Does moral advice from Large Language Models make people more prosocial toward outgroups?

Sonia Teo
University of California, Los Angeles

Falk Lieder
University of California, Los Angeles

Abstract

People increasingly ask large language models (LLMs) for moral advice, yet little is known about how such advice affects people's decisions. We investigated whether interacting with ChatGPT influences outgroup prosociality and decision extremity in us-vs-them dilemmas. In a preregistered between-subjects experiment (N = 460), participants either discussed a moral dilemma with GPT-5 or engaged in a control conversation about an unrelated neutral dilemma. Although GPT-5's decisions were significantly more prosocial toward outgroups than participants' decisions, its advice had no net effect on the prosociality or extremity of people's decisions. Nevertheless, consulting GPT-5 made participants significantly more confident about their final decisions. Exploratory analyses further suggested that GPT-5's advice had a convergence effect: participants whose initial decisions were either substantially more or substantially less prosocial than GPT-5's moved toward GPT-5's position. These findings suggest that LLM moral advice may primarily increase users' confidence without reliably increasing prosociality.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*