Belief Attribution as Mental Explanation: The Role of Accuracy, Informativity, and Causality

Lance Ying
Massachusetts Institute of Technology , Cambridge, Massachusetts, United States

Almog Hilel
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Ryan Truong
University of Texas at Austin, Austin, Texas, United States

Vikash Mansinghka
MIT, Cambridge, Massachusetts, United States

Joshua Tenenbaum
MIT, Cambridge, Massachusetts, United States

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

Abstract

A key feature of human theory-of-mind is the ability to attribute beliefs to other agents as mentalistic explanations for their behavior. But given the wide variety of beliefs that agents may hold about the world and the rich language we can use to express them, which specific beliefs are people inclined to attribute to others? In this paper, we investigate the hypothesis that people prefer to attribute beliefs that are good explanations for the behavior they observe. We develop a computational model that quantifies the explanatory strength of a (natural language) statement about an agent's beliefs via three factors: accuracy, informativity, and causal relevance to actions, each of which can be computed from a probabilistic generative model of belief-driven behavior. Using this model, we study the role of each factor in how people selectively attribute beliefs to other agents. We investigate this via an experiment where participants watch an agent collect keys hidden in boxes in order to reach a goal, then rank a set of statements describing the agent's beliefs about the boxes' contents. We find that accuracy and informativity perform reasonably well at predicting these rankings when combined, but that causal relevance is the single factor that best explains participants' responses.

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