CellPress
REVIEWS

Spotlight

Beyond the Average View of Dopamine

Angela J. Langdon [1,*] and Nathaniel D. Daw [1,*]

Dopamine (DA) responses are synonymous with the 'reward prediction error' of reinforcement learning (RL), and are thought to update neural estimates of expected value. A recent study by Dabney *et al.* enriches this picture, demonstrating that DA neurons track variability in rewards, providing a readout of risk in the brain.

Imagine that you want to track the average temperature in July. Each day you might take a measurement and average it with the readings of the previous days. You can maintain this running average and update it each day according to the 'prediction error' – the difference between your current estimate and the measured temperature that day. This approach, however, gives you no measure of the variability of the temperature – your estimate will be the same for a month with the same temperature every day and one in which the temperature fluctuates wildly around the same mean. Instead, imagine that a group of your friends performs the same learning procedure, but some of them are particularly influenced by days that are hotter than expected, others by surprisingly cold days. These idiosyncrasies in relative learning from positive versus negative prediction errors would produce either optimistic (warm) or pessimistic (cold) biases in their estimate of the mean, depending on how much the temperature varies from day to day. Collectively, the ensemble of biased temperature expectations across this group of friends would characterize the full temperature distribution for July.

Midbrain DA neurons have long been hypothesized to support a similar averaging process, called temporal difference (TD) learning, by conveying a prediction error for reward (RPE in place of temperature. This RPE is thought to allow the brain to learn to predict 'value' – the expected (i.e., mean) aggregate future reward – from trial-by-trial experience, and thereby guide choice. However, this classic model, according to which many neurons broadcast a common signal for learning a single quantity, has increasingly struggled with mounting evidence for all sorts of puzzling variation in DA responses. In a recent study, Dabney *et al.* [1] detail one such dimension of systematic variation: in the same way as your friends, different DA neurons reliably differ in their relative sensitivity to negative versus positive prediction errors. Dabney *et al.* reasoned that such unbalanced sensitivity should make the value estimates of the neurons, like those of your friends, span a range of variability-dependent biases from optimistic to pessimistic [2]. If so, the response of the population would implicitly represent not merely the mean reward but a set of 'expectiles' (an asymmetric generalization of the mean in the same way that a quantile is an asymmetric generalization of the median [3]), thus reflecting the entire distribution of outcomes used in the experiment (Figure 1). Indeed, after showing that the neurons complied with many specific predictions of this account, Dabney *et al.* [1] were ultimately able to reconstruct that distribution from the population responses. The success of this decoding shows that the spread of learning asymmetry across neurons is relatively broad, allowing representation of biased estimates that are far from the mean.

Would this information also be useful to the brain? And for what purpose? Classically, approaches to RL in artificial intelligence focus on predicting only the mean reward because this is all one needs to choose the action expected to be best on average. Dabney *et al.* relate the DA responses to a newer class of distributional RL algorithms [4] that work in the same way as your ensemble of friends to predict a large range of possible value estimates. Even if you still use only the mean for choice, the mere presence of such variation, in practice, seems to help deep neural networks to learn challenging tasks faster. However, it also raises the intriguing possibility of actually using pessimistic or optimistic estimates to guide choice – and even dynamically switching among them. The strategy of down- or up-weighting better or worse outcomes closely mirrors the classic formalism of risk sensitivity in economics, in which a chooser is more willing to gamble for a larger outcome if they overweight its value. Accordingly, distributional RL allows an engineer training a network to fly a drone to take risks in simulation, but then pilot more carefully when actually flying expensive equipment. There is also some intriguing evidence that biological organisms can modulate their risk sensitivity to circumstance: for instance, animals forage more desperately when in danger of starvation [5]. In human neuroimaging, links have also been demonstrated between asymmetric scaling of prediction errors and risk-sensitive choices, although so far only using a single prediction error instead of an ensemble [6]. Risk adjustment may also go wrong in mental illness: many diverse symptoms of anxiety disorders can be understood as reflecting pathologically pessimistic evaluation [7]. An exciting future direction for this work will be to examine whether the distributional response of DA neurons, shown here in the context of classical conditioning, might also relate to risk-sensitive choice behavior in the context of decision-making tasks.

Neurobiologically, the results also bring new data and a new formal perspective to bear on longstanding questions about the nature and degree of heterogeneity in

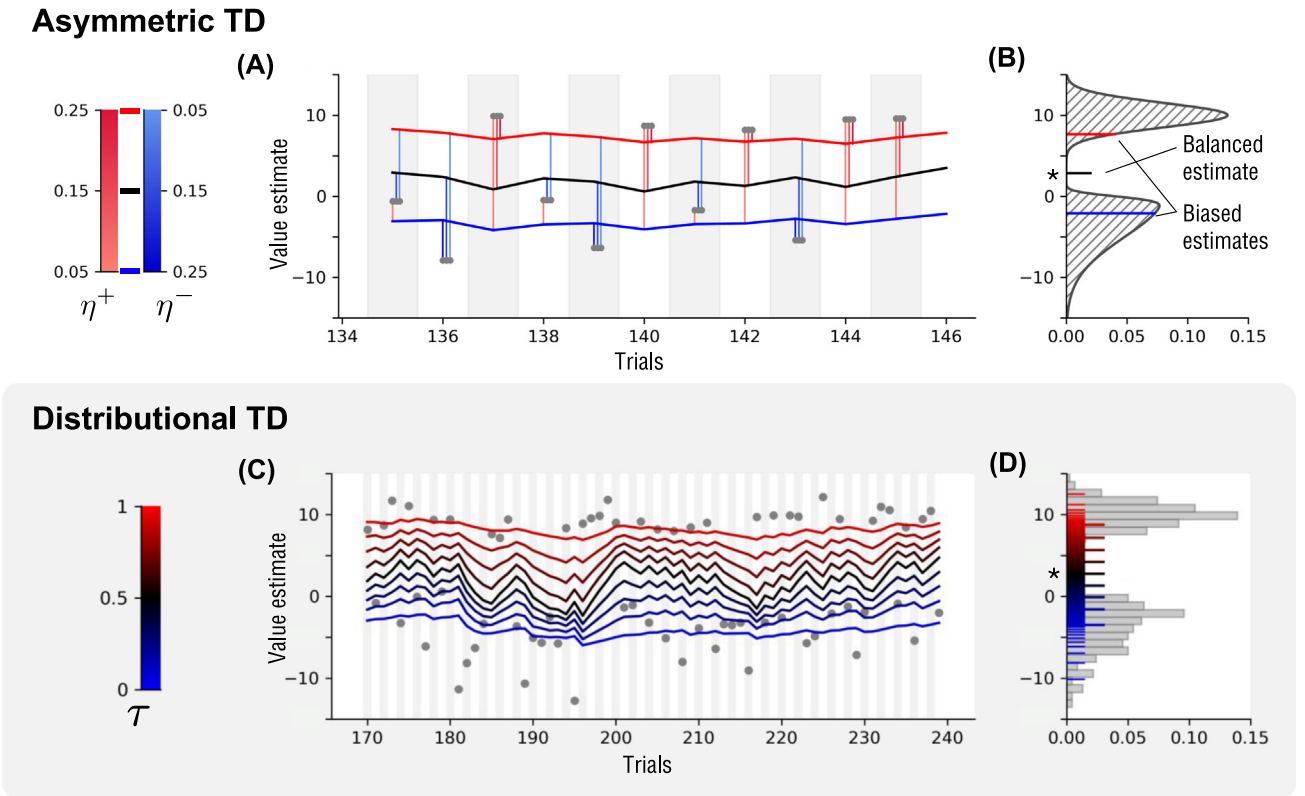

Figure 1. Distributional Temporal Difference (TD) Learning Is an Ensemble of Biased TD Learning Agents. Typically, TD learning assumes a single learning rate (here, η), thus the influence of positive and negative prediction errors is balanced, ensuring convergence of the value estimate to the mean return. (A) Asymmetric TD learning produces biased value estimates through different learning rates for positive and negative prediction errors. Divergence between optimistic (red) and pessimistic (blue) learners may be sufficient to produce prediction errors in opposite directions to the same outcome (for e.g., trial 138; outcomes in grey). Opposite phasic responses to identical reward amounts was one of the striking features of DA responses predicted by Dabney *et al.* according to distributional reinforcement learning (RL). (B) Classic TD learning (black) balances the influence of positive and negative prediction errors, and the value estimate will converge to the true mean of the distribution of outcomes (marked by a star). Asymmetric TD learning will converge to an expectile, the value at which positive and negative reward prediction errors (RPEs) are biased in proportion to the asymmetry of η^+ and η^-. (C) Distributional TD learning uses an ensemble of asymmetric TD learners that cover a range of learning rate biases $\left(\tau = \frac{\eta^+}{\eta^+ + \eta^-}\right)$, tracking many expectiles of the outcome distribution. (D) Distributed value estimates learned via distributional TD learning [here, 101 expectiles uniformly covering the range τ=[0,1)], compared with the observed distribution of outcomes (grey). A subset of biased value estimates corresponding to (C) are highlighted.

DA neuron responses, and the corresponding patterns of precision versus diffusion in their inputs and outputs [8]. In addition to recent reports of heterogeneity in the environmental and behavioral dimensions to which DA neurons respond [9], the current results establish variability in DA responses that relate to the single dimension of value. Perhaps surprisingly, the current data also indicate a notable degree of separation in the ensemble of feedback loops between these distinct value estimates and the prediction errors that train them. That is, there cannot be substantial crosstalk between value estimates learned from each pool of RPE-signaling DA neurons with a particular learning asymmetry, or the entire ensemble would collapse to the mean estimate predicted by classic TD learning. However, this implied separation actually raises a further computational puzzle because the full goal of TD learning is to predict not just a single number – such as the temperature or the single water bolus given the animals in each trial of these experiments – but is instead to chain these together in sequential steps to predict the long-run sum of future rewards.

Chaining together sequential predictions is what allows these algorithms (and, it is believed, the brain) to solve difficult real-world tasks such as mazes that involve multiple steps of choice. However, extending distributional RL from estimating a single outcome, as in this study, to the full problem of tracking variation over a sequence of outcomes actually does require

bringing together information across all the channels of the learning ensemble at each step of value update. It is as yet unclear how (or indeed whether) the DA system is wired up to square this circle. Perhaps one additional piece of the puzzle is that DA efflux in striatal target regions (thought to mediate learning-related plasticity) bears a complex relationship to spiking in DA neurons and features volumetric diffusion of released DA through the intracellular fluid [10]. All this provides another locus in the learning circuit in which functional asymmetry in prediction error-driven value updates, and anatomical crosstalk across channels, might arise. Whether a spectrum of biased value estimates arises through cellular or circuit mechanisms local to the ventral tegmental area, or is mediated through striatal, and perhaps even cortical, loops, the detailed results of Dabney *et al.* make it clear that DA responses are indeed able to maintain rich information about the experienced distribution of outcomes. Ultimately, whether and how this distributional information in DA might be used to guide decisions involving risk is the next test of the algorithm – to go beyond the average course of action and into the complex calculus of choice.

Acknowledgments

A.J.L. was supported by funding from the National Institutes of Health (National Institute on Drug Abuse R01DA042065) and the Templeton Foundation.

[1]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ 08544, USA

*Correspondence:
alangdon@princeton.edu (A.J. Langdon) and
ndaw@princeton.edu (N.D. Daw).

https://doi.org/10.1016/j.tics.2020.04.006

References

1. Dabney, W. *et al.* (2020) A distributional code for value in dopamine-based reinforcement learning. *Nature* 577, 671–675
2. Katahira, K. (2018) The statistical structures of reinforcement learning with asymmetric value updates. *J. Math. Psychol.* 87, 31–45
3. Waltrup, L.S. *et al.* (2014) Expectile and quantile regression – David and Goliath? *Stat. Modelling* 15, 433–456
4. Bellemare, M.G. *et al.* (2017) A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 449–458, ICML
5. Kacelnik, A. and Bateson, M. (2015) Risky theories – the effects of variance on foraging decisions. *Am. Zool.* 36, 402–434
6. Niv, Y. *et al.* (2012) Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* 32, 551–562
7. Zorowitz, S. *et al.* (2020) Anxiety, avoidance, and sequential evaluation. *Computat. Psychiatry* 4, 1–17
8. Tian, J. *et al.* (2016) Distributed and mixed information in monosynaptic inputs to dopamine neurons. *Neuron* 91, 1374–1389
9. Engelhard, B. *et al.* (2019) Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* 570, 509–513
10. Mohebi, A. *et al.* (2019) Dissociable dopamine dynamics for learning and motivation. *Nature* 570, 65–70

Forum

Seeing Visual Gamma Oscillations in a New Light

Eleonora Bartoli,[1] William Bosking,[1] and Brett L. Foster[1,2,*]

Gamma oscillations have been argued to support visual perception by synchronizing the processing and transfer of information within and across areas of visual cortex. Here, we highlight recent findings implicating the influence of color on visual gamma oscillations and how these observations may relate to local cortical tuning and organization.

Visually induced gamma oscillations are high-frequency (>30 Hz) fluctuations in electrical brain activity that have been proposed to support perception by synchronizing neural firing [1]. Here, we highlight recent explorations into the influence of stimulus color on gamma oscillations. Together with a novel discovery of visual circuit organization, they advance our understanding of the neural processes underlying visual gamma oscillations and their role in perception.

Historically, gamma oscillations have been studied by recording local field potentials (LFP) in early visual cortex in response to simple grating stimuli. Over the years, the properties of induced gamma oscillations have been shown to be dependent on low-level stimulus attributes. For example, large diameter and high-contrast grating stimuli increase the amplitude and peak frequency of induced gamma oscillations [2]. Prior work has linked the influence of grating size to surround-suppression phenomena, such that large-scale gamma oscillations occur under conditions where local inhibition is strong [3]. Indeed, computational models of gamma genesis are based on the interplay between excitation and inhibition [1], for which specific interneurons are critical in generating oscillatory population dynamics. Therefore, it is important to understand how these stimulus–response properties are engaged during natural vision, particularly because evidence for visual gamma oscillations in response to more complex image stimuli differs between studies and has been an area of debate [4,5].

Why are grating stimuli particularly effective at inducing gamma oscillations? Intuitively, this could be related to the well-known orientation tuning of cells in primary visual cortex (V1). However, effects seen at the single-cell level do not necessarily capture the full range of responses observed at the population LFP level. Recently, Hermes and colleagues [6] used a simple model to demonstrate that low variability of oriented edges falling within the receptive field predicts the occurrence of gamma oscillations in visual cortex to both gratings and natural images. These findings highlight the sensitivity of gamma oscillations to the spatial structure of visual stimuli, in particular spatially homogenous oriented edges, such as gratings, which