Characterizing the mechanisms of instructed reinforcement learning with fMRI pattern-similarity analysis

Euan Prentis
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Nathan Tardiff
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Sharon Thompson-Schill
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

Past work has made conflicting proposals about the mechanisms underlying instructed reinforcement learning (RL)specifically, that prefrontal cortex, representing instruction, either biases, attenuates, or overrides learning signals in the brain. We leverage the sensitivity of pattern-similarity analysis of fMRI data to distinguish between the qualitative features of these accounts. Participants learn the value of six novel stimuli after receiving false information that one is of high value. We track markers of value learning in visual cortex during a value-independent perceptual judgement task presented between intervals of RL. We predict that with learning, the correlation between activation patterns for similarly valued stimuli will increase. To characterize influences on learning, we examine how the rate at and direction in which these patterns change in similarity will be influenced by explicit instruction about stimulus value. This work will help us identify the principle cognitive and neural mechanisms underlying instructed RL.