Thinking about Thinking as Rational Computation

Marlene Berke
Yale University, New Haven, Connecticut, United States

Abi Tenenbaum
Yale University, New Haven, Connecticut, United States

Ben Sterling
Yale University, New Haven, Connecticut, United States

Julian Jara-Ettinger
Yale University, New Haven, Connecticut, United States

Abstract

Theory of Mind enables us to attribute mental states to others. But we not only make inferences about mental states (like what someone believes or wants), but about mental processes (like if someone is distracted or whether they remember something). Here, we present a computational formalization of these kinds of inferences. We propose that inferences about mental processes are structured around a principle of rational mental effort: the expectation that other people allocate mental resources rationally so as to minimize thinking costs incurred while pursuing their goals. We develop this theory into a computational model in the context of the Rush Hour puzzle game. In two behavioral experiments testing different inferences about mental processing, we find that our model predicts participant judgments. This work advances our understanding of the richness of the human mind's ability to think about other minds, and even about thinking itself.

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