FAKE: Faithful Automated Kantian Ethics

Abstract: Warnings from regulators, philosophers, and computer scientists about the dangers of unethical artifical intelligence have spurred interest in automated ethics. Machine ethics will best cohere with intuitions and guide action when it is faithful to existing philosophical literature. Translating complex ethical theories expressed in natural language to the rigid syntax of a computer program poses technical and philosophical challenges. In this paper, I present FAKE: 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 presents inviolable, context-agnostic, formal rules. 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. FAKE 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 computational ethics.

Areas: algorithm development, applications, philosophy