Choice values

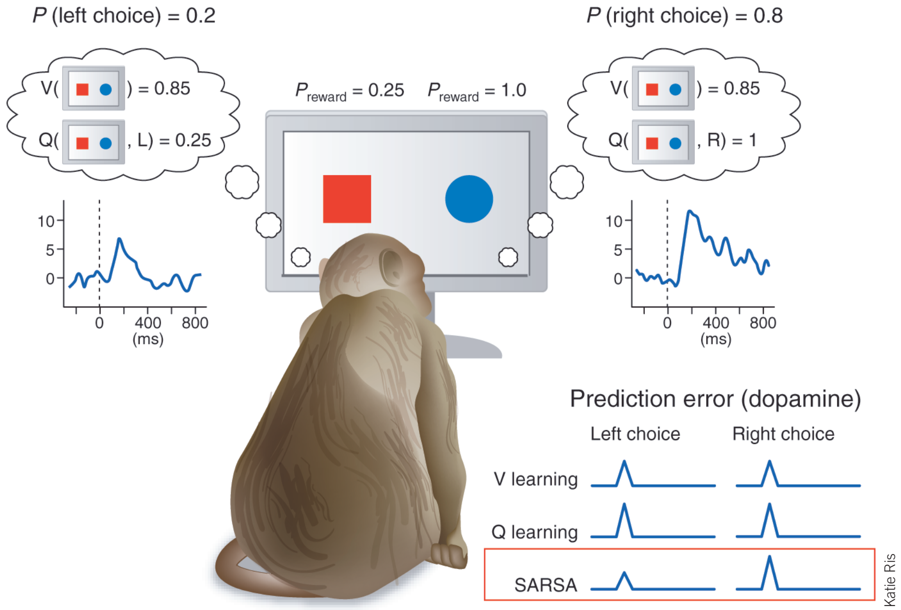
Yael Niv, Nathaniel D Daw & Peter Dayan

Dopaminergic neurons are thought to inform decisions by reporting errors in reward prediction. A new study reports dopaminergic responses as monkeys make choices, supporting one computational theory of appetitive learning.

A group of neurophysiologists, computer scientists, psychologists and economists has found common ground in trumpeting a detailed hypothesis about the involvement of the brain's dopaminergic system in reinforcement learning[1], which is the process of learning by trial and error to predict rewards (and punishments) and to make good decisions. This hypothesis is grounded in an impressive body of experimental evidence, notably recordings showing that dopaminergic neurons in behaving primates seem to carry an error signal that could be useful for learning to predict rewards and also to choose rewarding actions[2,3]. This presumed central role in appetitive choice aligns well with dopamine's involvement in synaptic plasticity, learned habits, drug addiction and various pathologies[4].

Maddeningly, however, the detailed neurophysiological evidence concerns prediction learning alone, leaving the story about decision making bereft of empirical guidance. This is because dopaminergic neurons have rarely been recorded when animals are making nontrivial choices between multiple rewarded options (that is, but for a few exceptions[5], doing much more than following instructions or passively receiving rewards). An experiment reported by Morris and colleagues in this issue[6] fills this gap and provides surprising and detailed information about the computations underlying decision making.

In the experiment, four different images were associated with fixed probabilities (0.25, 0.5, 0.75 and 1) of obtaining juice (or water) reward. In 'decision trials', monkeys saw pairs of these images and, after a couple of seconds, had to choose between them. Reward was given, or not, according to the probability attached to the chosen image. The monkeys knew these probabilities well, as the decision trials were embedded sparsely among a set of 'reference trials', in which the monkeys were presented with single images and responded for the associated chance of receiving juice. As in most previous dopamine recording studies, reference trials did not present the monkeys with a meaningful choice, and the particular prediction error reported by dopamine firing on presentation of an image simply corresponds to its associated reward probability[7].



Katie Ris

Figure 1 A monkey is presented with a choice between two options (decision trial). Each option is associated with a different probability of reward, so the overall worth of this choice can be evaluated in several different ways (thought balloons; see text). These different evaluations would lead to different prediction error signals associated with different reinforcement learning rules. The recordings of Morris *et al.* from dopamine neurons, putatively reporting these errors, show that the neural responses quickly reflect the distal future choice of the monkey (left and right traces; data from ref. 6). This supports the SARSA algorithm for prediction learning and action selection (bottom) against value learning and Q learning.

Faced with a decision trial, one might expect the monkeys to choose the richer option after pondering the pair for a while. Neither of these expectations was satisfied. Fortunately for science, but unfortunately for themselves, the monkeys adopted the (surprisingly common) suboptimal choice strategy of 'probability matching', pursuing richer and poorer options in rough proportion to their relative worth. This behavior was fortuitous because it allowed Morris *et al.* to record the activity of dopamine neurons on decision trials in which the monkey ultimately chose either the richer or the poorer option. The comparison of one with the other, and with the firing patterns from the single-image reference trials, provides a window onto the decision process.

The central finding is that a burst of dopaminergic responding at the outset of a decision trial nearly instantly reflects the average reward associated with the option that will ultimately be chosen, even though the monkey cannot actually submit its decision until some seconds later. Indeed, the neural response to the presentation of a pair of images is nearly the same in (average)

The authors are at the Gatsby Computational Neuroscience Unit, 17 Queen Square, London WC1N 3AR, UK. Yael Niv is also at the Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel.
e-mail: dayan@gatsby.ucl.ac.uk

time course and magnitude as the response on reference trials to the presentation of only the image the animal will choose. Within just 200 ms, the monkey has evidently made a decision and communicated it to the dopaminergic cells.

The beauty of this result is that it lays waste to a crowded field of computational ideas about appetitive choice; under a friendly interpretation, it leaves just one survivor. An immediate casualty is the idea that the dopamine signal might be directly involved in selecting actions[8]—instead, the firing apparently reflects a choice already made. Most other accounts of reinforcement learning assume that dopaminergic responses affect decisions only indirectly, by controlling learning. All the accounts agree that the dopaminergic prediction error at the start of a decision trial reports an evaluation of the overall value (predicted reward) of the trial. However, their substantially different approaches to learned choice are reflected in subtly different ways of assessing this value, which the results of Morris *et al.* exactly test. There are three main possibilities (**Fig. 1**) for what the value of a decision trial is, in terms of the values of the two images it comprises. It could be (i) the average of the values of the two options, weighted according to the probability that each would be chosen, (ii) the value of the better option, or (iii) the value of the image that is actually chosen on that trial. The data favor the last possibility.

The first option—that the values are averaged over the choices (V in **Fig. 1**)—would have been expected under the so-called actor-critic algorithm, which posits that a 'critic' with no knowledge of the actions can track the average value of situations (called 'states'); these values can, separately, be used as rewards to train an 'actor' that makes choices. This notion mirrors venerable ideas from psychology about the interaction between reward prediction (Pavlovian conditioning in the critic) and action choice (instrumental conditioning in the actor)[9], and seems nicely to parallel the anatomical division of the dopamine system and its targets into ventral (evaluation) and dorsal (action) components[10]. The central trick of the actor-critic algorithm is how it learns to choose actions using reward predictions that ignore actions altogether, instead averaging over them. However, the data of Morris *et al.* rule out this trick and show that the dopamine signal instead incorporates richer information, separately reporting the value of choosing either action at a state.

The value of taking a particular action at a state is called a Q value[11], and the two remaining ways to evaluate a decision trial can both be used to learn Q values. The more popular approach is Q learning[11], in which the prediction error associated with a decision (which is what the dopamine cells report) is determined by the Q value of the better option rather than the one actually chosen (**Fig. 1**, bottom). This is a very clever idea, as it decouples learning from the actual choice and allows optimal behavior to be acquired while exploring suboptimal alternatives. However, going by the data, this is evidently too clever for the dopamine cells, whose activity follows the reference activity for the action actually chosen. The remaining option is the class of algorithms that acquire Q values using a prediction error that reflects the value of the chosen option. It is these so-called SARSA (state-action-reward-state-action) algorithms[12] (**Fig. 1**) that this study favors.

In sum, the most natural conclusion from the neural data is that dopamine signals report prediction errors based on Q values for SARSA learning. A choice can be made between actions by favoring (perhaps still subject to randomness) the one with a larger Q value. This would account for the animal's fortuitous but flawed probability matching behavior. The behavior itself is both informative and surprising. Under methods such as the actor-critic, persistent performance of a suboptimal action is not possible. That it happens is additional evidence that choice is based on Q values. However, it is odd that the monkeys seemed never to adjust their behavior toward exclusive choice of the richer option. The common rationale for occasional suboptimal choices is to allow for exploration of unfamiliar alternatives, but no such exploration was necessary here because the images' values were stable over weeks of recording and anyway sampled extensively during reference trials.

As with any illuminating result, many open issues and interesting implications remain. First, even though dopamine seems not to be involved directly in the choice between options, it may influence other aspects of the selected action, such as the vigor with which it is executed[13]. Second, dopaminergic responses during decision trials, though similar on average to those on reference trials, were nevertheless much more variable. Structure in the signal may still remain undiscovered, perhaps including evidence of the monkey changing its mind during the waiting period. Third, from a reinforcement learning perspective, it is not straightforward to expect that the monkey will represent the state of a pair of images as simply the conjunction of the two single-image states. Fourth, it is now pressing to work out a SARSA-like algorithm that also respects the anatomical data on the dual dopamine and striatal systems that helped motivate the actor-critic model. A relative of the actor-critic algorithm called 'advantage learning', which has found some support in human functional magnetic resonance imaging (fMRI) studies of learned choice[10], seems not to do the trick, but a variant might.

Finally, the monkeys in this study were vastly overtrained, which allowed for careful study under uniform conditions, but at the cost of ensuring that no learning occurred during the experiment. The recorded reinforcement learning error signals were therefore apparently epiphenomenal. In studies of similar decision-making tasks, animals have been exposed to more complex and changing reward contingencies and continually updated their behavior in light of received rewards[14,15]. An obvious future direction is to understand how such behavioral changes relate, trial by trial, to recorded dopamine responses. The experiment of Morris *et al.* sets the stage for a new multidisciplinary enterprise of such studies of dopamine and decisions.

1. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
2. Houk, J.C., Adams, J.L. & Barto, A.G. in *Models of Information Processing in the Basal Ganglia* (eds. Houk, J.C., Davis, J.L. & Beiser, D.G.) Ch. 13, 249–270 (MIT Press, Cambridge, Massachusetts, 1995).
3. Schultz, W., Dayan, P. & Montague, P.R. *Science* **275**, 1593–1599 (1997).
4. Wise, R.A. *Nat. Rev. Neurosci.* **5**, 483–495 (2004).
5. Bayer, H.M. & Glimcher, P.W. *Neuron* **47**, 129–141 (2005).
6. Morris, G., Nevet, A., Arkadir, E. & Bergman, H. *Nat. Neurosci.* **9**, 1057–1063 (2006).
7. Morris, G., Arkadir, D., Nevet, A., Vaadia, E. & Bergman, H. *Neuron* **43**, 133–143 (2004).
8. McClure, S.M., Daw, N.D. & Montague, P.R. *Trends Neurosci.* **26**, 423–428 (2003).
9. Dickinson, A. *Contemporary Animal Learning Theory* (Cambridge Univ. Press, New York, 1980).
10. O'Doherty, J. *et al. Science* **304**, 452–454 (2004).
11. Watkins, C. *Learning from Delayed Rewards.* Thesis, Cambridge Univ. (1989).
12. Rummery, G.A. & Niranjan, M. in *Technical Report* CUED/F-INENG/TR 166 (Engineering Department, Cambridge Univ., 1994).
13. Niv, Y., Daw, N.D. & Dayan, P. in *Advances in Neural Information Processing Systems* (NIPS) Vol. 18 (eds. Weiss, Y., Schölkopf, B. & Platt, J.) 1019–1026 (MIT Press, Cambridge, Massachusetts, 2005).
14. Lau, B. & Glimcher, P.W. *J. Exp. Anal. Behav.* **84**, 555–579 (2005).
15. Corrado, G.S., Sugrue, L.P., Seung, H.S. & Newsome, W.T. *J. Exp. Anal. Behav.* **84**, 581–617 (2005).