Human-like learning in reasoning models: behavioral and neuroimaging evidence

Csaba Botos
University of Oxford, Oxford, Oxfordshire, United Kingdom

Sreejan Kumar
Columbia University, New York, New York, United States

Austin Andrews
University of Oxford, Oxfordshire, Oxford, United Kingdom

Laurence Hunt
University of Oxford, Oxford, United Kingdom

Christopher Summerfield
Oxford University, Oxford, United Kingdom

Rui Ponte Costa
University of Oxford, Oxford, United Kingdom

Joshua Tenenbaum
MIT, Cambridge, Massachusetts, United States

Marcelo Mattar
New York University, New York, New York, United States

Momchil Tomov
Harvard University, Cambridge, Massachusetts, United States

Abstract

Humans rapidly learn abstract knowledge when encountering novel environments and flexibly deploy this knowledge to guide efficient and intelligent action. Can modern AI systems learn and plan in a similar way? Leveraging a unique dataset of human gameplay with concurrent fMRI recordings, we evaluate model-free reinforcement learning agents, bayes-optimal model-based agents, and a frontier Large Reasoning Model, on two complementary dimensions: behavioral patterns and predictivity of human brain representations. Using encoding models, we assess how well each system's internal representations predict brain activity in regions previously implicated in theory-based reasoning. We find that the Large Reasoning Model most closely matches human behavioral patterns during game discovery and predicts brain activity in theory-coding regions an order of magnitude better than both model-free and model-based alternatives. Our results shed light on the computational principles underlying human-like rapid learning and planning.

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