Language Models as Informative Goal Priors in a Bayesian Theory of Mind

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

Paul Stefan Lunis
University of Central Florida, Orlando, Florida, United States

Nathalie Fernandez Echeverri
University of Maryland, College Park, Maryland, United States

Vikash Mansinghka
MIT, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

Bayesian models of theory of mind (ToM) have been successful in explaining how humans infer goals from the actions of other agents. However, they have typically been limited to small and fixed sets of possible goals specified in advance by the modeler, leaving open the question of how spontaneous goal inference occurs in rich and complex environments. To address this question, we posit that people are guided by informative, context-specific goal priors and proposals when making inferences about others. As proxies for these informed priors, we make use of context-conditioned large language models (LLMs) and integrate them into a Bayesian inverse planning framework. We find that LLMs can serve as usefully informative priors and proposals over goals compared to a structural baseline prior, allowing them to be used as models of the learned statistical knowledge that humans bring to bear in their inferences about others' goals.

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