UCI Center for Theoretical Behavioral Sciences

The Center for Theoretical Behavioral Sciences' Seminar Series presents:

David Thorstad, Ph.D.
Assistant Professor
Department of Philosophy
Vanderbilt University

Cognitive bias in large language models: Cautious optimism meets anti-Panglossian meliorism

Traditional discussions of bias in large language models focus on a conception of bias closely tied to unfairness especially as affecting marginalized groups. Recent work raises the novel possibility of assessing the outputs of large language models for a range of cognitive biases familiar from research in judgment and decisionmaking. My aim in this paper is to draw two lessons from recent discussions of cognitive bias in large language models: cautious optimism about the prevalence of bias in current models coupled with an anti-Panglossian willingness to concede the existence of some genuine biases and work to reduce them. I draw out philosophical implications of this discussion for the rationality of human cognitive biases as well as the role of unrepresentative data in driving model biases.

THURSDAY, October 10, 2024 | 4:00 - 5:00 PM

Social Science Plaza A, Room 2112