

A Step-by-Step Guide to Dopamine

Peter Dayan and Mark E. Walton

There is an odd irony associated with the by-now almost ineluctable tie between reinforcement learning's temporal difference learning rule (1) and the phasic activity of dopamine neurons (2). Although temporal difference learning was designed to enable the acquisition of whole sequences of actions and predictions, a task that its ancestors (3) would flub, there are very few direct tests of this characteristic. In a penetrating new study, Wassum and colleagues (4) measured and manipulated dopamine in a sequence learning task for sucrose reward, revealing four correlates of the neuromodulator: its fine- and gross-scale dynamics during learning, its involvement in two aspects of energizing behavior, and its necessity for learning.

The task is straightforward. Rats were placed in an operant chamber in which one (distal) lever was freely available to be pressed at any time. A response on this lever caused a second (proximal; mildly pretrained) lever to emerge, and when this was pressed, a sucrose pellet was delivered to a food well and the proximal lever was retracted. Over the course of training, the animals learned to make increasingly efficient sequential responses: distal press, proximal press, and consumption.

To determine how dopamine transmission is modulated during the development of this behavior, the authors used fast-scan cyclic voltammetry (FSCV) to measure changes in dopamine concentration in the ventral striatum. A strength of FSCV is that it allows reliable trial-by-trial monitoring of phasic dopamine transmission in individual animals. Wassum and colleagues used this to their advantage to demonstrate the dynamic modulation of dopamine as the animals improved their efficiency on the sequence. They used FSCV in three groups of animals. For one group (labeled day 0), measurements were made on the very first day they were introduced to the sequence task. For the second and third groups, FSCV was performed after moderate (day 5) or extensive (day 10) training. Furthermore, in a separate experiment, the authors used the nonspecific dopamine antagonist flupenthixol systemically to establish whether dopamine release was required for the appropriate sequence of actions to be learned and performed.

Treating this as an instrumental learning task (albeit with some important Pavlovian elements, such as the sound of the proximal lever extending or retracting), we might expect phasic dopamine to report an appetitively signed temporal difference prediction error. At least to a first approximation, this would imply that early in learning, dopamine release would increase to the unexpected reward following the proximal press; but late in learning this release would be absent because the reward would have been well predicted and so inspire no prediction error. Nonetheless, even at the end of training, unexpected, unearned rewards should still lead to dopamine release. This is exactly what was observed and is an important novel finding because past FSCV studies of free operant instrumental conditioning have used cocaine as a reward (5). The

pharmacology of that drug means that dopamine concentrations always rise following its delivery, even when predicted.

If the various elements of the task had been cued, then by standard temporal difference theory, we would expect the dopamine transients to transfer to earlier points in the trial, potentially acting as secondary reinforcers for the initial elements of the sequence of actions (such as pressing the distal lever) that by themselves do not lead to immediate reward. It is precisely this type of transfer that the Rescorla-Wagner rule (3) cannot manage. In a free operant task such as this, behavior is self-paced rather than cued, and the exact prediction that temporal difference learning would make about dopamine transients is not quite so clear. However, a latent decision to press must have been made prior to the press itself. In this case, at the end of learning, one might expect a dopamine transient just before the execution of the press (as indeed in a regular cued case but without any visible cue for alignment of traces), coincident with the resolution of temporal uncertainty about the future reward. As in previous free operant studies (5), this is indeed just what was observed, for both the distal and proximal levers. Something similar was observed in a task involving a cued lever press for sucrose reward, at least on trials on which the response was only weakly time locked to the cue (6).

Of course, as well as being associated with learning, dopamine has also been implicated in various aspects of the energization and vigor of behavior (7). Two aspects of this are apparent in the present study. One is that at all levels of training, under the antagonist flupenthixol, the overall rate of sequence initiation, and thus reward acquisition, was decreased. This is rather as expected on theoretical grounds (8). However, FSCV monitors changes in local, phasic dopamine transmission, whereas the drug has prolonged systemic and nonspecific actions on D1 and D2 receptors. Thus, actions of the neuromodulator in different regions and different time scales might be contributing to this effect over and above the rapid changes in dopamine observed electrochemically.

The other aspect of vigor is perhaps more straightforward, namely that the duration of sequences was inversely correlated with the magnitude of the initial dopamine transient particularly during initial acquisition, reminiscent of a similar, electrophysiological finding for a single action in monkeys (9). This, combined with the results on learning, will nicely warm the hearts of both sides of the seeming debate about the role of phasic dopamine (6), and particularly those of the ecumenicalists (10).

Additional features of these results will quicken the pulses of various other researchers. First, reinforcement learning theorists will be enthused at the prospect of analyzing the transients over the course of initial learning when the transfer is happening to provide insight into the details of trial-by-trial correlations between dopamine transients (and indeed behavior). They would also be intrigued at what happens in the moderate percentage of trials in which the well-trained animals act inefficiently, by pressing the distal lever twice in succession rather than pressing the newly inserted proximal lever. It might be possible to read from the dopamine signal in these cases whether this stems from some form of disengagement with the task, or from an active, stochastic, but evidently incorrect choice. Future work monitoring the progress of learning in individual animals using chronically implanted electrodes (11) will also provide crucial data to speak to both of these issues.

From the Gatsby Computational Neuroscience Unit (PD), London; and the Department of Experimental Psychology, University of Oxford (MEW), Oxford, United Kingdom.

Address correspondence to Peter Dayan, Ph.D., Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London WC1N 3AR, United Kingdom; E-mail: dayan@gatsby.ucl.ac.uk.

Received March 8, 2012; revised and accepted March 9, 2012.

Second, one of the intriguing observations is that after the most substantial training, the time of the peak of the dopamine transient preceding the press of the distal lever became progressively less tightly coupled to the time of the press itself. It has been suggested that transient dopamine release plays a causal role in flexible approach responses to important aspects of the environment such as levers (12). It would be most interesting to correlate the details of the behavior of the subjects to the timing of this peak: does it happen when the subjects apparently suddenly become reengaged in the prospect of working?

Third, voltammetrists might enjoy picking over the fine details of the timing and relative levels of dopamine at the four key time points in the task, before and after the distal lever press, and before and after the proximal lever press and reward. It was surprising to see that the dopamine transient to the earned reward apparently persisted almost unabated on day 5, even though behavior had seemingly nicely stabilized by then, and the sequential behavior was largely unaffected by flupenthixol administration. It must have been hard for the authors to resist the temptation of using different amounts of reward across the trials to engender a richer set of prediction errors on the trials.

Finally, students of behavior will be eager to dissociate Pavlovian and instrumental aspects of the task. As currently set, purely Pavlovian mechanisms should do a pretty good job at organizing behavior in this paradigm. That is, it would suffice that distal and proximal levers are associated with reward to encourage engagement, approach, and pressing; the instrumental contingency between pressing and outcome might not in fact be playing a key role in the task. Indeed, the ventral striatum is often considered to be a substrate of Pavlovian influences over action, in fact including phasic, dopaminergically influenced energization as in Pavlovian to instrumental transfer. By contrast, the dorsal striatum (and particularly the dorsolateral striatum) is implicated in instrumental action learning, although it is reported to be a much more elusive target for recording dopamine transients (P.E.M. Phillips, personal communication), at least within the confines of an operant chamber. It would be irony indeed that if underlying these wonderful results, which apparently offer strong support for neural reinforcement learning ideas, was actually no (at least Thorndikian) reinforcement at all.

We are very grateful to the authors of the original study for their help. Dr Dayan receives research funding from the Gatsby Charitable Foundation. Dr Walton receives research funding from the Wellcome Trust.

The authors report no biomedical financial interests or potential conflicts of interest.

1. Sutton RS (1988): Learning to predict by the methods of temporal differences. *Machine Learning* 3:9–44.
2. Montague PR, Hyman SE, Cohen JD (2004): Computational roles for dopamine in behavioural control. *Nature* 411:760–767.
3. Rescorla RA, Wagner AR (1972): A theory of Pavlovian conditioning. Variations in the effectiveness of reinforcement and nonreinforcement. In Black AH, Prokasy WF, editors. *Classical Conditioning II: Current Research and Theory*, New York: Appleton-Century-Crofts, 64–99.
4. Wassum KM, Ostlund SB, Maidment NT (2012): Phasic mesolimbic dopamine signaling precedes and predicts performance of a self-initiated action sequence task. *Biol Psychiatry* 71:846–854.
5. Stuber GD, Roitman MF, Phillips PEM, Carelli RM, Wightman RM (2005): Rapid dopamine signaling in the nucleus accumbens during contingent and noncontingent cocaine administration. *Neuropsychopharmacology* 30:853–863.
6. Roitman MF, Stuber GD, Phillips PE, Wightman RM, Carelli RM (2004): Dopamine operates as a subsecond modulator of food seeking. *J Neurosci* 24:1265–1271.
7. Berridge KC (2007): The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology (Berl)* 191:391–431.
8. Niv Y, Daw ND, Joel D, Dayan P (2007): Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl)* 191: 507–520.
9. Satoh T, Nakai S, Sato T, Kimura M (2003): Correlated coding of motivation and outcome of decision by dopamine neurons. *J Neurosci* 23: 9913–9923.
10. McClure SM, Daw ND, Montague PR (2003): A computational substrate for incentive salience. *TINS* 26:423–428.
11. Clark JJ, Sandberg SG, Wanat MJ, Gan JO, Horne EA, Hart AS, et al. (2008): Chronic microsensors for longitudinal, subsecond dopamine detection in behaving animals. *Nat Methods* 7:126–129.
12. Nicola SM (2010): The flexible approach hypothesis: Unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. *J Neurosci* 30: 16585–16600.