Ramping and State Uncertainty in the Dopamine Signal

John G. Mikhael[1,2*], HyungGoo R. Kim[3], Naoshige Uchida[3], Samuel J. Gershman[4]

[1]Program in Neuroscience, Harvard Medical School, Boston, MA 02115

[2]MD-PhD Program, Harvard Medical School, Boston, MA 02115

[3]Department of Molecular and Cellular Biology and Center for Brain Science, Harvard University, Cambridge, MA 02138

[4]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138

*Corresponding Author

Correspondence: john_mikhael@hms.harvard.edu

Abstract

Reinforcement learning models of the basal ganglia map the phasic dopamine signal to reward prediction errors (RPEs). Conventional models assert that, when a stimulus reliably predicts a reward with fixed delay, dopamine activity during the delay period and at reward time should converge to baseline through learning. However, recent studies have found that dopamine exhibits a gradual ramp before reward in certain conditions even after extensive learning, such as when animals are trained to run to obtain the reward, thus challenging the conventional RPE models. In this work, we begin with the limitation of temporal uncertainty (animals cannot perfectly estimate time to reward), and show that sensory feedback, which reduces this uncertainty, will cause an unbiased learner to produce RPE ramps. On the other hand, in the absence of feedback, RPEs will be flat after learning. These results reconcile the seemingly conflicting data on dopamine behaviors under the RPE hypothesis.

Keywords: dopamine, ramping, reinforcement learning, reward prediction error, state value, state uncertainty, sensory feedback

Introduction

Perhaps the most successful convergence of reinforcement learning theory with neuroscience has been the insight that the phasic activity of midbrain dopamine (DA) neurons tracks 'reward prediction errors' (RPEs), or the difference between received and expected reward (Schultz et al., 1997; Schultz, 2007a; Glimcher, 2011). In reinforcement learning algorithms, RPEs serve as teaching signals that update an agent's estimate of rewards until those rewards are well-predicted. In a seminal experiment, Schultz et al. (1997) recorded from midbrain DA neurons in primates and found that the neurons responded with a burst of activity when an unexpected reward was delivered. However, if a reward-predicting cue was available, the DA neurons eventually stopped responding to the (now expected) reward and instead began to respond to the cue, much like an RPE (see Results). This finding formed the basis for the RPE hypothesis of DA.

Over the past two decades, a large and compelling body of work has supported the view that phasic DA functions as a teaching signal (Schultz et al., 1997; Niv and Schoenbaum, 2008; Glimcher, 2011; Steinberg et al., 2013; Eshel et al., 2015). In particular, phasic DA activity has been shown to track the RPE term of temporal difference (TD) learning models, which we review below, remarkably well (Schultz, 2007a). However, recent results have called this model of DA into question. Using fast-scan cyclic voltammetry in rat striatum during a goal-directed spatial navigation task, Howe et al. (2013) observed a ramping phenomenon—

a steady increase in DA over the course of a single trial—that persisted even after extensive training. Since then, DA ramping has been observed during a two-armed bandit task (Hamid et al., 2016) and during the execution of self-initiated action sequences (Collins et al., 2016). At first glance, these findings appear to contradict the RPE hypothesis of DA. Indeed, why would error signals persist (and ramp) after a task has been well-learned? Perhaps, then, instead of reporting an RPE, DA should be reinterpreted as reflecting the value of the animal's current state, such as its position during reward approach (Hamid et al., 2016). Alternatively, perhaps DA signals different quantities in different tasks, e.g., value in operant tasks, in which the animal must act to receive reward, and RPE in classical conditioning tasks, in which the animal need not act to receive reward.

To distinguish among these possibilities, Kim et al. (2019) recently devised an experimental paradigm that dissociates the value and RPE interpretations of DA. As we show in the Methods, RPEs in the experiments considered above can be approximated as the derivative of value under the TD learning framework. By training mice on a virtual reality environment and manipulating various properties of the task—namely, the speed of scene movement, teleports, and temporary pauses at various locations—the authors could dissociate spatial navigation from locomotion and make precise predictions about how value should change vs. how its derivative (RPE) should change. The authors found that mice continued to display ramping DA signals during the task even without locomotion, and that the changes in DA behaviors were consistent with the RPE hypothesis and not with the value interpretation.

The body of experimental studies outlined above produce a number of unanswered questions regarding the function of DA: First, why would an error signal persist once an association is well-learned? Second, why would it ramp over the duration of the trial? Third, why would this ramp occur in some tasks but not others? Does value (and thus RPE) take different functional forms in different tasks, and if so, what determines which forms result in a ramp and which do not? Here we address these questions from normative principles.

In this work, we examine the influence of sensory feedback in guiding value estimation. Because of irreducible temporal uncertainty, animals not receiving sensory feedback (and therefore relying only on internal timekeeping mechanisms) will have corrupted value estimates regardless of how well a task is learned. In this case, value functions will be "blurred" in proportion to the uncertainty at each point. Sensory feedback, however, reduces this blurring as each new timepoint is approached. Beginning with the normative principle that animals seek to best learn the value of each state, we show that unbiased learning, in the presence of feedback, requires RPEs that ramp. These ramps scale with the informativeness of the feedback (i.e., the reduction in uncertainty), and at the extreme, absence of feedback leads to flat RPEs. Thus we show that

differences in a task's feedback profile explain the puzzling collection of DA behaviors described above.

We will begin the next section with a review of the TD learning algorithm, then examine the effect of state uncertainty on value learning. We will then show how, by reducing state uncertainty without biasing learning, sensory feedback causes the RPE to reproduce the experimentally observed behaviors of DA.

Results

Temporal Difference Learning

In TD learning, an agent transitions through a sequence of states according to a Markov process (Sutton, 1988). The value associated with each state is defined as the expected discounted future return:

$$V_t = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right], \tag{1}$$

where t denotes time and indexes states, r_t denotes the reward delivered at time t, and $\gamma \in (0, 1)$ is a discount factor. In the experiments we will examine, a single reward is presented at the end of each trial. For these cases, Equation (1) can be written simply as:

$$V_t = \gamma^{T-t} r, \tag{2}$$

for all $t \in [0, T]$, where r is the magnitude of reward delivered at time T. In words, value increases exponentially as reward time T is approached, peaking at a value of r at T (Figure 1B,D). Additionally, note that exponential functions are convex: The convex shape of the value function will be important in subsequent sections (see Kim et al. (2019) for experimental verification).

How does the agent learn this value function? Under the Markov property, Equation (1) can be rewritten as:

$$V_t = r_t + \gamma V_{t+1}, \tag{3}$$

which is referred to as the Bellman equation (Bellman, 1957). The agent approximates V_t with $\hat{V}_t$, which is updated in the event of a mismatch between the estimated value and the reward actually received. By

analogy with Equation (3), this mismatch (the RPE) can be written as:

$$\delta_t = r_t + \gamma \hat{V}_{t+1} - \hat{V}_t. \tag{4}$$

When δ_t is zero, Equation (3) has been well-approximated. However, when δ_t is positive or negative, $\hat{V}_t$ must be increased or decreased, respectively:

$$\hat{V}_t^{(t+1)} = \hat{V}_t^{(t)} + \alpha \delta_t^{(t)}, \tag{5}$$

where $\alpha \in (0, 1)$ denotes the learning rate, and the superscript denotes the learning step. Learning will progress until $\delta_t = 0$ on average. After this point, $\hat{V}_t = \gamma^{T-t} r$ on average, which is precisely the true value. (See the Methods for a more general description of TD learning and its neural implementation.)

Having described TD learning in the simplified case where the agent has a perfect internal clock and thus no state uncertainty, let us now examine how state uncertainty affects learning, and how this uncertainty is reduced with sensory feedback.

Value Learning Under State Uncertainty

Because animals do not have perfect internal clocks, they do not have complete access to the true time t (Gibbon, 1977; Church and Meck, 2003; Staddon, 1965). Instead, t is a latent state corrupted by timing noise, often modeled as follows:

$$\tau \sim \mathcal{N}(t, \sigma_t^2), \tag{6}$$

where τ is subjective (internal) time, drawn from a distribution centered on objective time t, with some standard deviation σ_t. We take this distribution to be Gaussian for simplicity (an assumption we relax in the Methods). Thus the subjective estimate of value $\hat{V}_\tau$ is an average over the estimated values $\hat{V}_t$ of each state t:

$$\hat{V}_\tau = \sum_t p(t|\tau)\ \hat{V}_t, \tag{7}$$

where $p(t|\tau)$ denotes the probability that t is the true state given the subjective measurement τ, and thus represents state uncertainty. We refer to this quantity as the uncertainty kernel (Figure 1A,C). Intuitively, $\hat{V}_\tau$ is the result of blurring $\hat{V}_t$ proportionally to the uncertainty kernel (Methods).

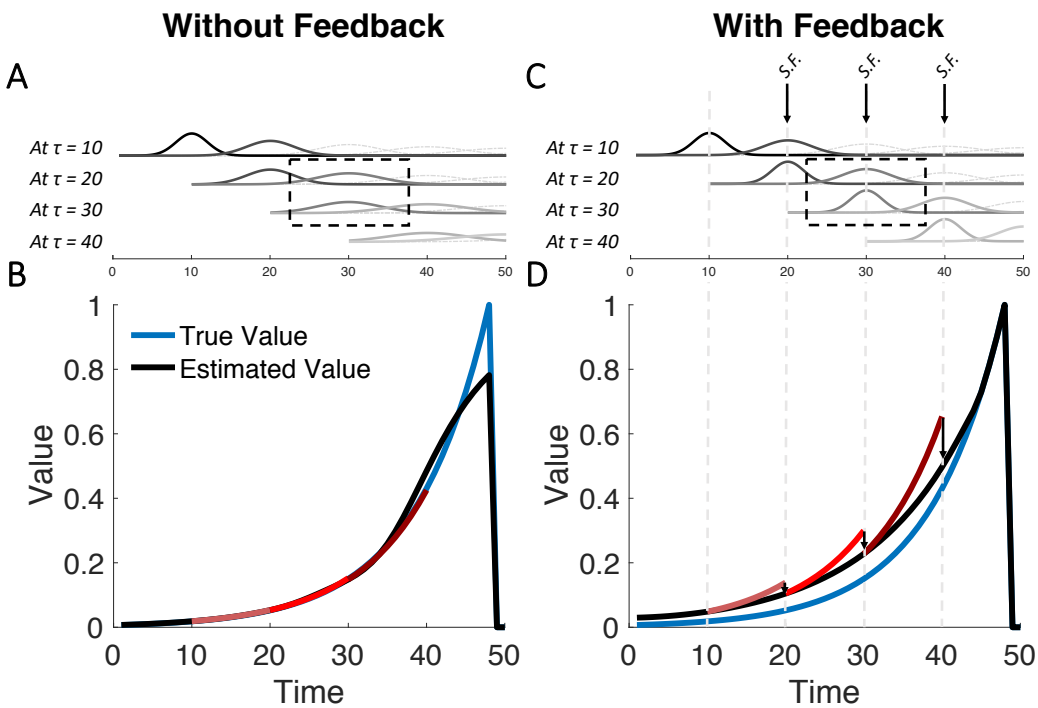

Figure 1: Sensory Feedback Biases Value Learning. (A) Illustration of state uncertainty in the absence of sensory feedback. Each row includes the uncertainty kernels at the current state and the next state (solid curves). Lighter gray curves represent uncertainty kernels for later states. Thus, similarly colored kernels on different rows represent uncertainty kernels for the same state, but evaluated at different timepoints (e.g., dashed box). In the absence of feedback, state uncertainty for a single state does not acutely change across time (compare with C). (B) Without feedback, value is unbiased on average. Red curves represent the predicted increase in value between the current state and the next state (10 and 20 for light red; 20 and 30 for red; 30 and 40 for brick red). After learning, this roughly equals an increase by γ^{-1} on average. (C) Sensory feedback reduces state uncertainty. Three instances of feedback are shown for illustration (*S.F.*; arrows). Note here that the kernels used to estimate value at the same state have different widths depending on whether they were evaluated before or after feedback. This results in different value estimates being used to compute the RPE at the current state and at the next state (Equations (8) and (9)). (D) As a result of sensory feedback, value at each state will be estimated based on an inflated version of value at the next state. Hence, after learning (when RPE is zero on average), estimated value will be systematically larger than true value. Red curves represent the predicted increase in value between the current state and the next state. After learning, this roughly equals an increase by γ^{-1} on average. See Methods for simulation details.

After learning (i.e., when the RPE is zero on average), the estimated value at every state will be roughly the estimated value at the next state, discounted by γ, on average (black curve in Figure 1B). A key requirement for this unbiased learning can be discovered by writing the RPE equations for two successive states:

$$\delta_\tau = r_\tau + \gamma \hat{V}_{\tau+1} - \hat{V}_\tau \tag{8}$$

$$\delta_{\tau+1} = r_{\tau+1} + \gamma \hat{V}_{\tau+2} - \hat{V}_{\tau+1}. \tag{9}$$

Notice here that $\hat{V}_{\tau+1}$ is represented in both equations. Thus, for value to be well-learned, a requirement

is that $\hat{V}_{\tau+1}$ not acutely change during the interval after computing δ_τ and before computing $\delta_{\tau+1}$. This requirement extends to changes in the uncertainty kernels: By Equation (7), if the kernel $p(t|\tau+1)$ were to be acutely updated due to information available at $\tau+1$ but not at τ, then $\hat{V}_{\tau+1}$ will acutely change as well. This means that $\hat{V}_\tau$ will be discounted based on $\hat{V}_{\tau+1}$ before feedback (i.e., as estimated at τ; red curves in Figure 1D) rather than $\hat{V}_{\tau+1}$ after feedback (i.e., as estimated at $\tau+1$; black curve). In the next section, we examine this effect more precisely.

Value Learning in the Presence of Sensory Feedback

How is value learning affected by sensory feedback? As each time τ is approached, state uncertainty is reduced due to sensory feedback (arrows in Figure 1C). This is because at timepoints preceding τ, the estimate of what value will be at τ is corrupted by both temporal noise and the lower-resolution stimuli associated with τ. Approaching τ in the presence of sensory feedback reduces this corruption. This, however, means that $\hat{V}_{\tau+1}$ will be estimated differently while computing δ_τ and $\delta_{\tau+1}$ (Equations (8) and (9); compare widths of similarly colored kernels beneath each arrow in Figure 1C), which in turn results in biased value learning.

To examine the nature of this bias, we note that averaging over a convex value function results in overestimation of value. Intuitively, convex functions are steeper on the right (larger values) and shallower on the left (smaller values), so averaging results in a bias toward larger values. Furthermore, wider kernels result in greater overestimation (Methods). Thus upon entering each new state, the reduction of uncertainty via sensory feedback will acutely mitigate this overestimation, resulting in different estimates $\hat{V}_{\tau+1}$ being used for δ_τ and $\delta_{\tau+1}$. Left uncorrected, the value estimate will be systematically biased, and in particular, value will be overestimated at every point (Figure 2A; Methods). An intuitive way to see this is as follows: The objective of the TD algorithm (in this simplified task setting) is for the value at each state τ to be γ times smaller than the value at $\tau+1$ by the time the RPE converges to zero (Equation (2)). If an animal systematically overestimates value at the next state, then it will overestimate value at the current state as well (even if sensory feedback subsequently diminishes the next state's overestimation). Thus the "wrong" value function is learned (Figure 2A,B).

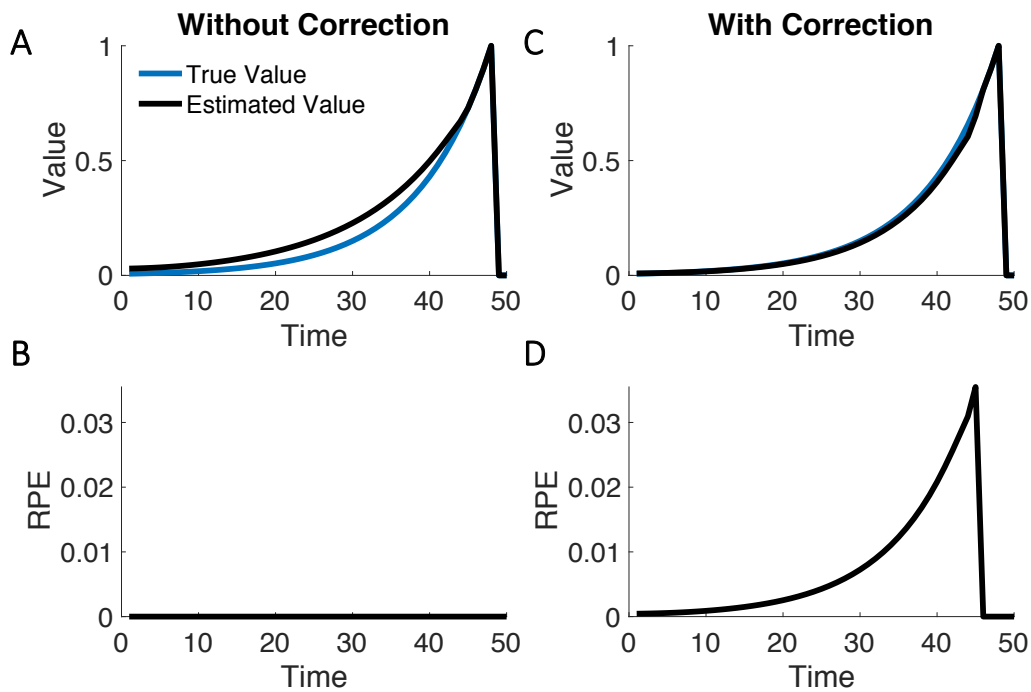

Figure 2: Unbiased Learning in the Presence of Feedback Leads to RPE Ramps. (A) Value at each state is learned according to an overestimated version of value at the next state. Thus, a biased value function is learned (see Figure 1D). (B) After learning, the RPE converges to zero. (C) With a correction term, the correct value function is learned. (D) The cost of forcing an unbiased learning of value is a persistent RPE. Intuitively, value at the current state is not influenced by the overestimated version of value at the next state (compare with A,B). By Equation (13), this results in RPEs that ramp. See Methods for simulation details.

To overcome this bias, an optimal agent must correct the just-computed RPE as sensory feedback becomes available. In the Methods, we show that this correction can simply be written as:

$$\hat{V}_t^{(t+1)} = \hat{V}_t^{(t)} + \alpha\delta_\tau^{(t)} p(t|\tau) - \beta\hat{V}_\tau^{(t)} p(t|\tau) \tag{10}$$

$$\simeq \hat{V}_t^{(t)} + \alpha\delta_\tau^{(t)} p(t|\tau) - \beta\hat{V}_t^{(t)}, \tag{11}$$

where the approximate equality holds for sufficient reductions in state uncertainty due to feedback, and

$$\beta = \alpha\left(\exp\left[\frac{(\ln\gamma)^2(l^2 - s^2)}{2}\right] - 1\right). \tag{12}$$

Here, the uncertainty kernel of $\hat{V}_{\tau+1}$ has some standard deviation l at τ and a smaller standard deviation s at $\tau + 1$. In words, as the animal gains an improved estimate of $\hat{V}_{\tau+1}$, it corrects the previously computed δ_τ with a feedback term to ensure unbiased learning of value (Figure 2C). Notice here that the correction term is a function of the reduction in variance $(l^2 - s^2)$ due to sensory feedback. In the absence of feedback,

the reduction in variance is zero (the uncertainty kernel for $\tau + 1$ cannot be reduced during the transition from τ to $\tau + 1$), which means $\beta = 0$.

How does this correction affect the RPE? By Equation (10), the RPE will converge to:

$$\delta_\tau = \frac{\beta}{\alpha} \hat{V}_\tau. \tag{13}$$

Therefore, with sensory feedback, the RPE ramps and tracks $\hat{V}_\tau$ in shape (Figure 2D). In the absence of feedback, $\beta = 0$; thus, there is no ramp.

In summary, when feedback is provided with new states, value learning becomes miscalibrated, as each value point will be learned according to an overestimated version of the next (Figure 2A). With a subsequent correction of this bias, the agent will continue to overestimate the RPEs at each point (RPEs will ramp; Figure 2D), in exchange for learning the correct value function (Figure 2C).

Relationship with Experimental Data

In classical conditioning tasks without sensory feedback, DA ramping is not observed (Schultz et al., 1997; Kobayashi and Schultz, 2008; Stuber et al., 2008; Flagel et al., 2011; Cohen et al., 2012; Hart et al., 2014; Eshel et al., 2015; Menegas et al., 2015, 2017; Babayan et al., 2018) (Figure 3A). On the other hand, in goal-directed navigation tasks, characterized by sensory feedback in the form of salient visual cues as well as locomotive cues (e.g., joint movement), DA ramping is present (Howe et al., 2013) (Figure 3C). DA ramping is also present in classical conditioning tasks that do not involve locomotion but that include either spatial or non-spatial feedback (Kim et al., 2019), as well as in two-armed bandit tasks (Hamid et al., 2016) and when executing self-initiated action sequences (Wassum et al., 2012; Collins et al., 2016).

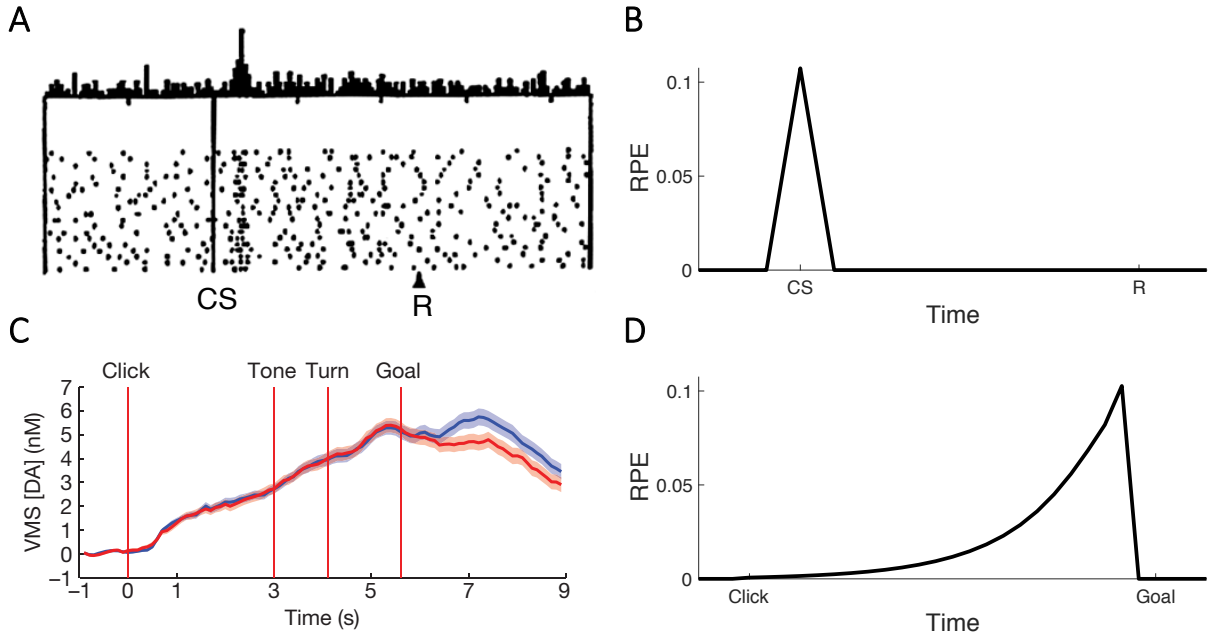

Figure 3: Differences in Feedback Result in Different RPE Behaviors. (A) Schultz et al. (1997) have found that after learning, phasic DA responses to a predicted reward (R) diminish, and instead begin to appear at the earliest reward-predicting cue (conditioned stimulus; CS). Figure from Schultz et al. (1997). (B) Our derivations recapitulate this result. In the absence of sensory feedback, RPEs converge to zero. (C) Howe et al. (2013) have found that the DA signal ramps during a well-learned navigation task over the course of a single trial. Figure from Howe et al. (2013). (D) Our derivations recapitulate this result. In the presence of sensory feedback, RPEs track the shape of the estimated value function. See Methods for simulation details.

As described in the previous section, sensory feedback—due to external cues or to the animal's own movement—can reconcile both types of DA behaviors with the RPE hypothesis: In the absence of feedback, there is no reduction in state uncertainty upon entering each new state ($\beta = 0$), and therefore no ramps (Equation (13); Figure 3B). On the other hand, when state uncertainty is reduced as each state is entered, ramps will occur (Figure 3D).

More generally, our results demonstrate that a measured DA signal whose shape tracks with estimated value need not be evidence against the RPE hypothesis of DA, contrary to some claims (Hamid et al., 2016; Berke, 2018): Indeed, in the presence of sensory feedback, δ_τ and $\hat{V}_\tau$ have the same shape. Thus, our derivation is conceptually compatible with the value interpretation of DA under certain circumstances, but importantly, this derivation captures the experimental findings in other circumstances in which the value interpretation fails.

Discussion

The role of DA in reinforcement learning has long been studied. While a large body of work has established phasic DA as an error signal (Schultz et al., 1997; Niv and Schoenbaum, 2008; Glimcher, 2011; Steinberg et al., 2013; Eshel et al., 2015), more recent work has questioned this view (Wassum et al., 2012; Howe et al., 2013; Hamid et al., 2016; Collins et al., 2016). Indeed, in light of persistent DA ramps occurring in certain tasks even after extensive learning, some authors have proposed that DA may instead communicate value itself in these tasks (Hamid et al., 2016). However, the determinants of DA ramps have remained unclear: Ramps are observed during goal-directed navigation, in which animals must run to receive reward (operant tasks; Howe et al., 2013), but can also be elicited in virtual reality tasks in which animals do not need to run for reward (classical conditioning tasks; Kim et al., 2019). Within classical conditioning, DA ramps can occur in the presence of navigational or non-navigational stimuli indicating time to reward (Kim et al., 2019). Within operant tasks, ramps can be observed in the period preceding the action (Totah et al., 2013) as well as during the action itself (Howe et al., 2013). These ramps are furthermore not specific to experimental techniques and measurements, and can be observed in cell body activities, axonal calcium signals, and in the DA concentrations (Kim et al., 2019).

We have shown in this work that under the RPE hypothesis of DA, sensory feedback may control the different observed DA behaviors: In the presence of sensory feedback, RPEs track the estimated value in shape (ramps), but they remain flat in the absence of feedback (no ramps). Thus DA ramps and phasic responses follow from common computational principles and may be generated by common neurobiological mechanisms.

Our derivation makes a number of testable predictions. In particular, our results predict that any type of information that reduces state uncertainty—for example, an auditory tone whose frequency reflects time to reward or a moving visual stimulus whose position reflects time to reward—will result in a DA ramp, and furthermore, the magnitude of the ramp will increase with the informativeness of the stimulus (i.e., with a greater reduction in state uncertainty; Equations (12) and (13)). Therefore, in trials where the change in the tone's frequency is less apparent, or the contrast of the visual stimulus is lower, the ramp will be blunted. At the extreme, when the tone's frequency does not change with time or the contrast is minimal, no ramp will be observed. At this point, the task is indistinguishable from the classical conditioning experiments of Schultz et al. (1997) discussed in the Introduction.

Our work takes inspiration from previous studies that examined the role of state uncertainty in DA responses

(Kobayashi and Schultz, 2008; Fiorillo et al., 2008; de Lafuente and Romo, 2011; Starkweather et al., 2017; Lak et al., 2017). For instance, temporal uncertainty increases with longer durations (Staddon, 1965; Gibbon, 1977; Church and Meck, 2003). This means that in a classical conditioning task, DA bursts at reward time will not be completely diminished, and will be larger for longer durations, as Kobayashi and Schultz (2008) and Fiorillo et al. (2008) have observed. Similarly, Starkweather et al. (2017) have found that in tasks with uncertainty both in *whether* reward will be delivered as well as *when* it is delivered, DA exhibits a prolonged dip (i.e., a negative ramp) leading up to reward delivery. Here, value initially increases as expected reward time is approached, but then begins to slowly decrease as the probability of reward delivery during the present trial becomes less and less likely, resulting in persistently negative prediction errors (see also Starkweather et al., 2018; Babayan et al., 2018). As the authors of these studies note, both results are fully predicted by the RPE hypothesis of DA. Hence, state uncertainty, due to noise either in the internal circuitry or in the external environment, is reflected in the DA signal.

A number of questions arise from our analysis. First, is there any evidence to support the benefits of learning the 'true' value function as written in Equation (2) (Figure 2C) over the biased version of value (Figure 2A)? We note here that under the normative account, the agent seeks to learn *some* value function that maximizes its well-being. Our key result is that this function—regardless of its exact shape—will not be learned well if feedback is delivered during learning, unless correction ensues. While we have chosen the exponential shape in Equation (2) after the conventional TD models, our results extend to any convex value function. Second, due to this presumed exponential shape, the ramping behaviors resulting from our analysis also look exponential, rather than linear (compare with experimental results). We nonetheless have chosen to remain close to conventional TD models and purely exponential value functions for ease of comparison with the existing theoretical literature. Perhaps equally important, the relationship between RPE and its neural correlate need only be monotonic and not necessarily equal. In other words, a measured linear signal does not necessarily imply a linear RPE, and a convex neural signal need not communicate convex information.

Third, while we have derived RPE ramping from normative principles, it is important to note that biases in value learning may also produce ramping. For instance, one earlier proposal by Gershman (2014) was that value may take a fixed convex shape in spatial navigation tasks; the mismatch between this shape and the exponential shape in Equation (2) produces a ramp (see Methods for a general derivation of the conditions for a ramp). Morita and Kato (2014), on the other hand, posited that value updating involves a decay term. Assuming such a decay term results in a relationship qualitatively similar to that in Equation (10), and thus RPE ramping (see also implementations in Mikhael and Bogacz, 2016; Cinotti et al., 2019). Ramping can similarly be explained by assuming temporal or spatial bias that decreases with approach to the reward, by

modulating the temporal discount term during task execution, or by other mechanisms (see Supplemental Information for derivations). In each of these proposals, ramps emerge as a 'bug' in the implementation, rather than as an optimal strategy for unbiased learning. These proposals furthermore do not explain the different DA patterns that emerge under different paradigms. Finally, it should be noted that we have not assumed any modality- or task-driven differences in learning (any differences in the shape of the RPE follow solely from the sensory feedback profile), although in principle, different value functions may certainly be learned in different types of tasks.

Alternative accounts of DA ramping that deviate more significantly from our framework have also been proposed. In particular, Lloyd and Dayan (2015) have provided three compelling theoretical accounts of ramping. In the first account, the authors show that within an actor-critic framework, uncertainty in the communicated information between actor and critic regarding the timing of action execution may result in a monotonically increasing RPE leading up to the action. In the second account, ramping modulates gain control for value accumulation within a drift-diffusion model (e.g., by modulating neuronal excitability (Nicola et al., 2000)). Under this framework, fluctuations in tonic and phasic DA produce average ramping. The third account extends the average reward rate model of tonic DA proposed by Niv et al. (2007). In this extended view, ramping constitutes a 'quasi-tonic' signal that reflects discounted vigor. The authors show that the discounted average reward rate follows $(1-\gamma)V$, and hence takes the shape of the value function in TD learning models. Finally, Howe et al. (2013) have proposed that these ramps may be necessary for sustained motivation in the operant tasks considered. Indeed, the notion that DA may serve multiple functions beyond the communication of RPEs is well-motivated and deeply ingrained (Schultz, 2007b, 2010; Berridge, 2007; Frank et al., 2007; Gardner et al., 2018). Our work does not necessarily invalidate these alternative interpretations, but rather shows how a single RPE interpretation can embrace a range of apparently inconsistent phenomena.

Methods

Temporal Difference Learning and Its Neural Correlates

Under TD learning, each state is determined by task-relevant contextual cues, referred to as features, that predict future rewards. For instance, a state might be determined by an internal estimate of time or perceived distance from a reward. We model the agent as approximating V_t by taking a linear combination

of the features (Schultz et al., 1997; Ludvig et al., 2008, 2012):

$$\hat{V}_t = \sum_d w_d x_{d,t}, \tag{14}$$

where $\hat{V}_t$ denotes the estimated value at time t, and $x_{d,t}$ denotes the dth feature at t. The learned relevance of each feature x_d is reflected in its weight w_d, and the weights are updated in the event of a mismatch between the estimated value and the rewards actually received. The update occurs in proportion to each weight's contribution to the value estimate at t:

$$w_d^{(t+1)} = w_d^{(t)} + \alpha \delta_t^{(t)} x_{d,t}, \tag{15}$$

where $\alpha \in (0, 1)$ denotes the learning rate, and the superscript denotes the learning step. In words, when a feature x_d does not contribute to the value estimate at t ($x_{d,t} = 0$), its weight is not updated. On the other hand, weights corresponding to features that do contribute to $\hat{V}_t$ will be updated in proportion to their activations at that time. This update rule is referred to as gradient ascent ($x_{d,t}$ is equal to the gradient of $\hat{V}_t$ with respect to the weight w_d), and it implements a form of credit assignment, in which the features most activated at t undergo the greatest modification to their weights.

In this formulation, the basal ganglia implements the TD algorithm termwise: Cortical inputs to striatum encode the features $x_{d,t}$, corticostriatal synaptic strengths encode the weights w_d (Houk et al., 1995; Montague et al., 1996), phasic activity of midbrain dopamine neurons encodes the error signal δ_t (Schultz et al., 1997; Niv and Schoenbaum, 2008; Glimcher, 2011; Steinberg et al., 2013; Eshel et al., 2015), and the output nuclei of the basal ganglia (substantia nigra pars reticulata and internal globus pallidus) encode estimated value $\hat{V}_t$ (Ratcliff and Frank, 2012).

We have implicitly assumed in the Results a maximally flexible feature set, the complete serial compound representation (Moore et al., 1989; Sutton and Barto, 1990; Montague et al., 1996; Schultz et al., 1997), in which every time step following trial onset is represented as a separate feature. In other words, the feature $x_{d,t}$ is 1 when $t = d$ and 0 otherwise. In this case, value at each timepoint is updated independently of the other timepoints, and each has its own weight. It follows that $\hat{V}_t = w_t$, and we can write Equation (15) directly in terms of $\hat{V}_t$, as in Equation (5).

Value Learning Under State Uncertainty

Animals only have access to subjective time, and must infer objective time given the corruption in Equation (6). The RPE is then:

$$\delta_\tau = r_\tau + \gamma \hat{V}_{\tau+1} - \hat{V}_\tau, \tag{16}$$

and this error signal is used to update the value estimates at each point t in proportion to its posterior probability $p(t|\tau)$:

$$\hat{V}_t^{(t+1)} = \hat{V}_t^{(t)} + \alpha \delta_\tau^{(t)} p(t|\tau). \tag{17}$$

Said differently, the effect of state uncertainty is that when the error signal δ_τ is computed, it updates the value estimate at a number of timepoints, in proportion to the uncertainty kernel.

Acute Changes in State Uncertainty Result in Biased Value Learning

Averaging over a convex value function results in overestimation of value. For an exponential value function, we can derive this result analytically in the continuous domain:

$$\int_t \gamma^{T-t} \mathcal{N}(t; \tau, \sigma_t^2) dt = \gamma^{T-\tau} \exp\left[\frac{(\ln \gamma)^2 \sigma_t^2}{2}\right], \tag{18}$$

where the second term on the left-hand side is greater than one. Intuitively, because the function is steeper on the right side and shallower on the left side, the average will be overestimated. Importantly, however, the estimate will be a multiple of the true value, with a scaling factor that depends on the width of the kernel (second term on left-hand side of Equation (18); note also that while we have assumed a Gaussian distribution, our results hold for any distribution that results in overestimation of value). Thus, with sensory feedback that modifies the width of the kernel upon transitioning from one state (τ) to the next ($\tau + 1$), there will be a mismatch in the value estimate when computing each RPE. More precisely, the learning rules are:

$$\hat{V}_\tau = \sum_t p(t|\tau, \sigma_t = s)\ \hat{V}_t \tag{19}$$

$$\hat{V}_{\tau+1} = \sum_t p(t|\tau + 1, \sigma_{t+1} = l)\ \hat{V}_t \tag{20}$$

$$\delta_\tau = r_\tau + \gamma \hat{V}_{\tau+1} - \hat{V}_\tau \tag{21}$$

$$\hat{V}_t^{(t+1)} = \hat{V}_t^{(t)} + \alpha \delta_\tau^{(t)} p(t|\tau, \sigma_t = s). \tag{22}$$

Notice that $\hat{V}_{\tau+1}$ takes different values depending on the state: When computing δ_τ,

$$\hat{V}_{\tau+1} = \sum_t p(t|\tau+1, \sigma_{t+1} = l)\ \hat{V}_t. \tag{23}$$

On the other hand, when computing $\delta_{\tau+1}$,

$$\hat{V}_{\tau+1} = \sum_t p(t|\tau+1, \sigma_{t+1} = s)\ \hat{V}_t. \tag{24}$$

How does this mismatch affect the learned value estimate? If averaging with kernels of different standard deviations can be written as multiples of true value, then they can be written as multiples of each other. The RPE is then

$$\delta_\tau = r_\tau + \gamma(a\hat{V}_{\tau+1,s}) - \hat{V}_{\tau,s}, \tag{25}$$

where we use the comma notation to denote that the two value estimates are evaluated with the same kernel width s, and a is a constant. By analogy with Equations (2) and (4), estimated value converges to $\hat{V}_\tau = (a\gamma)^{T-\tau} r$. Here, $a > 1$, so value is systematically overestimated. By the learning rules in Equations (19) to (22), this is because δ_τ is inflated by

$$\gamma \sum_t p(t|\tau+1, \sigma_{t+1} = l)\ \hat{V}_t - \gamma \sum_t p(t|\tau+1, \sigma_{t+1} = s)\ \hat{V}_t = \beta \hat{V}_\tau, \tag{26}$$

where β is defined in Equation (12).

An optimal agent will use the available sensory feedback to overcome this biased learning. Because averaging with a kernel of width l is simply a multiple of that with width s, it follows that a simple subtraction can achieve this correction (Equations (10) and (11)). Hence, sensory feedback can improve value learning with a correction term. It should be noted that with a complete correction to s as derived above, the bias is fully extinguished. For corrections to intermediate widths between s and l, the bias will be partially corrected but not eliminated. In both cases, because $\beta > 0$, ramps will occur.

RPEs Are Approximately the Derivative of Value

Consider the formula for RPEs in Equation (4). In tasks where a single reward is delivered at T, $r_t = 0$ for all $t < T$ (no rewards delivered before T). Because $\gamma \simeq 1$, the RPE can be approximated as

$$\delta_t \simeq \frac{\hat{V}_{t+1} - \hat{V}_t}{(t+1) - t}, \tag{27}$$

which is the slope of the estimated value. To examine the relationship between value and RPEs more precisely, we can extend our analysis to the continuous domain:

$$\begin{aligned}
\delta(t) &= \lim_{\Delta t \to 0} \frac{\gamma^{\Delta t}\hat{V}(t+\Delta t) - \hat{V}(t)}{\Delta t} \\
&= \dot{\hat{V}}(t) \lim_{\Delta t \to 0} \gamma^{\Delta t} + \hat{V}(t) \lim_{\Delta t \to 0} \frac{\gamma^{\Delta t} - 1}{\Delta t} \\
&= \dot{\hat{V}}(t) \lim_{\Delta t \to 0} \gamma^{\Delta t} + \hat{V}(t)(\ln \gamma) \lim_{\Delta t \to 0} \gamma^{\Delta t} \\
&= \dot{\hat{V}}(t) + \hat{V}(t) \ln \gamma,
\end{aligned} \tag{28}$$

where $\dot{\hat{V}}(t)$ is the time derivative of $\hat{V}(t)$, and the third equality follows from L'Hôpital's Rule. Here, $\ln \gamma$ has units of inverse time. Because $\ln \gamma \simeq 0$, RPE is approximately the derivative of value.

Sensory Feedback in Continuous Time

In the complete absence of sensory feedback, σ_t is not constant, but rather increases linearly with time, a phenomenon referred to as *scalar variability*, a manifestation of Weber's law in the domain of timing (Gibbon, 1977; Church and Meck, 2003; Staddon, 1965). In this case, we can write the standard deviation as $\sigma_t = wt$, where w is the Weber fraction, which is constant over the duration of the trial.

Set $l = w(\tau + \Delta\tau)$ and $s = w\tau$. Following the steps in the previous section,

$$\begin{aligned}
\delta(\tau) &= \lim_{\Delta \tau \to 0} \frac{\gamma^{\Delta \tau} e^{\frac{(\ln \gamma)^2}{2} w^2((\tau+\Delta\tau)^2 - \tau^2)} \hat{V}(\tau + \Delta\tau) - \hat{V}(\tau)}{\Delta \tau} \\
&= \dot{\hat{V}}(\tau) + \hat{V}(\tau) \ln \gamma + \hat{V}(\tau)(\ln \gamma)^2 w^2 \tau \\
&> \dot{\hat{V}}(\tau) + \hat{V}(\tau) \ln \gamma.
\end{aligned} \tag{29}$$

Hence, as derived for the discrete case, RPEs are inflated, and value is systematically overestimated.

RPE Ramps Result From Sufficiently Convex Value Functions

By Equation (28), the condition for ramping is $\dot{\delta}(t) > 0$, i.e., the estimated shape of the value function at any given point, before feedback, must obey

$$\ddot{\hat{V}}(t) + \dot{\hat{V}}(t) \ln \gamma > 0, \tag{30}$$

where $\ddot{\hat{V}}(t)$ is the second derivative of $\hat{V}(t)$ with respect to time. For an intuition of this relation, note that when $\gamma \simeq 1$, the inequality can be approximated as $\ddot{\hat{V}}(t) > 0$, which denotes any convex function. The exact inequality, however, has a tighter requirement on $\hat{V}(t)$: Since $\dot{\hat{V}}(t) \ln \gamma < 0$ for all t, ramping will only be observed if the contribution from $\ddot{\hat{V}}(t)$ (i.e., the convexity) outweighs the quantity $\dot{\hat{V}}(t) \ln \gamma$ (the scaled slope). For example, the function in Equation (2) does not satisfy the strict inequality even though it is convex, and therefore with this choice of $\hat{V}(t)$, the RPE does not ramp. In other words, to result in an RPE ramp, $\hat{V}(t)$ has to be 'sufficiently' convex.

Simulation Details

Value Learning Under State Uncertainty (Figure 1): For our TD learning model, we have chosen $\gamma = 0.9$, $\alpha = 0.1$, $n = 50$ states, and $T = 48$. In the absence of feedback, uncertainty kernels are determined by the Weber fraction, arbitrarily set to $w = 0.15$. In the presence of feedback, uncertainty kernels have a standard deviation of $l = 3$ before feedback and $s = 0.1$ after feedback. For the purposes of averaging with uncertainty kernels, value peaks at T and remains at its peak value after T, and the standard deviation at the last 4 states in the presence of feedback is fixed to 0.1. Intuitively, the animal expects reward to be delivered, and attributes any lack of reward delivery at $\tau = T$ to noise in its timing mechanism (uncertainty kernels have nonzero width) rather than to a reward omission. The learning rules were iterated 1000 times.

Value Learning in the Presence of Sensory Feedback (Figure 2): For our TD learning model, we have chosen $\gamma = 0.9$, $\alpha = 0.1$, $n = 50$ states, and $T = 48$. The learning rules were iterated 1000 times.

Relationship with Experimental Data (Figure 3): For our TD learning model, we have chosen $\gamma = 0.8$, $\alpha = 0.1$, and Weber fraction $w = 0.05$. For the navigation task, kernels have standard deviation $l = 3$ before feedback and $s = 0.1$ after feedback. In the experimental paradigms, trial durations were approximately 1.5 seconds (Schultz et al., 1997) and over 5 seconds (Howe et al., 2013). Thus for (B) and (D), we have arbitrarily set $n = 10$ and 25 states, respectively, between trial start and reward. The learning rules were

iterated 1000 times.

Full implementations can be found at www.github.com/jgmikhael/ramping.

Acknowledgments

The project described was supported by National Institutes of Health grants T32GM007753 and T32MH020017 (JGM), R01 MH110404 and MH095953 (NU), U19 NS113201-01 (SJG and NU), the Simons Collaboration on the Global Brain (NU), and a research fellowship from the Alfred P. Sloan Foundation (SJG). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Simons Collaboration on the Global Brain. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

J.G.M. and S.J.G. developed the model. H.R.K. and N.U. conceived that the structure of state uncertainty may influence the shape of estimated value functions and thus RPEs. J.G.M., H.R.K., N.U., and S.J.G. contributed to the writing of the paper. J.G.M. analyzed and simulated the model, and wrote the first draft.

Declaration of Interests

The authors declare no competing interests.

Data and Code Availability

Source code for all simulations can be found at www.github.com/jgmikhael/ramping.

Supplemental Information

1 Alternative Causes of Ramping

In the main text, we argue that ramping follows from normative principles. In this section, we illustrate that various types of biases ('bugs' in the implementation) may also lead to RPE ramps.

Ramping Due to State-Dependent Bias

Assume the animal persistently overestimates the amount of time or distance remaining to reach its reward (or, equivalently, that it underestimates the time elapsed or the distance traversed so far), and that this overestimation decreases as the animal approaches the reward. For instance, since the receptive fields of place cells decrease as the animal approaches reward (O'Keefe and Burgess, 1996), the contribution of place cells immediately behind the approaching animal in its estimate of value may outweigh that from the place cells in front of it. It will simplify our analysis to set $T = 0$ without loss of generality, and allow time to progress from the negative domain ($t < 0$) toward $T = 0$. In the continuous domain and for the simple case of linear overestimation, we can write this as

$$\hat{V}(t) = \gamma^{-\eta t} r, \tag{31}$$

where $\eta > 1$ is our overestimation factor. Therefore, by Equation (28),

$$\begin{aligned} \delta(t) &= \dot{\hat{V}}(t) + \hat{V}(t) \ln \gamma \\ &= (\ln \gamma)(1 - \eta)\gamma^{-\eta t} r, \end{aligned} \tag{32}$$

which is monotonically increasing. Hence, the RPE should ramp. Equivalently, in the discrete domain,

$$\begin{aligned} \delta_t &= \gamma \hat{V}_{t+1} - \hat{V}_t \\ &= \gamma \gamma^{-\eta(t+1)} r - \gamma^{-\eta t} r \\ &= \gamma^{-\eta t}(\gamma^{1-\eta} - 1) r. \end{aligned} \tag{33}$$

Here, $\delta_{t+1} > \delta_t$. Hence, the RPE should ramp.

Ramping Due to State-Dependent Discounting of Estimated Value

Assume the animal underestimates $\hat{V}(t)$ by directly decreasing the temporal discount term γ. Then if $\hat{V}(t) = (\eta\gamma)^{T-t} r$, with $\eta \in (0, 1)$, we can write in the continuous domain:

$$\begin{aligned} \delta(t) &= \dot{\hat{V}}(t) + \hat{V}(t) \ln \gamma \\ &= (-\ln \eta)(\eta\gamma)^{T-t} r, \end{aligned} \tag{34}$$

which is monotonically increasing. Hence, the RPE should ramp. Equivalently, in the discrete domain, if $\hat{V}_t = (\eta\gamma)^{T-t} r$ with $\eta \in (0, 1)$, we can write

$$\delta_t = (\eta\gamma)^{T-t} \left(\frac{1}{\eta} - 1 \right) r, \tag{35}$$

and

$$\delta_{t+1} = (\eta\gamma)^{-1} \delta_t. \tag{36}$$

Here, $\delta_{t+1} > \delta_t$. Hence, the RPE should ramp.

References

Babayan, B. M., Uchida, N., and Gershman, S. J. (2018). Belief state representation in the dopamine system. *Nature communications*, 9(1):1891.

Bellman, R. (1957). Dynamic programming. *Princeton University Press*.

Berke, J. D. (2018). What does dopamine mean? *Nature neuroscience*, page 1.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, 191(3):391–431.

Church, R. M. and Meck, W. (2003). A concise introduction to scalar timing theory. *Functional and neural mechanisms of interval timing*, pages 3–22.

Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R., and Khamassi, M. (2019). Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific reports*, 9(1):6770.

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, 482(7383):85–88.

Collins, A. L., Greenfield, V. Y., Bye, J. K., Linker, K. E., Wang, A. S., and Wassum, K. M. (2016). Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. *Scientific reports*, 6.

de Lafuente, V. and Romo, R. (2011). Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. *Proceedings of the National Academy of Sciences*, 108(49):19767–19771.

Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. *Nature*, 525:243–246.

Fiorillo, C. D., Newsome, W. T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nature neuroscience*, 11(8):966.

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., Akers, C. A., Clinton, S. M., Phillips, P. E., and Akil, H. (2011). A selective role for dopamine in stimulus–reward learning. *Nature*, 469(7328):53.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., and Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, 104(41):16311–16316.

Gardner, M. P., Schoenbaum, G., and Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B*, 285(1891):20181645.

Gershman, S. J. (2014). Dopamine ramps are a consequence of reward prediction errors. *Neural computation*, 26(3):467–471.

Gibbon, J. (1977). Scalar expectancy theory and weber's law in animal timing. *Psychological review*, 84(3):279.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15647–15654.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., Kennedy, R. T., Aragona, B. J., and Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, 19:117–126.

Hart, A. S., Rutledge, R. B., Glimcher, P. W., and Phillips, P. E. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *Journal of Neuroscience*, 34(3):698–704.

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generate ans use neural signals that predict reinforcement. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of information processing in the basal ganglia*. MIT Press, Cambridge.

Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E., and Graybiel, A. M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *nature*, 500(7464):575.

Kim, H. R., Malik, A. N., Mikhael, J. G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., Gershman, S. J., and Uchida, N. (2019). A unified framework for dopamine signals across timescales. *bioRxiv*.

Kobayashi, S. and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *Journal of neuroscience*, 28(31):7837–7846.

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., and Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, 27(6):821–832.

Lloyd, K. and Dayan, P. (2015). Tamping ramping: Algorithmic, implementational, and computational explanations of phasic dopamine signals in the accumbens. *PLoS computational biology*, 11(12):e1004622.

Ludvig, E., Sutton, R. S., Kehoe, E. J., et al. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2012). Evaluating the td model of classical conditioning. *Learning & behavior*, 40(3):305–319.

Menegas, W., Babayan, B. M., Uchida, N., and Watabe-Uchida, M. (2017). Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *Elife*, 6:e21886.

Menegas, W., Bergan, J. F., Ogawa, S. K., Isogai, Y., Venkataraju, K. U., Osten, P., Uchida, N., and Watabe-Uchida, M. (2015). Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *Elife*, 4:e10032.

Mikhael, J. G. and Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. *PLoS computational biology*, 12(9):e1005062.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *The Journal of neuroscience*, 16(5):1936–1947.

Moore, J., Desmond, J., and Berthier, N. (1989). Adaptively timed conditioned responses and the cerebellum: a neural network approach. *Biological cybernetics*, 62(1):17–28.

Morita, K. and Kato, A. (2014). Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. *Frontiers in neural circuits*, 8:36.

Nicola, S. M., Surmeier, D. J., and Malenka, R. C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annual review of neuroscience*, 23(1):185–215.

Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520.

Niv, Y. and Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in cognitive sciences*, 12(7):265–272.

O'Keefe, J. and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, 381(6581):425.

Ratcliff, R. and Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural computation*, 24(5):1186–1229.

Schultz, W. (2007a). Behavioral dopamine signals. *Trends in neurosciences*, 30(5):203–210.

Schultz, W. (2007b). Multiple dopamine functions at different time courses. *Annu. Rev. Neurosci.*, 30:259–288.

Schultz, W. (2010). Review dopamine signals for reward value and risk: basic and recent data. *Behav. Brain Funct*, 6:24.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Staddon, J. (1965). Some properties of spaced responding in pigeons. *Journal of the Experimental Analysis of Behavior*, 8(1):19–28.

Starkweather, C. K., Babayan, B. M., Uchida, N., and Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience*, 20(4):581–589.

Starkweather, C. K., Gershman, S. J., and Uchida, N. (2018). The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. *Neuron*, 98:616–629.

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience*, 16(7):966–973.

Stuber, G. D., Klanker, M., de Ridder, B., Bowers, M. S., Joosten, R. N., Feenstra, M. G., and Bonci, A. (2008). Reward-predictive cues enhance excitatory synaptic strength onto midbrain dopamine neurons. *Science*, 321(5896):1690–1692.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

Sutton, R. S. and Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement.

Totah, N. K., Kim, Y., and Moghaddam, B. (2013). Distinct prestimulus and poststimulus activation of vta neurons correlates with stimulus detection. *Journal of neurophysiology*, 110(1):75–85.

Wassum, K. M., Ostlund, S. B., and Maidment, N. T. (2012). Phasic mesolimbic dopamine signaling precedes and predicts performance of a self-initiated action sequence task. *Biological psychiatry*, 71(10):846–854.