Context dependent, model-based reinforcement learning explains changes to reward devaluation over time

Trent Toulouse and Suzanna Becker
Department of Psychology, Neuroscience and Behaviour, McMaster University

One of the most exciting developments in linking machine learning and neuroscience was the discovery that the firing rate of dopamine cells in the ventral tegmental area correlates with reward prediction error, an important term in the temporal-difference (TD) machine learning algorithm [1]. It has been hypothesized that the neurological underpinnings of instrumental conditioning may be well modeled by TD learning or similar cached value algorithms. An alternative view is that the changes in firing rates of dopamine cells is not a prediction-error term but rather a measure of the improbability of a reward, signaling surprise and salience [2]. We use this conceptualization of dopamine to construct a model-based approach to reinforcement learning that explains behavioural and cell recording data that are contradictory to the standard TD approach.

Model-based and cached value based algorithms make different predictions about sensitivity to reward devaluation and changes in dopamine cell firing rates in response to changes in the probability of reward delivery. Cached value algorithms like TD learning will only adjust action policy in response to a change in the value of the reward when the new reward value is experienced paired with the action. Model-based approaches do not need to experience the new reward paired with the action in order to adjust action policy. The effects of devaluation reported in the literature are mixed, and depend upon context and amount of training. For example, moderately trained rats remain sensitive to devaluation while extensively trained rats are insensitive to devaluation [3].

Other experiments have looked at the change in dopamine cell firing rates in response to changes in the probability of reward delivery. Tobler et al. (2005) measured the cell firing rates in the ventral tegmental area of Macaque monkeys in response to a reward delivery. Three different visual cues served as conditioned stimuli where the magnitude of the reward differed by a factor of 10 but the probability of reward delivery was equal. The firing rate of the dopamine cells did not increase across conditions, which suggests that they are sensitive to reward probability and not reward magnitude.

Attempts have been made to reconcile the conflicting behavioral data by proposing that animals use both a model-based approach and a cached value algorithm simultaneously. Daw et al. (2005) proposed that the computationally costly model-based algorithm is used early on when uncertainty is high, and after training when uncertainty is low the more efficient cached value algorithm is used. However, this works only if it is assumed that the dopamine signal is a reward prediction-error which does not match all of the cell recording data available. Our proposed algorithm employs a model-based approach to instrumental learning which integrates context as a major element of learning reward value.

In our approach conditioning and associative learning tasks are constructed as a 4-tuple Markov Decision Process with state (s), actions (a), transition probability ($T(s,a,s')$), and reward function ($R(s,a)$) as parameters. The states and actions are known and then the transition and reward functions are modeled using Bayesian inference, with each iteration through the task updating the estimated values. Action policy at any state is based on comparing Q-values defined as:

$$Q(s,a)=E[R(s,a)]+\gamma\sum_{s'}T(s,a,s')max_{a}Q(s',a')$$

When a reward is received the dopamine cells change their firing rates based on the log of the estimated probability for that reward value as well as the estimated maximum future expected reward .

$$\Delta f_i = (-c \cdot \log(p(R)) + p(\sum_{s'} max_a Q(s', a'))) \cdot \sum_{s'} max_a Q(s', a')$$

The firing rate increases if an unexpected reward is received, and this increase in firing rate slowly moves to the earliest predictor of reward based on increasing confidence in the estimate of the reward value.

By comparing the size of this signal relative to the confidence in the estimated reward it is easy to see if something unexpected has occurred. Early on in the learning paradigm when the confidence in the estimate is low the dopamine signal should not trigger context dependent memory. But once the confidence in the estimate for the reward probability is high a smaller signal cues that the change in reward is context dependent and should be stored in a separate episodic memory.

Using this approach the conflicting data on devaluation is not due to shifts between two different neurological systems but rather one system that applies context cues to discriminate between generalizable difference in expected and actual consequences, or changes that are context specific. The dopamine signal helps this differentiation by signaling significant departures from expected reward probability, thus serving as a signal of surprise and salience. This approach can explain the conflicting behavioral data while better matching cell recording data.

References

[1] W. Schultz. Predictive Reward Signal of Dopamine Neurons. Journal of Neurophysiology, 80(1):1–27, 1998.

[2] A. Smith, M. Li, S. Becker, and S. Kapur. Dopamine, prediction error and associative learning: A model-based account. Network: Computation in Neural Systems, 17(1):61–84, 2006.

[3] N.D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience, 8(12):1704–1711, 2005.

[4] Tobler, P., Fiorillo, C., & Schultz, W. Adaptive coding of reward value by dopamine neurons. Science, 307(5715):1642-1645, 2005.