Competition Between Reinforcement and Similarity in Category Learning

John McDonnell
New York University

Todd Gureckis
New York University

Abstract: Reinforcement learning has become a popular framework for modeling human learning and decision making in a variety of tasks. However, it is less understood how issues of selective attention and conceptual representation (i.e., categories) interact with such trial and error learning to influence behavior in humans. Using a probabilistic categorization task, we examine the influence that categories have on learning from reinforcement. In particular, we evaluate the hypothesis that in complex, high-dimensional state spaces, participants will adopt a clustering scheme to represent highly similar situations. To the degree that any particular set of stimuli are mentally "clustered," we predicted that the probability of making a given response should match the mean probability of reinforcement for that group. By pitting similarity and reinforcement against one another in one condition and making them congruent in another, we were able to manipulate participant's acquired task representations, showing that stimulus similarity can cause "reward averaging" across stimuli in some situations. Overall, this work has important implications both for contemporary theories of human learning (based on RL) as well as for understanding the role that cognitive abilities such as categorization might play in human reinforcement learning in more complex, sequential tasks.