CAKE: Computational and Automated Kantian Ethics

Abstract: As artifically intelligent agents become increasingly powerful, the need for philosophically sophisticated automated ethical reasoning becomes more pressing. Representing a robust ethical theory to a computer poses technical and ethical challenges. I argue that, out of the three major ethical traditions, Kantian ethics is most natural to formalize because it presents inviolable, context-agnostic, formal rules. In this paper, I present CAKE, an implementation of automated Kantian ethics that is faithful to the rich Kantian philosophical tradition. I formalize Kant's categorical imperative in 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. Not only is CAKE an early step towards philosophically rich ethical AI agents, it also establishes that computational tools can help philosophers reach ethical insights.

Areas: algorithm development, applications, philosophy