When it's not out of line to get out of line: Principles of universalizability, welfare, and harm

Joseph Kwon
MIT , Cambridge, Massachusetts, United States

Tan Zhi-Xuan
Massachusetts Institute of Technology , Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Sydney Levine
Allen Institute for Artificial Intelligence, Seattle, Washington, United States

Abstract

How do we know when it's OK to break moral rules? We propose that — alongside well-studied outcome-based measures of welfare and harm — people sometimes use universalization, asking "What if everyone felt at liberty to ignore the rule?" We develop a virtual environment where agents stand in line to gather water. Subjects judge agents who get out of line to try to get water more quickly. If subjects use universalization, they would need to imagine all agents getting out of line and going straight for the water in each environment. To test this prediction, we model an action's universalizability by simulating what would happen if every agent tried to follow a path directly to the water, then evaluating the effects. We also investigate the role of several outcome-based measures, including welfare aggregation and harm-based measures. We find that universalizability plays a important role in rule-breaking judgments alongside outcome-based concerns.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.