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, it will make the ethical reasoning underlying such decisions nuanced, precise and reliable. 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.

Armed with relatively uncontroversial facts about the world, my system is able to correctly resolve these moral dilemmas because it is grounded in philosophical literature. Moreover, because I automate an explicit ethical theory, the ethical reasoning underlying my sys-

tem'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. AI agents routinely make decisions with ethical implications without explicitly performing ethical reasoning and, in many cases, without human supervision. For example, the Alleghany Family Screening tool can automatically trigger an investigation into a potential case of child neglect, a decision that can uproot entire families and is known to be biased against poor people of color (Eubanks, 2018). Not only are machines making moral decisions without actually performing ethical reasoning, they're doing so without human involvement. 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 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. The trolley problem demonstrates that the moral dilemmas that artifical agents face are not entirely new, so solutions to these problems should take advantage of philosophical progress. The more faithful that automated ethics is to philosophical literature, the more reliable and nuanced 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. Even philosophers who agree that a specific ethical theory, like Kantian ethics, is true, debate the theory's details.[1] 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.

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, to help automated agents navigate the world, which I will refer to as automated ethics or machine ethics interchangeably. Second, to help human philosophers reason about philosophy, which I call computational ethics.

2 System Components

My system consists of three major 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 that undergirds my system, and explain the consequences of each of the three choices I make.

These specific components determine the features and limitations of the specific implementation of automated ethics that I present, but other choices of components, such as another ethical theory, a different logic, or a different theorem prover could be made. Given that my system is a proof-of-concept, I choose system components that are natural and intuitively fit together, but other choice could also be valid and perhaps even superior.

I do not claim that I have chosen the best ethical theory or the best logic or the best interactive theorem prover, and flaws with these components merely demonstrate that my specific choices were incorrect, but do not indict logic-programming-based automated ethics as I implement it in this thesis. 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. Future work could and should address the challenges I outline in this section. The more ethical theories that computational tools can handle, the more valuable computational philosophy becomes both for philosophers and for AI agents. Insofar as my project serves as an early proof-of-concept for computational ethics, I choose to automate an ethical theory that poses fewer challenges than others.

I first present deontological ethics, then consequentialism, and finally virtue ethics. For each tradition, I present a crash course for non-philosophers and then explain some obstacles to automation, arguing that these obstacles are weakest in the case of deontology. Finally, I will present the specific deontological theory I am automating (Kantian ethics) and will argue that it is comparatively easier to formalize. I will also outline the specific debates in the literature that my formalization takes a stance on and potential challenges for formalizing deontology.

2.1.1 Deontological Ethics

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

Deontology is immediately an attractive candidate for formalization because computers tend to understand rules; programming languages are designed to teach computers algorithms. Deontological ethical theories give inviolable rules that an automated agent can apply.

Moreover, because deontological theories focus on the action itself, they require relatively little data. A deontological moral judgement does not require as much information about context, consequences, or moral character as the other theories presented later in this section. All that matters is the action and some limited set of circumstances in which it is performed. I will later argue that, in the case of the specific deontological ethic that I implement (Kantian ethics), the action's representation is space efficient.

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. Many modern deontologists agree on Kant's ethic but disagree in how to interpret it. In this paper, I will formalize Kantian ethics. Deontology's comparatively greater focus on Kant means that the choice of Kant as a guiding figure will be less controversial to deontologists than, for example, the choice of Bentham as the guiding figure of consequentialism. Moreover, at the end of this section, I also argue that internal debates in the part of Kantian ethics that I focus on tend to be less controversial than those in the consequentialist or virtue ethical traditions.

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

[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.

moral worth of an action from the action itself. Some debates in the consequentialist tradition include which consequences of an action matter, what exactly constitutes a "good" consequence, and how we can aggregate the consequences of an action over all the individuals involved.

Which Consequences Matter

Because consequentialism evaluates the state of affairs following an action, this kind of ethical reasoning requires more knowledge about the state of the world than deontology. Consequentialism requires knowledge about some or all consequences following an action. This requires that an automated ethical system somehow collect a subset of the infinite consquences of following an action, a difficult, if not impossible, task. Moreover, compiling this database of consequences requires answering difficult questions about which consequences were actually caused by an action.[3] Evaluating an effect of an action requires knowledge about the state of the world before and after an action and knowledge about the action itself. Consequentialism requires knowledge about the situation in which the act is performed and following the act, whereas deontology mostly requires knowledge about the act itself. 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 the debate about circumstances and consequences in a consequentialist system. An automated consequentialist system must make such judgements about the act itself, the circumstances in

[3] David Hume argues that many straightforward accounts of causation face difficulties (Hume, 2007), and philosohpers 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 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.

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 individuals' desires. 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 of the above theories of good require that a moral reasoner understand complex features about individuals' preferences, desires, or sensations in order to evaluate a moral action, making automated consequentialist ethics difficult. Evaluating a state of affairs requires judgements about whether a state of affairs actually satisifes the relevant criteria for goodness. These judgements are controversian, and any consequentialist decision requires many of these judgements for each individual involved. As systems involve more people, making these judgements quickly becomes difficult, posing a scaling challenger. 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 is 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. For example, under a simple welfare model, each per-

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

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

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

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 useful 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 theories vary from Aristotle's virtues of courage and temperance Aristotle (1951) to 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 a virtue ethical theory is often a person's entire moral character. While deontologists evaluate the act itself and utilitarians evaluate the consequences of an act, virtue ethicists evaluate the actor's moral character and their disposition towards the act. Virtues are character traits and evaluating an action as virtuous or not requires understanding the agent's character and disposition while acting. If states of affairs require complex representations, an agent's ethical character and disposition are even more difficult to represent to a computer. Consequentialism posed a data-collection problem in evaluating and representing states of affairs, but virtue ethics poses a conceptual problem about the formal nature of moral character. Formalizing the concept of character appears to require significant philosophical and computational

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

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

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

Machine learning approaches like the Delphi system (Jiang et al., 2021) mentioned in 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 or correlations in datasets that don't actually cohere with the trends and causes that a human being would identify (Puiutta and Veith, 2020). While there is significant activity and progress in explainable machine learning, interactive theorem provers are designed to be explainable at the outset. Indeed, Isabelle can show the axioms and lemmas it used in constructing a proof, allowing a human being to reconstruct the proof independently if they wish. This is not an intractable problem for machine learning approaches to computational ethics, but is one reason to prefer logical approaches.[7]

[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

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

The will operates by adopting, or willing, maxims, which are its perceived reasons for acting. Kant defines a maxim as the "subjective principle of willing," or the reason that the will *subjectively* gives to itself for acting (Kant, 1785, 16 footnote 1). There is debate about what exactly must be included in a maxim, but many philosophers agree that a maxim consists of some combination of circumstances, act, and goal.[8] One example of a maxim is "When I am hungry, I will eat a doughnut in order to satisfy my sweet tooth." When an agent wills this maxim, they decide to act on it. They commit themselves to the end in the maxim (e.g. satisfying your sweet tooth). They represent their action, to themselves, as following the principle given by this maxim. Because a maxim captures an agent's principle of action, Kant evaluates maxims as obligatory, prohibited, or permissible. He argues that certain maxims have a form or logical structure that requires any rational agent to will them, and these maxims are obligatory.

Related Work.

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

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

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

The first formulation of the categorical imperative is the Formula of Universal Law (FUL), which reads, "act only according to that maxim through which you can at the same time will that it become a universal law" (Kant, 1785, 34). This formulation generates the universalizability test, which tests the moral value of a maxim by imagining a world in which it becomes a universal law and attempting to will the maxim in that world. If there is a contradiction in willing the maxim in a world in which everyone universally wills the maxim, the maxim is prohibited. Velleman presents a concise argument for the FUL. He argues that reason is universally shared among reasoners. For example, all reasoners have equal access to the arithmetic logic that shows that "2+2=4" (Velleman, 2005, 29). The chain of reasoning that

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

makes this statement true is not specific to any person, but is universal across people. Therefore, if I have sufficient reason to will a maxim, so does every other rational agent. There is nothing special about the operation of my practical reason that other reasoners don't have access to. Practical reason is shared, so in adopting a maxim, I implicitly state that all reasoners across time also have reason to adopt that maxim. Therefore, because I act on reasons, I must obey the FUL. Notice that this fulfills the above criterion for a categorical imperative: the FUL is derived from a property of practical reason itself and thus derives authority from the will's authority over itself, as opposed to some external authority.

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

Ease of Automation

Kantian ethics is an especially candidate for formalization because the categorical imperative, particularly the FUL, is a property of reason related to the form or structure of a maxim, or a formal principle of practical reason. It does not require any situational knowledge or contingent beyond the circumstances included in the maxim itself and thus requires far less contingent facts than other ethical theories. Instead, it is purely a property of the proposed principle for action. This formalism makes Kantian ethics an attractive candidate for formalization. While other ethical theories often rely on many facts about the world or the actor, Kantian ethics simply relies on the form of a given maxim. A computer evaluating a maxim doesn't require any knowledge about the world beyond what is contained in a maxim. A maxim is the only input that the computer needs to make a moral judgement. Automating Kantian ethics merely requires making the notion of a maxim precise and representing it to the computer. This distinguishes Kantian ethics from consequentialism and virtue ethics, which, as I argued above, require far more knowledge about the world or the agent to reach a moral decision.

Not only does evaluating Kantian ethics focus on a maxim, a maxim itself is an object with a thin representation for a computer, as compares to more complex objects like states of affairs or moral character. Later in my project, I argue that a maxim can be represented simply as a tuple of circumstances, act, and goal.[10] This representation is simple and efficient, especially when compared to the representation of a causal chain or a state of affairs or moral character. A maxim is a principle with a well-defined form, so representing a maxim to the computer merely requires capturing this form. This property not only reduces the computational complexity (in terms of time and space) of representing a maxim, it also make the system easier for human reasoners to interact with. A person crafting an input to a Kantian automated agent needs to reason about relatively simple units of evaluation, as opposed to the more complex features that consequentialism and virtue ethics require. I will make the comparision to consequentialism and virtue ethics explicit below.

Difficulties in Automation

My choices to interpret maxims and the Formula of Universal Law in a particular way represent debates in Kantian ethics over the meanings of these terms that I take a stance on. Another debate in Kantian ethics is the role of "common-sense" reasoning. Kantian ethics requires common-sense reasoning to determine which circumstances are "morally relevant" in the formulation of a maxim. Many misunderstandings in Kantian ethics are due to badly formulated maxims, so this question is important for an ethical reasoner to answer. My system does not need to answer this question because I assume a well-formed maxim as input and apply the categorical imperative to this input, but if my system were ever to be used in a faully automated agent, answering this question would require significant computational and philosophical work. For more, see Section AI Ethics.

Common-sense reasoning is also relevant in applying the universalizability test itself. Consider an example maxim tested using the Formula of Universal Law: "When broke, I will falsely promise to repay a loan to get some quick cash." This maxim fails the universalizability test because in a world where everyone falsely promises to repay loans, no one will believe

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

promises anymore, so the maxim will no longer serve its intended purpose (getting some quick cash). Making this judgement requires understanding enough about the system of promising to realize that it breaks down if everyone abuses it in this manner. This is a kind of common sense reasoning that an automated Kantian agent would need. This need is not unique to Kantian ethics; consequentialists agents need this kind of common sense to determine the consequences of an action and virtue ethical agents need this kind of common sense to determine which virtues an action reflects. Making any ethical judgement requires relatively robust conceptions of the action or situation at hand, falsely promising in this case. The advantage of Kantian ethics is that this is all the common sense that it requires, whereas a consequentialist or virtue ethical agent will require much more. All moral theories evaluating falsely promising will a robust definition of the convention of promising, but consequentialism and virtue ethics will also require additional information about consequences or character that Kantian ethics will not. Thus, although the need for common sense poses a challenge to automated Kantian ethics, this challenge is more acute for consequentialism or virtue ethics so Kantian ethics remains within the closest reach of automation.

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

A deontic logic is a special kind of modal logic designed to reason about obligation. Standard deontic logic (Cresswell and Hughes, 1996; McNamara and Van De Putte, 2021) replaces $\Box$ with the obligation operator O, and $\diamond$ with the permissibility operator P. Using

the Kripke semantics for O, Op is true at w if p is true at all ideal deontic alternatives to w. The O operator in SDL takes a single argument (the formula that is obligatory), and is thus called a monadic deontic operator.

While SDL is appreciable for its simplicity, it suffers a variety of well-documented paradoxes, including contrary-to-duty paradoxes [11]. In situations where duty is 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, or DDL, which improves on SDL Carmo and Jones (2013). It introduces a dyadic obligation operator $O\{A|B\}$ to represent the sentence "A is obligated in the context B". This gracefully handles contrary-to-duty conditionals. The obligation operator uses a neighborhood semantics Scott (1970); MONTAGUE (1970), instead of the Kripke semantics. Carmo and Jones define a function *ob* that maps from worlds to sets of sets of worlds. Intuitively, each world is mapped to the set of propositions obligated at that world, where a proposition p is defined as the worlds at which the p is true.

DDL also includes other modal operators. In addition to $\Box$ and $\Diamond$, DDL also has a notion of actual obligation and possible obligation, represented by operators O_a and O_p respectively. These notions are accompanied by the corresponding modal operators $\Box_a, \Diamond_a, \Box_p, \Diamond_p$. These operators use a Kripke semantics, with the functions av and pv mapping a world w to the set of corresponding actual or possible versions of w.

For more of fine-grained properties of DDL see (Carmo and Jones, 2013) or this project's source code. DDL is a heavy logic and contains modal operators that aren't necessary for my

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

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

analysis. While this expressivity is powerful, it may also cause performance impacts. DDL has a large set of axioms involving quantification over complex higher-order logical expressions. Proofs involving these axioms will 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.

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

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

2.3.1 System Definition

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

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

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

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

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

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

2.3.2 Axiomatization

For a semantic embedding, axioms are modelled as restrictions on models of the system. In this case, a model is specificied by the relevant accessibility relations, so it suffices to place conditions on the accessibility relations. These axioms can be quite unweildy, so luckily I was able to lift BFP's implementation of Carmo and Jones's original axioms directly (Benzmüller et al., 2021). Here's an example of an axiom:

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

— 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-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.

2.3.3 Syntax

The syntax that I will work with is defined as abbreviations. Each DDL operator is represented as a HOL formula. Isabelle automatically unfolds formulae defined with the `abbreviation` command whenever they are applied. While the shallow embedding is performant (because it uses Isabelle's original syntax tree), abbreviations may hurt performance. In some complicated proofs, we want to control definition unfolding. Benzmueller and Parent told me that the performance cost of abbreviations can be mitigated using a definition instead.

Modal operators will be useful for my purposes, but the implementation is pretty stan-

dard.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Validity will be useful when discussing metalogical/ethical properties.

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

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

2.3.4 Syntactic Properties

One way to show that a semantic embedding is complete is to show that the syntactic specification of the theory (axioms) are valid for this semantics - so to show that every axiom holds at every world. Benzmueller, Farjami, and Parent provide a complete treatment of the com-

pleteness of their embedding, but I will include selected axioms that are particularly interesting here. This section also demonstrates many of the relevant features of Isabelle/HOL for my project.

Consistency

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

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

— Nitpick found a model for card i = 1:

Empty assignment

Nitpick Blanchette and Nipkow (2010) can generate models or countermodels, so it's useful to falsify potential theorems, as well as to show consistency. by simp indicates the proof method. In this case, simp indicates the Simplification proof method, which involves unfolding definitions and applying theorems directly. HOL has $True$ as a theorem, which is why this theorem was so easy to prove.

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.

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.

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.

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.

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

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.

P. McNamara and F. Van De Putte. Deontic Logic. In E. N. Zalta, editor, *The Stanford*

Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, spring 2021 edition, 2021.

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

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.

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.