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 have no notion of 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 automated ethics, or the development of machines that can perform ethical reasoning. Previous work in automated ethics rarely engages with existing philosophical literature. Given that ethics is the study of how to navigate the world, automated ethics should look to philosophical literature to guide the development of AI agents that can responsibly navigate the world. All decisions are moral decisions, including the ones that AI agents are actively making today. Automated ethics that draws on sophisticated philosophical literature makes the ethical reasoning underlying such decisions nuanced, precise and reliable, but 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 case of joking and the example of a murderer knocking on someone's door asking about the location of their intended victim. Armed with relatively uncontroversial facts about the world, my system correctly resolves these moral dilemmas.

My system is able to resolve complex ethical dilemmas because it is grounded in philo-

sophical 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

The development of autonomous artificial agents has spurred interest in computers that can perform ethical reasoning, known as automated moral agents. AI agents are making decisions in increasingly consequential contexts, such as medical diagnoses, driving, and criminal sentencing, and therefore must perform ethical reasoning in order to navigate moral dilemmas responsibly. 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. Because ethics is the study of navigating the world responsibly, AI agents routinely make moral decisions without explicitly performing ethical reasoning. Moreover, artifical agents are making many such moral decisions without human supervision, such as hiring algorithms that filter job applicants' resumes, a decision that can impact people's livlihoods and often involves implicit prejudices. Machine ethics, also called automated ethics, is 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. The above example of the malfunctioning self-driving car is an instance of Phillipa Foot's trolley problem (Foot, 1967), in which a bystander watching a

runaway trolley can pull a lever to kill one instead of three. 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 AI faces 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, as seen in the example of Delphi, which 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 often 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. 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, acting on this judgement becomes difficult. 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 (**?**) 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). Ethical theories are almost always described in natural language, so automated ethics must first make ethics 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 on a single "correct" 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 ethics according to 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 the 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 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

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

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 generally evaluate how faithful an implementation of automated Kantian ethics is. My framework includes meta-ethical tests and application tests inspired by philosophical literature. My testing framework shows that my formalization substantially improves on prior work and can be generalized to 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, it can also help philosophers make philosophical progress.

I present a faithful implementation of automated Kantian ethics, a testing framework to evaluate how well my implementation coheres with expected properties of Kantian ethics (derived from philosophical literature), and examples in which my system performs sophisticated moral reasoning. My system consists of a logic in which I formalize an ethical theory and an implementation of this formalization in an interactive theorem prover.

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 $O\{A|B\}$ to represent the sentence "A is obligated in the context B." The introduction of context allows DDL to reason about violations of duty. 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 raw DDL 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. To this end, I present 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.

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.

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.

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. 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.

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.

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.

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.

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.

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

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

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.