Surprise! Dopamine signals mix action, value and error

Anne G E Collins & Michael J Frank

Two studies invite us to reconsider the nature of striatal dopamine signals. Accumbens dopamine appears to signal the value of overt action and prediction errors arise from deviations in these signals.

The neuromodulator dopamine (DA) is crucial to motivation, learning and action; its dysfunction is implicated in motor disorders, addiction and apathy. However, how dopamine operates across these domains remains poorly understood. On one hand, a large amount of literature focuses on the role of dopamine in motivation to exert physical effort to obtain a reward[1] or the incentive-motivating properties of stimuli that drive 'wanting'[2]. On the other hand, an influential theory proposes that phasic changes in dopaminergic signals reflect reward prediction errors (RPEs), or outcomes that are better or worse than expected[3]; these are used to potentiate corticostriatal synapses, thereby implementing a form of reinforcement learning. Although both theories have enjoyed much empirical support[4], two new studies[5,6] recording dynamics of accumbens DA concentration during instrumental learning suggest that an integration of the two (**Fig. 1a**) may be more appropriate.

Syed *et al.*[6] tested the effect of overt action selection on DA signals in an instrumental task. Rats were trained to distinguish cues indicating whether they would be rewarded for selecting an action (pressing a lever) or for withholding action (staying put until a fixed delay). The unexpected appearance of a cue indicated a change in expected future reward and was therefore expected to lead to a phasic increase in DA, independently of the required action. Using fast-scan cyclic voltammetry, the authors reported that DA levels varied with RPEs, but only in those trials requiring overt action to obtain reward. These signals were strongly diminished or absent for cues signaling the need to inhibit action—despite equivalent behavioral success and equivalent delays between cues and rewards. In those trials, phasic increases in DA related to reward expectation were nevertheless observed following movement initiation to collect the reward, after successful waiting triggered by the cue, and phasic decreases in DA were observed when the animals failed to withhold responses and cues indicated that rewards would not be obtained. Thus bidirectional RPEs were observed, but only in anticipation of, or following, active actions.

Dopamine generally increases motor activity, especially in rewarding contexts (Pavlovian approach), so DA release could lead to maladaptive learning when action needs to be inhibited to obtain a reward[7]. Thus, these findings may indicate that the DA system is smarter than we thought: gating its release by action selection would help to mitigate such Pavlovian learning biases. More broadly, the findings may provide preliminary evidence for a more general credit-assignment mechanism, whereby active action selection by basal ganglia can disinhibit phasic DA release[8] to enhance reward learning about those outcomes that it has caused[9]. Indeed, human active learning is consistent with such a mechanism: subjects exhibit learned preferences for freely chosen rewarded actions over those that were not chosen, even given identical reinforcement histories, and these preferences are modified by dopaminergic genetic variants[9].

Hamid *et al.*[5] provocatively recasts the quantity signaled by accumbens DA in terms of reward value, rather than RPEs *per se*. Although the predominant evidence for RPEs comes from DA neuron electrophysiology, a few voltammetry studies, including that described above (albeit modulated by action), have largely provided converging support for these bidirectional signals[10]. However, voltammetry measurements exhibit slow drifts across time, and the documented phasic signals are hence always evaluated relative to the pre-cue baseline. Because RPEs are inherently deviations in reward value, Hamid *et al.*[5] reasoned that this analysis could obscure the true nature of DA signals in terms of value. Moreover, a previous study suggested that DA levels ramp up as animals approach a rewarding location, as is consistent with a signal related to reward expectation[11].

To address this issue, Hamid *et al.*[5] performed a series of experiments in which they assessed and manipulated accumbens DA levels while rats learned to select between actions yielding different reward probabilities. First, using microdialysis (over a longer timescale of many trials), they reported that DA concentration was best related to the ongoing minute-by-minute reward rate and predicted task engagement, supporting the role of DA in mediating value-dependent motivation. However, by itself, this could reflect either a slow time-varying motivational component or the effect of multiple phasic learning signals.

Next they investigated the within-trial dynamics of DA using voltammetry. The main finding was that when rats were engaged in the task, DA levels exhibited progressive ramping from the initial cue, consistent with a reinforcement-learning model in which value progressively increases because of the discounting of predicted rewards that are further into the future. They also reported abrupt DA fluctuations at events carrying new information (task cues, reward cues), with the size of these jumps decreasing with learning. Thus, the voltammetry signal appears to mix value (ramping) signals with (abrupt, and declining with learning) RPEs. However, the authors noted that the abrupt signals could simply reflect deviations in value, with no need to invoke a separate RPE. They performed a careful analysis to disentangle these possibilities, supporting the value theory. Indeed, the declining 'RPE' signals with learning could be attributed to an increasing baseline from one trial to the next in the presence of a fixed peak (**Fig. 1b**), rather than a decline in the height of the peak in the presence of a fixed baseline. This clever analysis implies that baseline DA evolves with estimated expected value and that prediction errors simply reflect the changes in such value estimates before and after a task event.

But do these DA transients affect learning nonetheless, and/or perhaps directly guide motivational choice? To answer this question, the authors used optogenetics to causally probe the function of DA during distinct task periods. They first showed that phasic stimulation of accumbens DA release following choice acted to reinforce the rat's choice, and that phasic inhibition of DA acted to punish that choice (decrease its subsequent probability), independent of overt reward. This

Anne G.E. Collins and Michael J. Frank are in the Department of Cognitive, Linguistic and Psychological Sciences and Brown Institute for Brain Sciences, Brown University, Providence, Rhode Island, USA. Anne G.E. Collins is also in the Department of Psychology, University of California, Berkeley, California, USA.
e-mail: anne_collins@brown.edu or michael_frank@brown.edu

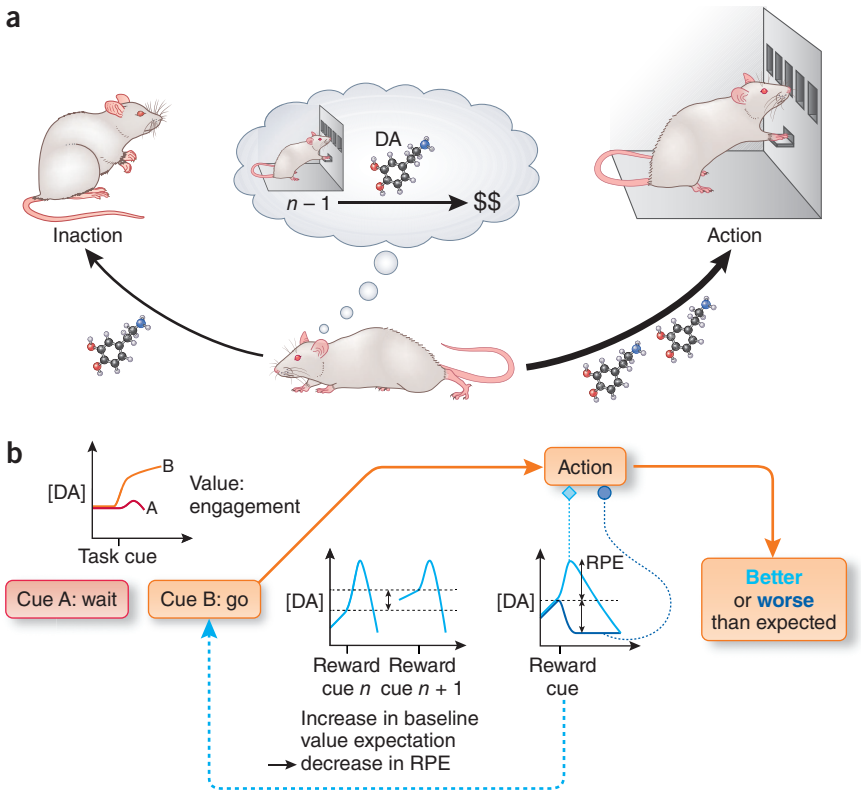

Figure 1 Functions of dopamine in action and learning. (**a**) DA concentration signals the value of overt action and directly invigorates choice accordingly. DA can also signal an RPE during reward on trial $n-1$, reinforcing the value of the action so that it is invigorated on trial n. (**b**) Dynamics of dopamine during task events. A cue signaling future reward elicits initial jump and ramping in DA, but only when an overt action is required. Reward is anticipated by ramping DA encoding expected value. Reward occurrence or omission is signaled by a jump in value, mimicking an RPE, which serves as a bidirectional learning signal (dashed lines) for both the value of the predictive cue and the action. Phasic dopamine bursts increase subsequent value estimates and thus baseline DA on trial $n+1$, diminishing the size of future phasic dopamine bursts.

is perhaps itself the clearest demonstration yet that abrupt changes in striatal DA do act as a bidirectional instrumental learning signal, within a single task procedure and using physiologically relevant magnitudes and durations of DA stimulation or inhibition. Moreover, the same stimulation protocol when applied at trial onset (rather than during the outcome) did not influence learning, but instead changed the likelihood that animals would engage in the task. Increases in DA diminished latencies to engage and approach one of the action ports, and inhibition of DA increased such latencies. Thus, during choice, causal manipulations of DA appear to influence the effect of expected value on task engagement, whereas during outcomes these manipulations induce learning.

Together, both studies highlight the important role of action and motivated task engagement in dopaminergic reward-related signaling. Effects on action or motivation and RPEs are typically considered in isolation, from different traditions and literatures, but these studies provide timely evidence for interactions between these factors and begin to bridge our understanding of dopamine across them. Indeed, many empirical studies using DA manipulations in both humans and animals conflate motivational and learning interpretations: a drug modulation of reinforcement learning curves, for example, could be explained either by changes in learning from RPEs or by increased emphasis on learned reward values during choice. Conversely, apparent motivational effects can sometimes be attributed to learning. Theoretical models have suggested that DA modulates both learning and motivated choice via common mechanisms acting on D1 and D2 dopamine receptor–containing striatal neurons, and have shown the need to consider both factors and their interactions to account for a variety of findings across species[12]. Such models will now need to consider the implications of dynamically changing DA transients such as the ramping observed by Hamid *et al.*[5].

As provocative findings often do, these studies raise many new questions. For example, Hamid *et al.*[5] did not manipulate the level of effort *per se*, but did manipulate the value of task engagement, and they propose a parsimonious theory that DA, at all scales, represents expected value and thus potentiates both choice (engagement, motivation) and learning, with RPEs signaled by abrupt changes. We highlight here three points that will require further investigation to probe the limits of this as compared to existing theories.

First, the ramping value signals observed here and in ref. 11 differ from those observed in the many electrophysiology studies in which spiking changes as a function of prediction errors, but not value *per se*. How can these discrepancies be reconciled? Hamid *et al.*[5] note that striatal DA release can be modulated presynaptically (for example, by local circuits) so that striatal DA concentration does not solely reflect spiking of afferent neurons, but the mechanism by which this circuit would convey time-varying values remains to be clarified. Second, if DA concentration represents value rather than RPEs, it raises the question of how target structures differentiate between absolute and relative changes in DA for implementing learning. Indeed, depending on prior expectations, the very same 'value' (and hence DA level) can be a positive or negative RPE, and efficient learning requires treating these differently. One possibility is that tonically active neurons, a cholinergic population that pause during windows of dopaminergic RPEs, may signal when to learn and can further enhance the contrast between tonic and phasic DA signals[13], although this remains speculative.

Finally, on a broader level, RPE theory has been so successful in part because it has generated clear and testable predictions, leading to experimental designs that parametrically manipulate factors that affect RPE with compelling results[14], even satisfying an axiomatic description of RPEs independent of any specific implementation[10]. Moreover, RPE models can still show ramping under certain circumstances and depending on assumptions[15], so some purists may prefer the parsimony of a single RPE mechanism until further evidence for value theory accrues. Thus, any new theory should be subject to the same rigor and falsifiable predictions for follow-up work. For example, an intriguing and counterintuitive prediction of the DA value theory proposed by Hamid *et al.*[5] is that if one could exogenously and selectively increase (or decrease) the DA baseline while allowing DA to respond endogenously to rewarding events, this should impair (versus improve) learning by reducing the local change in DA. Regardless of the outcome of these or other tests, these papers provide an exciting new bridge to reinvigorate our enthusiasm

and update our understanding of dopamine in action, motivation and learning.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Salamone, J.D. *et al. Curr. Top. Behav. Neurosci.* published online, doi:10.1007/7854_2015_383 (1 September 2015).
2. Berridge, K.C. *Eur. J. Neurosci.* **35**, 1124–1143 (2012).
3. Montague, P.R., Dayan, P. & Sejnowski, T.J. *J. Neurosci.* **16**, 1936–1947 (1996).
4. Robbins, T.W. & Everitt, B.J. *Psychopharmacology (Berl.)* **191**, 433–437 (2007).
5. Syed, E.C.J. *et al. Nat. Neurosci.* **19**, 34–36 (2016).
6. Hamid, A.A. *et al. Nat. Neurosci.* **19**, 117–126 (2016).
7. Guitart-Masip, M. *et al. Neuroimage* **62**, 154–166 (2012).
8. Lobb, C.J., Troyer, T.W., Wilson, C.J. & Paladini, C.A. *Front. Syst. Neurosci.* **5**, 25 (2011).
9. Cockburn, J., Collins, A.G. & Frank, M.J. *Neuron* **83**, 551–557 (2014).
10. Hart, A.S., Rutledge, R.B., Glimcher, P.W. & Phillips, P.E.M. *J. Neurosci.* **34**, 698–704 (2014).
11. Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E.M. & Graybiel, A.M. *Nature* **500**, 575–579 (2013).
12. Collins, A.G.E. & Frank, M.J. *Psychol. Rev.* **121**, 337–366 (2014).
13. Cragg, S.J. *Trends Neurosci.* **29**, 125–131 (2006).
14. Fiorillo, C.D. & Tobler, P.N. *Science* **299**, 1898–902 (2003).
15. Gershman, S.J. *Neural Comput.* **26**, 467–471 (2014).

Parietal and prefrontal: categorical differences?

Daniel Birman & Justin L Gardner

A working memory representation goes missing in monkey parietal cortex during categorization learning, but is still found in the prefrontal cortex.

If you can imagine reading to the end of this sentence and forgetting what was written at the beginning, you may start to appreciate how critical working memory is to much of what we take to be higher cognition. Indeed, a life without such short-term memory would lurch between disconnected events, threatening not just our cognitive abilities, but the core continuity of our conscious selves. The finding in the 1980s that, during delay periods in which monkeys remembered the location of an instructed eye movement, prefrontal[1] and parietal[2] neurons display persistent, spatially specific activity cemented the idea that these cortical areas are allied in serving this critical memory function. However, in this issue of *Nature Neuroscience*, Sarma *et al.*[3] report parietal neurons in the lateral intraparietal area (LIP) to be unexpectedly and puzzlingly forgetful, whereas their counterparts in the prefrontal cortex are not.

This finding comes from experiments probing another hallmark cognitive function for which parietal and prefrontal neurons appear to share responsibility: categorization. Categorization is our ability to generalize properties of, say, an apple across many exemplars with incidental differences in size, color or shape. Without categorization, each and every apple might have to be individually memorized. Categorization is clearly a foundational cognitive capacity; we use it not just when we recognize an apple, but when we distinguish specific states of importance, such as whether it is edible or rotten. We might imagine that categorical decisions are just as critical for a monkey as for a human.

Daniel Birman and Justin L. Gardner are in the Department of Psychology, Stanford University, Stanford, California, USA.
e-mail: jlg@stanford.edu

But how exactly does one get a monkey to make categorical decisions repeatedly and in a controlled way so that the neural representations can be systematically studied?

Continuing in the tradition of their laboratory[4], Sarma *et al.*[3] have developed formidable skills in training monkeys to do just this. In the current work, they shaped behavior incrementally, starting with a sequence of simpler tasks before making the leap into categorical decisions. The authors trained monkeys to perform a delayed match-to-sample task in which the monkeys were required to remember the direction of a patch of briefly presented moving dots and release a lever only if the direction of a second dot patch, presented after a short delay, exactly matched the remembered direction. After monkeys reached criterion performance, they were trained on the full categorization task, which was identical except for one crucial difference. In the categorization task, the monkeys were trained to release the lever not when the two directions were identical but when they were in the same experimenter chosen category (that is, moved in the same direction relative to an arbitrary category boundary). A series of studies has provided abundant support that, after such categorization training, neurons in LIP and prefrontal cortex show a beautifully simple and stable representation of the category, preferring stimuli in either one category or the other[4].

The new insight concerning working memory came from a relatively simple proposition that led to an unexpected result: record the activity of neurons before as well as after the training of the categorization task. Given that, after training, LIP neurons encode category during the delay period, might one expect that a working memory representation encoding direction would be present for the delayed match-to-sample task? Oddly, the answer is no. Despite the fact that, to perform the task, the monkey must have some working memory of the direction of the dots, neurons in LIP showed near chance-level encoding of motion direction during the delay period. This lack of delay-period selectivity is not a result of information about motion direction not entering LIP; during stimulus presentation, the motion direction was clearly represented. The neurons just seem to forget the direction when it most matters. Notably, the nearby medial superior temporal area, MST, which also receives direction-selective information from the medial temporal motion selective area MT, has by contrast been shown to carry working memory representations of motion direction[5].

Although clearly a provocative finding, task and training are intertwined in the experiments as outlined above, making it unclear which accounts for the difference in LIP activity. Parietal working memory activity might arise only after sufficient training. Given that the delayed match-to-sample task was trained first, it is possible that lack of working memory representation comes not from a difference in task, but from a lack of training. The authors have provided a detailed view into their training structure, which helps to allay this concern. In particular, the monkeys were extensively trained on the delayed match-to-sample task (hundreds of daily sessions and hundreds of thousands of trials), and their performance had plateaued. Although sorting out training and task by design awaits future replication studies, clearly the monkeys had substantial experience with the delayed match-to-sample task, therefore suggesting that task, and not training, accounts for the difference.

Potential alternative explanations aside, pause for a second to consider the extensive training regime. Why does it take monkeys so long to learn so little? It seems intuitive that a