Panel Session
What Does Dopamine Say: Clues from Computational Modeling

The Role of Dopamine in the Temporal Difference Model of Reinforcement Learning

Read Montague*

Baylor College of Medicine, Houston, TX, USA

Background: Reinforcement learning models now play a central role in modern attempts to understand how the brain categorizes and values events traditionally framed by psychology as rewards and punishments. These models provide a way to design and interpret of reward expectancy experiments in humans across a wide range of rewarding dimensions. They also provide a connection to computational models of optimizing control, and hence connect the neurobiology of reward processing to simple forms of decision-making, even decision-making about social exchanges. A central signal in these computational accounts is the reward prediction error signal encoded by burst and pause responses in midbrain dopamine neurons. Numerous experiments have now provided strong evidence for the existence of such reward prediction error signals. Despite these successes, there is a missing piece to this story. The missing piece is a learning signal known as regret. By regret, we mean the difference between what 'could have been obtained' and what 'actually was obtained'.
Methods: We used several event related fMRI and hyperscan-fMRI experiments to probe both reward prediction error signals and regret signals in humans subjects. We studied the reward prediction error signals using a simple conditioning paradigm where a light predicted the temporally consistent arrival of a juice squirt in the mouth of 25 human subjects. We also probed the existence of reward prediction error signals in another domain, social exchange, using a two-person economic exchange game (a trust game) and hyperscan-fMRI (n=96 subjects). A third experiment was carried out on to probe neural correlates of regret single human subjects carrying out an investment task (basically a gambling game).
Results: All three experiments revealed strong correlates of these computational learning signals: reward prediction error and the regret signal. In both cases strong responses were observed in the ventral striatum, and in the case that choices were actually made by the subjects the prediction error signal activated ventral portions of the caudate nucleus consistent with previous reports using different tasks. The regret experiment showed exceptionally strong responses in the ventral putamen and also responses in Lateral Interparietal Sulcus area (LIP) that correlated with the value of the market fluctuation. In the trust experiment, we observe a signal in the ventral caudate that displays features of a reward prediction error signal.
Discussion: These results address three major issues. (1) They show that reward prediction error signals possess detectable correlates in human brains using functional magnetic resonance imaging. (2) They show that reward prediction error signals show up in the ventral putamen when no action is required by the subject to obtain reward and the ventral caudate and putamen when an action is required. (3) Regret signals are treated by the brain as real losses and drive changes in behavior (behavioral results) and furthermore that this signals represent another form of learning signal, a counterfactual reward error signal, that has detectable neural correlates in the striatum, thus suggesting one physical substrate for the experience of regret. (4) Collectively, these results show the utility of using computational models to search for neural correlates of signals involved in reward learning and perturbed by disease. This approach provides a new direction to more traditional methods of searching for neural correlates of reasonable psychological categories.

Dopamine Encodes a Quantiative Reward Prediction Error for For Reinforcement Learning

Paul W Glimcher*, O'Dhaniel A Mullette-Gillman, Hannah M Bayer, Brian Lau and Robb Rutledge

Neural Science, New York University, New York, NY, USA

Background: There is much evidence that the activity of midbrain dopamine neurons is correlated with the reward prediction error signal postulated by all reinforcement learning models. There has, however, been little effort devoted to testing the hypothesis that the activity of these neurons specifically encodes the reward prediction error term of any particular model or that the activity of these neurons can account for behaviors related to reinforcement learning processes. Our laboratory has attempted to address this with a three pronged approach. First, we have developed behavioral tools for quanifying reinforcement learning in humans and primates. Second, we have linked the trial-by-trial activity of dopamine neurons, measured in awake behaving primates, to the history of recent rewards which serve as the input data for reinforcement learning. Third, we have examined how changes in dopamine unit activity influence the ways in which the history of recent rewards influence behavior.
Methods: Reinforcement learning combines information about previous rewards and punishments in order to place values on actions. This is, however, not the only class of information that can influence the desirability of an action. Biases and the history of ones own choices (irrespective of the rewards that they have yielded) can also influence choice. We therefore developed a mathematical technique for analyzing the choices made by monkeys that allows us to determine the specific contribution previous rewards and punishments make to decision-making; a quantitative estimate of the reinforcement learning process. We performed this analysis on monkeys and humans performing a Matching-Law task of the type pioneered by Herrnstein. Our single unit approach is broadly similar. Once again we ask, here by linear regression, how the firing rates of single dopamine neurons are related to the previous history of rewards. If the behavioral and neuronal processes are identical then these two sets of measurements should also be identical.
Results: Our behavioral studies indicate that the segment of choice behavior which is driven by the history of recent rewards is strongly influenced by recent rewards and weakly influenced by rewards that are more distant in time. This weakening with time occurs with an exponential decay having a time course of about 7 trials, exactly as predicted by reinforcement learning theories like the TD model of Sutton and Barto. We find that the weighting function which relates the firing rates of dopamine neurons to the magnitudes and times of previous rewards precisely matches both the theoretical weighting function predicted by Sutton and Bartos model and the behaviorally derived estimates of the reinforcement learning process described above. Most recently we have begun to explore how changes in the activity of dopamine neurons influence our sophisticated behavioral measures of the reinforcement learning process. To this end we have examined the behavior of monkeys who receive electrical stimulation in the substantia nigra and of human Parkinsons patients, both on and off medication, during reinforcement learning tasks.
Discussion: Our results support the conclusion that midbrain dopamine neurons carry a reward prediction error of precisely the type required by reinforcement learning models. This activity appears sufficient to account for behavioral measurements of reinforcement learning and the contribution that these processes make to behavior.

Implications of the Temporal Difference Reinforcement Learning Model for Addiction and Relapse

A David Redish*

Neuroscience, University of Minnesota, Minneapolis, MN, USA

Temporal difference reinforcement learning (TDRL) algorithms have gained popularity to explain both behavior and the firing paterns of

dopaminergic cells. These models learn to predict value (expected predicted reward). If the agent (the animal or simulation) knows the value of the consequences of its actions, it can act to maximize that value. Estimated value is updated through a value-error term δ, defined as the difference between expected and observed changes in value. Addictive drugs have been hypothesized to access the same neurophysiological mechanisms as natural learning systems. A non-compensable drug-induced dopamine increase will drive a TDRL model to over-select actions leading to drug receipt. In this model, the agent incorrectly assigns ever-increasing value to drugs due to the noncompensable dopamine signal. Because willingness to pay is proportional to estimated value, as the estimated value approaches infinity, the willingness to pay increases proportionally. This willingness to pay provides an explanation for addicts continued attempt to find drugs, even at the expense of great and terrible costs. Because responses are so easily renewed after extinction, extinction cannot entail unlearning of the original association (Pavlov 1927, Bouton LearnMem 2004). Because standard TDRL models are generalizations of standard associative models, they do not differentiate learning from unlearning: a missing reward produces $\delta<0$, which produces a decrease in value (expectation of reward), which produces a decrease in action-selection. We propose instead that acquisition and extinction are driven by separate processes: Acquisition entails the development of an association, is based on phasic increases in dopamine, and is learned through increases in the value-estimate. Once this association has been learned, it is permanently stored and cannot be unlearned. Extinction entails the development of a new state space, which has no associated value-estimate. Tonically low δ (signaled by repeated pauses in dopamine neuronal firing) produces a splitting of the state space, such that a new state s' is created which can be differentiated from s. Evidence for dopamine antagonists producing representational instability has been found in frontal cortex (Zahrt et al. JNsci 1997), auditory cortex (Bao et al. Nature 2001), and hippocampus (Kentros et al. Neuron 2004). Relapse, then, occurs when the neural representation returns to the original representation which leads to the addictive path to drug-use. As with extinction processes, this implies that relapse will be particularly sensitive to context and other cues which can drive the representation back to the original representation. This learning-theory explanation of relapse is independent of whether the association produces positive desire for drugs or negative symptoms which need to be relieved. In either case, relapse occurs when the representation returns to the state s and makes the pathway to drug use available again. Reward/aversion can be categorized into four separate processes: reward (positive value larger than expected), disappointment (lack of expected positive value), aversion (negative value larger than expect), and relief (lack of expected negative value). We suggest that they arise from different neurological mechanisms and have different neurological consequences. Whether aversion and relief work in similar ways to reward and disappointment is unknown at this time, but the similarity of extinction processes on negative value consequences (e.g. cue leads to shock) to positive value consequences (e.g. cue leads to food) suggest that they might.

Dopamine-Norepinephrine Interactions: Exploitation versus Exploration

Jonathan D Cohen*, Samuel M McClure, Mark S Gilzenrat and Gary Aston-Jones

Psychology, Center for the Study of Brain, Mind and Behavior, Princeton University, Princeton, NJ, USA

Background: Adaptive behavior involves a trade-off between exploiting known sources of reward and exploring the environment for new, potentially more valuable ones. Research over the past decade suggests that DA mediates a learning signal that reinforces responses predictive of reward. Reinforcement learning (RL) models have successfully described DA activity in stable environments, but have not addressed more realistic conditions in which contingencies between responses and rewards may change — requiring that previously learned associations be ignored and new ones discovered. For such adaptations, RL models require additional apparatus ("annealing mechanisms") that detect when the environment has changed and promote exploration of new behaviors. Models of DA currently lack such mechanisms. However, recent studies suggest that the locus coeruleus-norepinephrine (LC-NE) system may serve this role. These studies have revealed two modes of LC function: a phasic mode, selectively favoring responses to task-relevant events, and a tonic mode producing a more generalized enhancement of responding. These findings suggest that the LC-NE system may implement an annealing mechanism for DA-mediated RL. This theory presupposes that the LC has access to evaluations of current task utility necessary to adjudicate between exploitation (high utility) and exploration (low utility). Recent anatomic findings support this, indicating that the two primary cortical projections to LC are from orbitofrontal and anterior cingulate cortex — areas consistently implicated in the evaluation of rewards and costs, respectively.
Methods: We implemented a model of interactions between DA-mediated RL (using the method of temporal differences), cortical mechanisms for decision making and evaluation of utility (reward rate and conflict), and an LC-NE annealing mechanism (simulating the dynamics of LC-NE activity). All of the mechanisms were drawn from previous models that accurately simulate relevant behavioral and physiological findings concerning these systems. We tested the model in a reversal conditioning experiment using a target detection task, in which the target identity was periodically reversed. The model's performance was examined with and without the LC-NE system, and was compared with behavioral and LC recordings from a non-human primate performing the same task.
Results: Without the LC-NE system, the model rapidly learned the initial target but took a protracted amount of time (several hundred trials) to learn to respond accurately following reversals. Introducing the LC-NE system dramatically improved learning following reversals (within 25-50 trials). Reversals were associated with transient decreases in LC phasic responding and increases in baseline firing (shift to tonic mode), followed by a return to the LC phasic mode as the new contingency was acquired. Both the performance of, and dynamics of LC activity in the model closely matched empirical observations in the same task performed by a monkey.
Discussion: The model demonstrates how DA-NE interactions may support the self-regulation of exploitation vs. exploration, a function critical to adaptive learning and decision making. More generally, it highlights the importance of interactions between neuromodulatory systems, above and beyond their individual functions. This is likely to have direct relevance to psychiatric disorders, which almost certainly involve disturbances of interactions between neuromodulatory systems that go beyond the simple excesses or deficits of individual systems commonly postulated by many existing theories.

Panel Session
The Role of Feeding Neuropeptides in Alcohol and Drug Dependence

Galanin and Opioid Peptides in Relation to Alcohol Intake and Dietary Fat: Possible Positive Feedback Mechanisms

Sarah Leibowitz*, Olga Karatayev, Valerie Gaysinskaya, Pedro Rada, Michael Lewis, Nicole Avena, Carmen Carrillo and Bartley Hoebel

Rockefeller University, New York, NY, USA

Recent experiments in our lab have demonstrated a close link between hypothalamic feeding-stimulatory peptides and both fat consumption and an associated rise in circulating triglycerides (TG). When injected into the paraventricular nucleus (PVN), which is involved in controlling food intake, the peptide galanin (GAL) and the opioids, enkephalin (ENK) and dynorphin (DYN), stimulate feeding