Do Large Language Models Have a Planning Theory of Mind? Evidence from MindGames: a Multi-Step Persuasion Task

Jared Moore
Stanford University, Stanford, California, United States

Rasmus Overmark
University of St. Andrews, St. Andrews, United Kingdom

Ned Cooper
Australian National University, Canberra, New South Wales, Australia

Beba Cibralic
University of Cambridge, Cambridge, United Kingdom

Nick Haber
Stanford, Stanford, California, United States

Cameron Jones
UC San Diego, La Jolla, California, United States

Abstract

Recent evidence suggests that Large Language Models (LLMs) display Theory of Mind (ToM) abilities. However, experiments with LLMs typically assess only *spectatorial* ToM, where LLMs merely predict other agents' behavior, rather than *planning*. In contrast, ToM in humans also contributes to dynamically *planning action* and *intervening* on others' mental states. We present a novel task of such a 'planning theory of mind' (PToM), which requires agents to infer an interlocutor's beliefs and desires and persuade them to alter their behavior. We find that humans significantly outperform o1 (an LLM) at our task, even though o1 outperforms humans in a baseline condition which requires minimal mental state inferences. The results suggest that LLM performance at other ToM tasks may be attributable to simpler predictive abilities, while people excel at counterfactual planning when reasoning about others' behavior. Our paper is here: https://jaredmoore.org/mindgames

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*