Moral Dynamics: A Computational Model of Moral Judgment

Felix Sosa
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Tomer Ullman
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Samuel Gershman
Harvard University, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Tobias Gerstenberg
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

Previous work on morality has proposed psychophysical and/or qualitative models for moral judgment. While these models capture the data found in their respective studies, we believe they miss the underlying concepts on which people base their moral judgments. Here, we propose a quantitative model of morality grounded in our current understanding of intuitive theories of physics, psychology, and causality.

We detail how peoples intuitions of physics and causality can be used to infer the desire and intent of an agent to bring about or prevent harm and how this process can qualitatively predict empirical findings of previous work on moral judgment and quantitatively predict results in new scenarios involving an agent harming or helping another.