A Faithful Implementation of Automated Kantian Ethics

Abstract: Warnings from regulators, philosophers, and computer scientists about the dangers of unethical artificial intelligence have spurred interest in the development of machines that can perform ethical reasoning. However, previous work in automated ethics rarely engages with existing philosophical literature. Philosophically sophisticated ethical theories are necesssary for nuanced and reliable judgements, but faithfully translating these complex ethical theories from natural language to the rigid syntax of a computer program poses technical and philosophical challenges. In this paper, 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, context-agnostic, formal rule. 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. My system is an early step towards philosophically mature ethical AI agents and it can make nuanced judgements in complex ethical dilemmas because it is grounded in philosophical literature. Moreover, because my system uses an interactive theorem prover, its judgements are explainable.

Area: Algorithm Development
Keywords: AI ethics, interactive theorem provers, deontic logic, Kant