of time and space) of representing a maxim, but it also makes the system easier for human reasoners to interact with. A person crafting an input to a Kantian automated agent needs to reason about relatively simple units of evaluation, as opposed to the more complex features that consequentialism and virtue ethics require.

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

Difficulties in Automation

One debate in Kantian ethics is the role of "common-sense" reasoning. Kantian ethics requires common-sense reasoning to determine which circumstances are "morally relevant" in the formulation of a maxim. Many misunderstandings in Kantian ethics are due to badly formulated maxims, so this question is important for an ethical reasoner to answer.[10] 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. Using my system in a fully automated moral agent will require answering this question, a challenging computational and philosophical task.

Common-sense reasoning is also relevant when applying the universalizability test itself. Consider the example maxim "When broke, I will falsely promise to repay a loan to get some quick cash." This maxim fails the universalizability test because in a world where everyone falsely promises to repay loans, no one will believe promises anymore, so the maxim will no longer serve its intended purpose (getting some quick cash). Making this judgement requires understanding enough about the system of promising to realize that it breaks down if everyone abuses it in this manner. This is a kind of common sense reasoning that an automated Kantian agent would need. This need is not unique to Kantian ethics; consequentialists agents need common sense to determine the consequences of an action and virtue ethical agents need common sense to determine which virtues an action reflects. Making any ethical judgement requires robust conceptions of the action at hand, falsely promising in this case. The advantage of Kantian ethics is that this is all the common sense that it requires, whereas a consequentialist or virtue ethical agent will require much more[11]. All moral theories evaluating falsely promising will a robust definition of promising, but consequentialism and virtue ethics will also require additional information that Kantian ethics will not. Thus, although

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

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

the need for common sense poses a challenge to automated Kantian ethics, this challenge is more acute for consequentialism or virtue ethics.

2.1.5 The Formula of Universal Law

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

The first formulation of the categorical imperative is the formula of universal law (FUL), which reads, "act only according to that maxim through which you can at the same time will that it become a universal law" (Kant, 1785, 34). The second formulation of the categorical imperative is the formula of humanity (FUH): "So act that you use humanity, in your own person, as well as in the person of any other, always at the same time as an end, never merely as a means." (Kant, 1785, 41). This formulation is often understood as requiring us to acknowledge and respect the dignity of every other person. The third formulation of the categorical imperative is the formula of autonomy (FOA), which Korsgaard describes as, "we should so act that we may think of ourselves as legislating universal laws through our maxims" (Korsgaard, 2012, 28). While closely related to the FUL, the FOA presents morality as the activity of perfectly rational agents in an ideal "kingdom of ends," guided by what Kant calls the "laws of freedom."

I choose to focus on the FUL[12], because it is the most formal and thus the easiest to formalize and implement. Onora O'Neill explains that the formalism of the FUL allows for greater precision in philosophical arguments analyzing its implications and power (O'Neill, 2013, 33). This precision is particularly useful in a computational context because any formalism necessarily makes its content precise. The FUL's precision reduces ambiguity, allowing me to remain faithful to philosophical literature on Kant. Precision reduces the need to make

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

choices to resolve debates and ambiguities. Minimizing these choices minimizes arbitrariness in my formalization and puts it on solid philosophical footing.

Though Kantians study all formulations of the categorical imperative, Kant argues in *Groundwork* that the three formulations of the categorical imperative are equivalent (Kant, 1785). While this argument is disputed Johnson and Cureton (2021), for those who believe it, the stakes for my choice of the FUL are greatly reduced. If all formulations are equivalent, then a formalization of the FUL lends the exact same power as a formalization of the second or third formulation of the categorical imperative.

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

2.2 Dyadic Deontic Logic

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

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

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

While SDL is appreciable for its simplicity, it suffers a variety of well-documented paradoxes, including contrary-to-duty paradoxes. [13] In situations where duty is violated, the logic breaks down and produces paradoxical results. Thus, I use an improved deontic logic instead of SDL for this work.

I use as my base logic Carmo and Jones's Dyadic Deontic Logic (DDL), which improves

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

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

These premises contradict themselves, because items (2)-(4) imply that Tom ought not help his neighbors. The contradiction results because the logic cannot handle violations of duty mixed with conditionals. (Chisholm, 1963; Rönnedal, 2019)

on SDL (Carmo and Jones, 2013). It introduces a dyadic obligation operator $O\{A|B\}$ to represent the sentence "A is obligated in the context B". The introduction of context allows DDL to gracefully handle contrary-to-duty conditionals by modifying the context. The obligation operator uses a neighborhood semantics, instead of the Kripke semantics (Scott, 1970; MONTAGUE, 1970). While Kripke semantics requires that an obligated proposition hold at all worlds, the neighborhood semantics defines a different set of neighbors, or morally relevant alternatives, for each world. To represent this, Carmo and Jones define a function ob that maps a given context (or world) to the propositions that are obligatory at this world, where a proposition p is defined as the worlds at which the p is true. DDL is thus both modal and deontic; statements about obligations are true or false at a world according to the neighbordhood semantics, and different obligations may hold at different worlds. This property is particularly relevant to my work because the universalizability test requires reasoning about alternative worlds, such as the world of the universalized maxim.

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

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

2.3 Isabelle/HOL

The final component of my project is the automated theorem prover I use to automate my formalization. Isabelle/HOL is an interactive proof assistant built on Haskell and Scala (Nipkow et al., 2002). It allows the user to define types, functions, definitions, and axiom systems. It has built-in support for both automatic and interactive/manual theorem proving. To demonstrate the power and usage of Isabelle and make DDL more precise, I walk through my reimplementation of Benzmueller, Farjami, and Parent's implementation of DDL in Isabelle/HOL (Benzmüller et al., 2021; Benzmüller et al., 2019).

2.3.1 System Definition

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

typedecl i — This is an Isabelle comment. i is the type for a set of worlds.
— This is a line of actual code used in my implementation. For the rest of the thesis, text typeset like this represents Isabelle code.

type-synonym $t = (i \Rightarrow bool)$ — t represents a set of DDL formulae.
— A set of formulae is defined by its truth value at a set of worlds. For example, the set $\{True\}$ is true at any set of worlds.

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

consts $ob::t \Rightarrow (t \Rightarrow bool)$ — set of propositions obligatory in this context
— $ob\ (context)\ (term)$ is *True* if the term is obligatory in this context

In a semantic embedding, axioms are modelled as restrictions on models of the system.

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

axiomatization where

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

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

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

2.3.2 Syntax

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

— Nitpick found a model for card i = 1:

Empty assignment

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

3 Implementation Details

In this section, I present the details of my implementation of automated Kantian ethics. I take a logic-programming approach in which I formalize the FUL in Dyadic Deontic Logic and then implement this logic in Isabelle/HOl. I also present my testing framework, which demonstrates one way to evaluate how faithful an implementation of automated ethics is to philosophical literature. This testing framework shows that my system outperforms two other possible implementations of automated Kantian Ethics.

3.1 Formalization and Implementation of the FUL

Before formalizing the FUL, I must define and implement the relevant logical background. Dyadic Deontic Logic can express obligation and prohibition, but it cannot represent more complex features of moral judgement like actions, subject, maxims, and ends. I augment DDL by adding representations of these concepts, drawn from philosophical literature.

3.1.1 Logical Background

Kantian ethics is centered on practical reason because it is action-guiding; the categorical imperative is a moral rule that agents can use to decide between potential actions. Thus, before I even begin to formalize a specific formulation of the categorical imperative, I must define the notions of subjects and actions. Concretely, I need to add logical background so that my logic can express sentences like, "x does action."

typedecl *s* — I declare a new type, *s* as the type for a "subject," i.e. the subject of a sentence.

The `typedecl` keyword indicates that I am defining a new atomic type, which is not composed of pre-existing types but is instead a new kind of thing altogether. I try to minimize the number of atomic types, so this will be one of the few new types that I define. To define a subject, it suffices to declare a new type to represent a subject, as that is all that is needed to apply the Formula of Universal Law. Notice that I have not defined any properties of this type,

such as the idea that a subject must be rational or human. Throughout my project, I will use bare syntactic units like types and constants to define new ideas and will add the minimum necessary definitions, or "thin" definitions, to achieve the desired results. By avoiding complex definitions of a subject, I can avoid murky philosophical debates about the nature of agency and reduce the potential for controversy or errors.

In this interpretation, the defining feature of a subject is that it can act, a relatively uncontroversial notion. I represent that below by allowing subjects to substitute into sentences, a property that I will use to represent the notion of different people performing certain actions.

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

3.1.2 Maxim

Recall from Section 2.1.5 that I formalize a version of the categorical imperative called the Formula of Universal, which reads "act only according to that maxim by which you can at the same time will that it should become a universal law" Kant (1785). In order to faithfully formalize the FUL, I must make precise the notions of of "willing a maxim" and violating the FUL. I draw on reliable definitions of willing, maxims, and the FUL from Kantian literature and represent them in DDL.

The central unit of evaluation for Kantian ethics is a "maxim," which Kant defines as "the subjective principle of willing," or the principle that the agent understands themselves as acting on Kant (1785). Modern Kantians differ in their interpretations of this definition. I adopt O'Neill's view, derived from Kant's example maxims, that a maxim includes the act,

the circumstances, and the agent's purpose of acting or goal (O'Neill, 2013).

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

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

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

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

O'Neill (O'Neill, 2013, 37) argues that 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" (O'Neill, 2013, 37). She also includes a purpose, end, or goal in the maxim because human activity, is guided by a rational will, and is thus inherently purposive (Kant, 1785, 4:428). A rational will does not act randomly (else it would not be rational), but instead in the pursuit of ends which it deems valuable[14]. The inclusion a maxim's end is essential for the version of the FUL that I will implement, explained in Section Practical Contradiction.

SHOULD THIS GO HERE OR IN THE LIMITATIONS SECTION OF THE DISCUSSION O'Neill's inclusion of circumstances is potentially controversial because it leaves open the question of what qualifies as a relevant circumstance for a particular maxim. This is gives rise to "the tailoring objection" (Kitcher, 2003, 217) [15], under which maxims are arbitrarily specified to pass the FUL. For example, the maxim "When my name is Jane Doe, I will

[14] Some argue that a maxim should also include the agent's motive or motivation and I address this concern in Appendix 4.1.

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

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 Jane Doe" is not morally relevant to the act and goal. This solution requires determining what qualifies as a relevant circumstance.

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

The difficulty in determining relevant circumstances is a limitation of my system and may require that a human being formulate the maxim or that future work develop heuristics to classify circumstances as morally relevant. For proponents of the "human-in-the-loop" model of AI ethics, in which ethical AI requires that humans guide machines, this kind of human involvement may be a feature (Lukowicz, 2019). In this model, a human being must create a representation for the dilemma they wish to test, translating a complex situation into a flat logical structure. This parallels the challenge that programmers face when translating the complexity of reality to a programming langauge or computational representation. 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.

Another solution to this problem may be to develop heuristics to classify circumstances as morally relevant. For example, one such attempt could define a moral closeness relation between an action, a goal, and circumstances. This heuristic would define morally relevant circumstance as those that reach a certain closeness threshhold with the action and the goal. Determining morally relevant circumstances, either using heuristics or human involvement,

is a ripe area for future work.

To will a maxim is to adopt it as a principle to live by, or to commit oneself to the action for the sake of the end in the relevant circumstances. Korsgaard argues 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). I formalize this idea in Definition 2.

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

$$will\ M\ s \equiv \forall w\ (C \longrightarrow A\ (s))\ w$$

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

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

abbreviation *will* :: *maxim* ⇒ *s*⇒ *t* (*W* - -)

where *will* ≡ λ(*c*, *a*, *g*) *s*. (*c* ⟶ (*a s*))

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

3.1.3 Practical Contradiction Interpretation

In order to actually evaluate the moral status of a maxim, I must precise the idea of failing the universalizability test or a non-universalizable maxim. Kantians debate the correct interpretation of the Formula of Universal Law because Kant himself appears to interpret the criterion in different ways. My project uses Korsgaard's practical contradiction interpretation, broadly accepted as correct within the philosophical community (Ebels-Duggan, 2012).

Recall that the Formula of Universal Law is to "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant, 1785). To determine if a maxim can be willed as a universal law, we can use the "universalizability

test", which requires imagining a world in which everyone has willed the maxim. If willing the maxim in such a world generates a contradiction, then the action is prohibited.

One interpretation of the FUL, the logical contradiction interpretation, prohibits maxims that are logically impossible when universalized. Under this view, falsely promising to repay a loan fails the universalizability test because, in the universalized world, the practice of promising would die out, so making a false promise would be impossible.

One problem with this view is that it cannot handle natural acts. Korsgaard appeals to Dietrichson (1964) to construct the example natural act of a mother killing her children that tend to cry at night so that she can get some sleep. Universalizing this maxim does not generate a logical contradiction, but it is clearly wrong. Because killing is a natural act, it can never be logically impossible so the logical contradiction view cannot prohibit it.

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

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

Using the practical contradiction interpretation, the FUL states, "If, when universalized, a maxim is not effective, then it is prohibited." This requires defining effectiveness and universalization. If an agent wills an effective maxim, then the maxim's goal is achieved, and

if the agent does not will it, then the goal is not achieved.

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

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

I can implement this in Isabelle using another abbreviation.

abbreviation *effective* :: *maxim⇒s⇒ t* (*E - -*)

where *effective* ≡ λ(*c*, *a*, *g*) *s*. ((*will* (*c*, *a*, *g*) *s*) ≡ *g*)

— A maxim is effective for a subject if the goal is achieved if and only if the subject wills the maxim. Once again, I use an abbreviation to conveniently refer to this Isabelle formula.

The former direction of the implication is intuitive: if the act results in the goal, it was effective in causing the goal. This represents sufficient causality. The latter direction represents necessary causality, or the idea that, counterfactually, if the maxim were not willed, then the goal would not be achieved (Lewis, 1973a).[16] 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.

abbreviation *universalized*::*maxim⇒t* **where**

universalized ≡ λ*M*. (λ*w*. (∀ *p*. (*W M p*) *w*))

abbreviation *holds*::*maxim⇒t* **where**

holds ≡ λ(*c*, *a*, *g*). *c*

abbreviation *not-universalizable* :: *maxim⇒s⇒bool* **where**

not-universalizable ≡ λ*M s*. ∀ *w*. ((*universalized M*) ⟶ (¬ (*E M s*))) *w*

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

— 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.[17]

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

abbreviation *permissible*::$maxim \Rightarrow s \Rightarrow t$
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**
$non\text{-}contradictory\ A\ B\ c\ w \equiv ((O\{A|c\} \wedge O\{B|c\})\ w) \longrightarrow \neg((A \wedge (B \wedge c))\ w \longrightarrow False)$
— Terms A and B are non contradictory in circumstances c if, when A and B are obligated in circum-

[17] In the rest of this section, for convenience, I will use the phrase "subject s willing maxim M is obligatory" interchangeably with "maxim M is obligatory for subject s." I will use "maxim M is obligatory" to refer to M being obligatory for any arbitrary subject, which I will show to be equivalent to M being obligatory for a specific subject.

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

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

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

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

FULo 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 ≡ $\lambda(c, a, g).\ \lambda s.\ \lambda w.\ (\neg\ (c \rightarrow g)\ w) \wedge (\neg\ (c \rightarrow a\ s)\ w)$

— 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 ≡ $\forall M$::*maxim*. $\forall s$::*s*. ($\forall w$. *well-formed M s w*) ⟶ (*not-universalizable M s* ⟶ ⊨ (*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 = ((λx._)(i_1 := True), (λx._)(s_1 := (λx._)(i_1 := False)), (λx._) (i_1 := False)) λ w. p = (λx._)(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 FULo 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 FULo 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 FULo 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 FULo, 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

4 Appendix

4.1 Maxims and Motives

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.

References

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

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

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

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

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

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

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

E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5):2332–2337, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1911517117. URL https://www.pnas.org/content/117/5/2332.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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