Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia

Brian Lau[1], Tiago Monteiro[2] and Joseph J Paton[2]

Computational models of reinforcement learning (RL) strive to produce behavior that maximises reward, and thus allow software or robots to behave adaptively [1]. At the core of RL models is a learned mapping between 'states' — situations or contexts that an agent might encounter in the world — and actions. A wealth of physiological and anatomical data suggests that the basal ganglia (BG) is important for learning these mappings [2,3]. However, the computations performed by specific circuits are unclear. In this brief review, we highlight recent work concerning the anatomy and physiology of BG circuits that suggest refinements in our understanding of computations performed by the basal ganglia. We focus on one important component of basal ganglia circuitry, midbrain dopamine neurons, drawing attention to data that has been cast as supporting or departing from the RL framework that has inspired experiments in basal ganglia research over the past two decades. We suggest that the parallel circuit architecture of the BG might be expected to produce variability in the response properties of different dopamine neurons, and that variability in response profile may not reflect variable functions, but rather different arguments that serve as inputs to a common function: the computation of prediction error.

Addresses
[1] Brain and Spine Institute, Paris, France
[2] Champalimaud Research, Lisbon, Portugal

Corresponding authors: Lau, Brian (brian.lau@upmc.fr), Paton, Joseph J (joe.paton@neuro.fchampalimaud.org)

Current Opinion in Neurobiology 2017, **46**:241–247

This review comes from a themed issue on **Computational neuroscience**

Edited by **Adrienne Fairhall** and **Christian Machens**

http://dx.doi.org/10.1016/j.conb.2017.08.015

Prediction errors

The BG consists of the striatum, the external and internal segments of the globus pallidus (GPe, GPi), the subthalamic nucleus (STN), and the substantia nigra pars reticulata (SNr). Information from a broad array of cortical territories arrives at the striatum and is sent directly, or indirectly through the GPe and STN, to the GPi and SNr [4], which in turn project to thalamic and brainstem nuclei to influence behavior. Midbrain DA neurons in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) project densely to striatum. Early recordings indicated that — by contrast to neurons in all other BG nuclei — DA neuron responses were not obviously linked to movement parameters [5,6]. Rather, they respond to stimuli and rewards in a manner consistent with encoding a reward prediction error (RPE) defined in temporal difference (TD) reinforcement learning algorithms [7]. Thus, DA neurons respond phasically to unpredicted rewards, and this response transfers to a predictive stimulus with learning [8]. Subsequent experiments demonstrated that DA responses quantitatively matched the predictions of TD models [9,10].

More recent work in mice has confirmed that, in the context of simple classical conditioning behavioral paradigms, the vast majority of DA neurons encode RPE [11], exhibiting phasic responses proportional to the arithmetic difference between expected and received reward [12,13]. Selective stimulation of dopamine neurons in both rodents and primates indicates that positive RPEs drive learning in accordance with TD models [14,15•,16]. Although DA neurons can exhibit pauses in activity on omission of predicted reward [17,18] or to aversive stimuli [19], their role as negative RPE is debated [20]. Recently, however, it was shown that transient optogenetic inhibition of DA neurons can produce behavioral changes consistent with the insertion of a negative RPE in the context of a modified pavlovian overexpectation paradigm [21]. In other behavioral contexts, dopamine neurons reflect prediction errors that integrate information about the expected timing of reward predicting cues [22,23••] as well as animal's varying belief about the context in which a prediction error has arisen [24,25]. In the more complex scenario of sensory guided decision-making, phasic dopamine responses reflect stimulus difficulty, and trial by trial variability in dopamine responses is correlated with judgments, suggesting that DA neurons compute prediction error using the same sensory representation the subject uses to guide decisions [23••,24,26]. Taken together, the evidence across species and behavioral contexts indicates that dopamine neurons encode TD prediction errors that can drive reinforcement

learning. But is it the only type of prediction error dopamine neurons encode?

TD models learn the long-run value of future events in a model-free way, without storing information about the outcomes themselves. This is computationally simple, but raises the question of how dopamine figures into behaviors that are model-based, relying on knowledge of the environment and predicted events and outcomes. Experiments designed to distinguish between model-free and model-based behavior indicate that dopamine also seems to play an important role in model-based behavior [27–29]. DA neurons encode inferred value in a task where reward contingencies switch in blocks and reward obtained from one stimulus predicts whether reward is associated with a second stimulus [30]. Inferred relationships between stimuli can also be demonstrated using sensory preconditioning, which starts with pairing two neutral stimuli (A and B, say) in the absence of reward. If B is then paired with reward, presenting A in isolation produces conditioned responses despite A never being rewarded. The inference that A is linked to reward requires integrating information over separate experiences. Recordings from VTA show that the same DA neurons that encode model-free values (phasic response to B) also encode the model-based values of A [31••]. Moreover, optogenetic inhibition of VTA DA neurons disrupts stimulus–stimulus learning between A and B, and optogenetic activation reinstates learning between A and B that has been behaviorally blocked [32••].

Taken together, these data indicate that dopamine neurons can represent more complex prediction errors than previously appreciated. Indeed, richer behavioral experiments are revealing that DA computations extending beyond evaluating obtained rewards may be the rule rather than the exception. In choice tasks, DA responses were previously shown to correlate with TD prediction errors associated with the chosen action [33]. Recently, DA responses have also been shown to reflect the value of unchosen actions [34,35], although with a slightly delayed timing suggesting that different pathways may be involved in this computation [34]. In singing birds, VTA DA neurons encode a novel performance error representing the difference between sensory feedback and an internal representation of a song [36••]. This last study seems to represent clear evidence of DA neurons encoding an error between a predicted sequence derived from an internal model [37] and the behavioral sequence that is produced and experienced.

. . . and beyond?

Although prediction errors go a long way towards explaining dopamine response properties, some observations do not fit so neatly into the prediction error framework. An example is the short-latency phasic responses some DA neurons produce to novel or sufficiently intense stimuli [38–40]. These responses are difficult to explain in a prediction error framework since they can be elicited by stimuli that are not associated with rewards, and do not explicitly predict anything. Recent work suggests that DA neurons with these responses are distinct and serve a function separate from those computing prediction errors. Menegas *et al.* recorded dopamine terminal responses throughout the striatum in mice and found that DA axon responses in the ventral striatum did not respond to novel stimuli until they were paired with reward, after which these responses resembled classical TD errors [41••]. By contrast, DA axon responses in the posterior tail of the striatum responded strongly to novel stimuli but did not encode reward prediction errors. These authors previously showed that DA neurons projecting to the posterior tail of the striatum represent a distinct class based on the inputs they receive [42], suggesting that novelty and RPE are computed in different circuits. A similar distinction is observed in primates, where some neurons in the SNc respond to novel visual stimuli but not reward, while another group do not respond to novel visual stimuli but instead respond like classical TD error neurons [43,44]. These two groups of neurons are spatially segregated, with the novelty-responding neurons projecting to the tail of the caudate. The match between species is not exact, as in primates the novelty-responding neurons did in fact acquire the value of the stimuli (apparently without extinguishing them), which does not seem to be the case in mice.

Another example of DA signals that do not square so obviously with prediction error are tonic or quasi-tonic signals such as ramps or other more sustained activity profiles that seem to integrate information over longer timescales than the phasic bursts associated with prediction errors. Tonic modulation of DA has been suggested to be important for modulating vigor, which may correspond to the frequency of actions in a free operant setting [45], or intensity of movement itself [46,47]. Voltammetric recordings of DA release in the nucleus accumbens display ramps of activity reflecting the proximity and value of rewarding goals in mazes [48], and the average reward over trials or expected future reward on slow and fast timescales, respectively, during a value-based decision-making task [49]. It remains unclear whether such signals represent a challenge to the general prediction error framework. It is possible that the dynamics of DA reuptake and clearance from the extracellular space [50] are such that extracellular DA levels produced by phasic RPEs encoded in the spiking activity of DA neurons correspond to average reward rates or expected future reward. In addition, variants of TD learning models can reproduce some of these findings [51,52], and these data may provide specific predictions that can be used to test how specific RL algorithms are implemented in the brain.

Lastly, although the encoding of movement parameters was largely abandoned as a significant predictor of DA