ABSTRACT

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

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

1 Introduction

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

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

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

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

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

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

Contributions

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

1. In Section 2.1, I make a philosophical argument for why Kantian ethics is the most natural of the three major ethical traditions (deontology, virtue ethics, utilitarianism) to formalize.

2. In Section 3.1, I present a formalization of the practical contradiction interpretation of Kant's Formula of Universal Law in Dyadic Deontic Logic. I implement this formalization in the Isabelle/HOL theorem prover. My implementation includes axioms and

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

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

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

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

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

Automated Kantian Ethics

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

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

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

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

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

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

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

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

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