Abstract

Learning from trial-and-error is facilitated by prediction errors – signals reflecting discrepancy between expected and experienced results of our decisions. Positive prediction errors promote approach behaviors, while negative prediction errors lead to avoidance. One of the most influential findings of modern neuroscience was the discovery of prediction error coding in dopaminergic neurons. Using functional magnetic resonance imaging (fMRI), we can identify brain regions broadcasting prediction errors in various learning scenarios. In this thesis, I provide a holistic description of prediction error correlates in the reward-seeking and punishment-avoiding contexts. To elicit and investigate prediction errors, I used a probabilistic reversal learning task and scanned a group of healthy subjects using fMRI. I merged three complementary perspectives – behavioral, localization, and network – to comprehensively characterize the brain's implementation of reinforcement learning. On the behavioral level, I found that learning speed depends only on the sign of the prediction error and not on the experimental context. In line with the dual system hypothesis, activation analysis localized two independent sets of brain regions signaling positive-going and negative-going prediction errors. Whole-brain network analysis revealed a multi-scale community structure with a separate striatal reward network emerging at a finer topological scale and a ventromedial prefrontal network emerging at a coarser topological scale. I also found that the integration between large-scale networks increased when switching from positive to negative prediction error processing. The pattern of large-scale network reconfiguration followed the predictions of the Global Workspace hypothesis.