Passing the Moral Turing Test

Julia Haas
Rhodes College, Memphis, Tennessee, United States

Abstract

The translation problem in moral AI asks how insights into human norms and values can be translated into a form suitable for implementation in artificial systems. I argue that if my answer to a question about the human mind is right, then the translation problem is more tractable than previously thought. Specifically, I argue that we can use principles from reinforcement learning to study human moral cognition, and that we can use principles from the resulting evaluative moral psychology to design artificial systems capable of passing the Moral Turing Test (Allen, 2000). I illustrate the core features of my proposal by describing one such environment, or gridworld, in which an agent learns to trade-off between monetary profit and fair dealing, as characterized in behavioral economic paradigms. I conclude by highlighting the core technical and philosophical advantages of such an approach for modeling moral cognition more broadly construed.