Experimenting with Carmo and Jones' DDL

Lavanya Singh

April 2, 2021

Contents

1 Introduction

Leibniz dreamed of modelling all knowledge in the language of formal logic, so that all reasoning could be automated. As machines become increasingly capable of mathematical, strategic, and scientific reasoning and Leibniz's dream becomes closer to reality, one key gap remains. Machines lack the ability to perform any meaningful kind of ethical reasoning. Computational ethics is a young but attractive field for two primary reasons. First, the proliferation of artifically intelligence and autonomous agents is creating and will continue to create a demand for ethical autonomous agents. The call for "ethical AI" can, in one sense, be answered by the creation of an automated ethical reasoner. Second, just as automated mathematical reasoning allows mathematicians to explore new proofs, automated ethical reasoning is a tool that philosophers can use when reasoning about ethics. Many contradictions or paradoxes with an ethical system may not be immediately obvious to the human eye, but can be easily tested using an automated theorem prover.

Modelling ethics without sacrificing the intracacies and complexities of an ethical theory is a challenging computational and philosophical problem. Many ethical AI systems encode ethics as a series of constraints, and they maximize their objective with respect to such constraints. This approach to ethics fails to capture much of the complexity of any plausible ethical system. Any faithful model of an ethical theory will require machinery more complex than a single constraint satisfaction problem.

In addition to the computational machinery, computational ethics also requires a sophisticated ethical theory to model. Constraint satisfaction systems often default to some version of utiliarianism, the principle of doing the most good for the most people. Alternatively, they model basic moral principles such as "do not kill," without modelling the theory that these principles originated from. Modelling a more complex ethical theory will not only enable smarter philosophical machines, it will also empower philosophers to study more complex ethical issues with the computer's help.

The ideal candidate ethical theory will be both philosophically interesting and robust and easy to formalize. Kantian ethics, often described as "formal," is such a candidate. The categorical imperative, Kant's universal law of morality, is a moral rule that can be used to guide action in all spheres of life. Kant's original presentation itself is methodical and formal, and the theory lends itself well to formalization.

In this paper, I present, implement, and test three formalizations of Kant's categorical imperative in the Isabelle/HOL theorem prover. I start with Carmo and Jones' Dyadic Deontic Logic as a base logic and model each formalization as an extension of DDL. Section 4.3 implements and tests the naive implementation, a control group that is not much stronger than DDL itself. Section 4.4 examines a more sophisticated implementation inspired

by Moshe Kroy's partial formalization of the categorical imperative. Section 4.5 explores ????

I contribute implementations of three different interpretations of the categorical imperative, examples of how each implementation can be used to model and solve ethical scenarios, and tests that examine ethical and logical properties of the system, including logical consistency, consistentency of obligation, and possibility of permissibility. The implementations themselves are usable models of ethical principles and the tests represent the kind of philosophical work that formalized ethics can contribute.

2 The Problem

3 My Idea

4 Details

4.1 Naive Implementation/Control Group

4.2 Kroy's Partial Formalization

5 Related Work

6 Conclusion

end