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. Despite the fact that ethical philosophers are devoted to developing robust ethical theories, previous work in automated ethics rarely engages with existing philosophical literature. Faithfully translating complex, philosophically sophisticated ethical theories expressed in 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 not only an early step towards philosophically mature ethical AI agents, but it can also help philosophers reach new ethical insights, thus paving the way for the use of computational tools to address philosophical problems.

Areas: algorithm development, applications, philosophy, ethics and automation