Current Biology

CellPress

Article

The role of state uncertainty in the dynamics of dopamine

John G. Mikhael,[1,2,7,8,*] HyungGoo R. Kim,[3,4,5,7] Naoshige Uchida,[5] and Samuel J. Gershman[6]
[1]Program in Neuroscience, Harvard Medical School, Boston, MA 02115, USA
[2]MD-PhD Program, Harvard Medical School, Boston, MA 02115, USA
[3]Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon 16419, Republic of Korea
[4]Department of Biomedical Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
[5]Department of Molecular and Cellular Biology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
[6]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: john_mikhael@hms.harvard.edu
https://doi.org/10.1016/j.cub.2022.01.025

SUMMARY

Reinforcement learning models of the basal ganglia map the phasic dopamine signal to reward prediction errors (RPEs). Conventional models assert that, when a stimulus predicts a reward with fixed delay, dopamine activity during the delay should converge to baseline through learning. However, recent studies have found that dopamine ramps up before reward in certain conditions even after learning, thus challenging the conventional models. In this work, we show that sensory feedback causes an unbiased learner to produce RPE ramps. Our model predicts that when feedback gradually decreases during a trial, dopamine activity should resemble a "bump," whose ramp-up phase should, furthermore, be greater than that of conditions where the feedback stays high. We trained mice on a virtual navigation task with varying brightness, and both predictions were empirically observed. In sum, our theoretical and experimental results reconcile the seemingly conflicting data on dopamine behaviors under the RPE hypothesis.

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.[1–3] 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.[1] 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 (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.[1,3–6] 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.[2] However, recent results have called this model of DA into question. Using fast-scan cyclic voltammetry in the rat striatum during a goal-directed spatial navigation task, Howe et al.[7] 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,[8] during the execution of self-initiated action sequences,[9] and in the timing of movement initiation.[10] 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.[8] 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, we recently devised an experimental paradigm that dissociates the value and RPE interpretations of DA.[11] We began with the insight that, in the experiments considered above, RPEs can be approximated as the derivative of the value under the TD learning framework[12] (STAR Methods). This implies that, to effectively arbitrate between the value and RPE interpretations, one only needs to devise experiments where the value and its derivative are expected to behave very differently. Indeed, by training mice on a virtual reality environment and manipulating various properties of the task—namely, the speed of scene movement and the presence of forward teleportations and temporary pauses—we could make precise predictions about how the value should change versus how its derivative (RPE) should change. We found that the changes in DA behaviors were consistent with the RPE hypothesis and not with the value interpretation. The virtual reality task further allowed us to dissociate spatial navigation from locomotion (running), as one view of ramps had been that they are specific to operant tasks and that DA conveys qualitatively different information in operant versus classical conditioning tasks. However, we found that mice continued to display ramping DA signals during the task even without locomotion (i.e., when the mice did not run for reward). We confirmed these key results at the levels of somatic spiking of DA neurons, axonal calcium signals, and DA concentrations at neuronal terminals in the striatum. Taken together, these findings strongly support the RPE hypothesis of DA.

The body of experimental studies outlined above produces 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.

We begin this work by examining the influence of sensory feedback in guiding value estimation. Because of the 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 time point 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 the differences in a task’s feedback profile explain the puzzling collection of DA behaviors described above. To experimentally verify our hypothesis, we trained mice on a virtual navigation task in which the brightness of the virtual track was varied. As predicted by our framework, when the scene was darkened over the course of the trial (putatively decreasing the sensory feedback), DA exhibited a “bump” or a ramp up followed by a ramp down. Furthermore, the magnitude of signals during the ramp-up phase was globally greater than that of the corresponding ramp in conditions when the scene brightness remained high, as predicted by the theory.

We will begin the next section with a review of the TD learning algorithm and 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. Finally, we will specifically control the sensory feedback by manipulating the brightness of the track in a virtual navigation task, thereby uncovering DA bumps.

RESULTS

Temporal difference learning

In TD learning, an agent transitions through a sequence of states according to a Markov process.[13] The value associated with each state is defined as the expected discounted future return:

$$V_t = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right], \qquad \text{(Equation 1)}$$

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:

$$V_t = \gamma^{T-t} r, \qquad \text{(Equation 2)}$$

for all $t \in [0, T]$, where r is the magnitude of reward delivered at time T. In other words, value increases exponentially as reward time T is approached, peaking at a value of r at T (Figures 1D and 1F). Additionally, note that exponential functions are convex; the convex shape of the value function will be important in subsequent sections (see Kim et al.[11] for an experimental test of this property).

How does the agent learn this value function? Under the Markov property, the value at any time t, defined in Equation 1, can be rewritten as a sum of the reward received at t and the discounted value at the next time step:

$$V_t = r_t + \gamma V_{t+1}, \qquad \text{(Equation 3)}$$

which is referred to as the Bellman equation.[14] In other words, the value at time t is the sum of rewards received at t and the promise of future rewards. To learn V_t, the agent approximates it with $\widehat{V}_t$, 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:

$$\delta_t = r_t + \gamma \widehat{V}_{t+1} - \widehat{V}_t. \qquad \text{(Equation 4)}$$

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

$$\widehat{V}_t^{(n+1)} = \widehat{V}_t^{(n)} + \alpha \delta_t^{(n)}, \qquad \text{(Equation 5)}$$

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, $\widehat{V}_t = \gamma^{T-t} r$ on average, which is precisely the true value (see STAR Methods for a more general description of TD learning and its neural implementation).

Model overview

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 and sensory feedback affect learning. Our extension of the TD model to account for this case will involve three key ingredients:

(1) First, state uncertainty results in value overestimation. Intuitively, uncertainty about the state results in uncertainty about the value. However, the convexity of the value

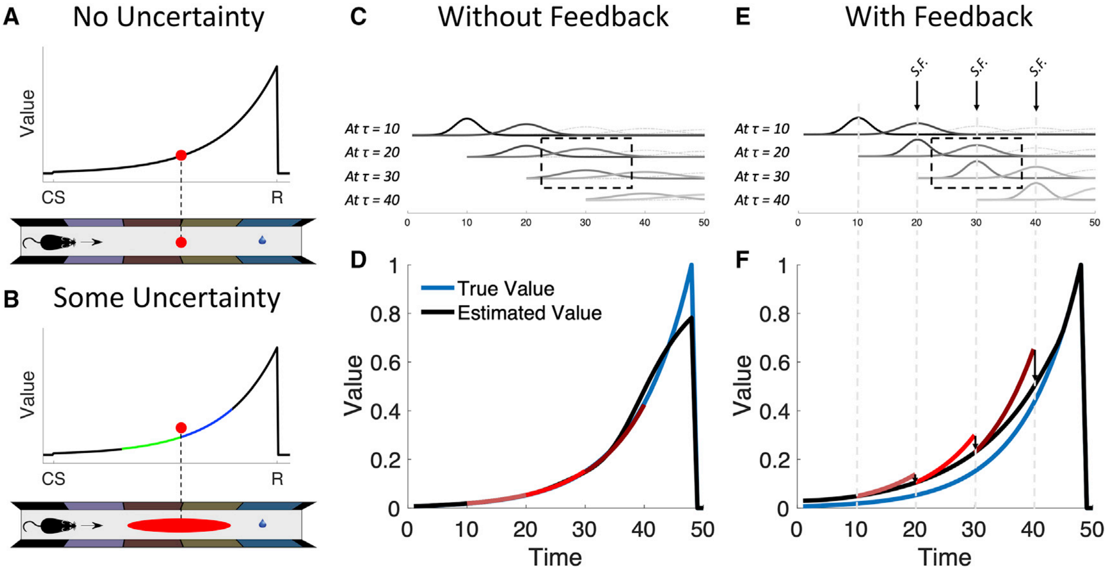

Figure 1. Sensory feedback biases value learning

(A) In the absence of state uncertainty, each state (red dot on maze) is mapped to its value (red dot on value function).

(B) On the other hand, when some state uncertainty is present (red ellipse on maze), the animal overestimates the value (red dot above value function). This is because convex functions are shallower to the left (green) and steeper to the right (blue), and the estimated value is a weighted average of the points on the green and blue segments.

(C) 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 (E).

(D) Without feedback, value is unbiased on average. Red curves illustrate the overestimated predicted increase in value between the current state and the next state (red curves; three examples extending over 10 states each for illustration only, as all 50 states are experienced on every trial). After learning, this roughly equals an increase by γ^{-1} on average.

(E) Sensory feedback reduces state uncertainty. Three instances of partial feedback (incomplete reduction in kernel widths) 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).

(F) 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 the true value. Red curves illustrate the overestimated value prediction. After learning, this roughly equals an increase by γ^{-1} on average. The illustration corresponds to a near-complete reduction in state uncertainty (lower kernel in the dashed box with near-zero width). See STAR Methods for simulation details.

function creates a bias, as early portions of the function are shallower than later portions (Figures 1A and 1B). This overestimation is greater with (a) greater uncertainty and (b) proximity to the reward.

(2) Second, sensory feedback that reduces this uncertainty biases learning. According to the TD algorithm, the agent takes a difference between two value estimates: one of the current state and another of the next state (Equation 4). If the agent systematically receives new information (in the form of sensory feedback) to reduce the uncertainty about the next state upon transitioning to it, then the learned value will be systematically biased.

(3) Third, the agent can correct this bias in the estimated value. In the TD algorithm, this can be written as a decay term that depends on the reduction in uncertainty due to sensory feedback and results in a persistent, positive RPE. This RPE is greater with (a) a greater reduction in uncertainty and (b) proximity to the reward. In other words, the RPE ramps. For the special case of tasks without feedback, the correction is null, and no ramps are observed.

Value learning under state uncertainty

Because animals do not have perfect internal clocks, they do not have complete access to the true time t.[15–17] Instead, t is a latent state corrupted by timing noise, often modeled as follows:

$$\tau \sim \mathcal{N}\left(t, \sigma_t^2\right), \qquad \text{(Equation 6)}$$

where τ is subjective 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 STAR Methods). Thus, the subjective estimate of value $\widehat{V}_\tau$ is an average over the estimated values $\widehat{V}_t$ of each state t:

$$\widehat{V}_\tau = \sum_t p(t|\tau)\widehat{V}_t, \qquad \text{(Equation 7)}$$

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 (Figures 1C and 1E). Intuitively, $\widehat{V}_\tau$ is the result of blurring $\widehat{V}_t$ proportionally to the uncertainty kernel (STAR Methods).

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 1D). A key requirement for this unbiased learning can be discovered by writing the RPE equations for two successive states:

$$\delta_\tau = r_\tau + \gamma \widehat{V}_{\tau+1} - \widehat{V}_\tau \qquad \text{(Equation 8)}$$

$$\delta_{\tau+1} = r_{\tau+1} + \gamma \widehat{V}_{\tau+2} - \widehat{V}_{\tau+1}. \qquad \text{(Equation 9)}$$

Notice here that $\widehat{V}_{\tau+1}$ is represented in both equations. In other words, $\widehat{V}_{\tau+1}$ must be computed at two separate time points: at τ (where it represents the value of the next state) and at $\tau+1$ (where it represents the value of the new, current state). The TD equations, in their standard form, require that $\widehat{V}_{\tau+1}$ remain the same regardless of when it is computed to achieve unbiased value learning. Said differently, for a value to be well learned, a requirement is that $\widehat{V}_{\tau+1}$ should 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 $\widehat{V}_{\tau+1}$ will acutely change as well. This means that $\widehat{V}_\tau$ will be discounted based on $\widehat{V}_{\tau+1}$ before feedback (i.e., as estimated at τ; red curves in Figure 1F) rather than $\widehat{V}_{\tau+1}$ after feedback (i.e., as estimated at $\tau+1$; black curve). In the next section, we will examine this effect more precisely, and we will show that any such acute change (here, due to sensory feedback) will cause an unbiased agent to produce ramping RPEs.

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 1E). This is because at time points preceding τ, the estimate of what the 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 $\widehat{V}_{\tau+1}$ will be estimated differently while computing δ_τ and $\delta_{\tau+1}$ (Equations 8 and 9; compare widths of similarly shaded kernels beneath each arrow in Figure 1E)—a violation of the requirement mentioned above, 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 values (Figures 1A and 1B). Intuitively, convex functions are steeper on the right (larger values; blue segment in Figure 1B) and shallower on the left (smaller values; green segment in Figure 1B); so, averaging results in a bias toward larger values. Furthermore, wider kernels result in greater overestimation (STAR Methods). Thus, upon entering each new state, the reduction of uncertainty via sensory feedback will acutely mitigate this overestimation, resulting in different estimates $\widehat{V}_{\tau+1}$ being used for δ_τ and $\delta_{\tau+1}$. Left uncorrected, the value estimate will be systematically biased, and in particular, the value will be overestimated at every point (Figure 2A; STAR 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 the value at the next state, then it will overestimate the 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 (Figures 2A and 2B).

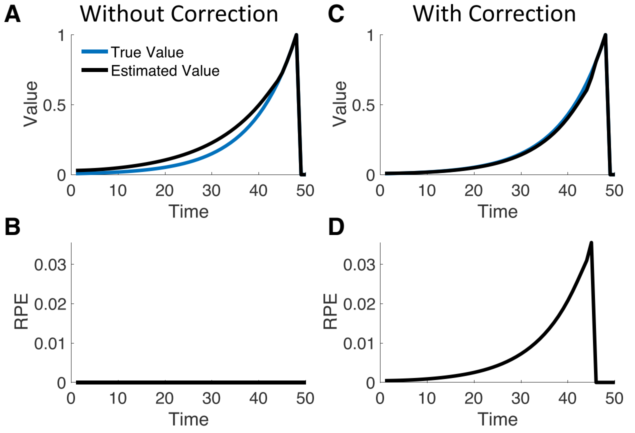

Figure 2. Unbiased learning in the presence of feedback leads to RPE ramps

(A) In a hypothetical task with sensory feedback, but in which correction does not occur, the value at each state is learned according to an overestimated version of value at the next state. Thus, a biased (suboptimal) value function is learned (see Figure 1F).

(B) After learning, the RPE converges to zero.

(C) With a correction term, the correct value function is learned instead.

(D) The cost of forcing an unbiased learning of value is a persistent RPE. Intuitively, the value at the current state is not influenced by the overestimated version of the value at the next state (compare with A and B). By Equation 13, this results in RPEs that ramp. See STAR Methods for simulation details.

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

$$\widehat{V}_t^{(n+1)} = \widehat{V}_t^{(n)} + \alpha\delta_\tau^{(n)} p(t|\tau) - \beta \widehat{V}_\tau^{(n)} p(t|\tau) \qquad \text{(Equation 10)}$$

$$\simeq \widehat{V}_t^{(n)} + \alpha\delta_\tau^{(n)} p(t|\tau) - \beta \widehat{V}_t^{(n)}, \qquad \text{(Equation 11)}$$

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

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

Here, the uncertainty kernel of $\widehat{V}_{\tau+1}$ has some standard deviation l at τ and a smaller standard deviation s at $\tau+1$. In other words, as the animal gains an improved estimate of $\widehat{V}_{\tau+1}$, it corrects the previously computed δ_τ with a feedback term to ensure unbiased learning of the 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? With enough learning, the RPE converges when the estimated value no longer

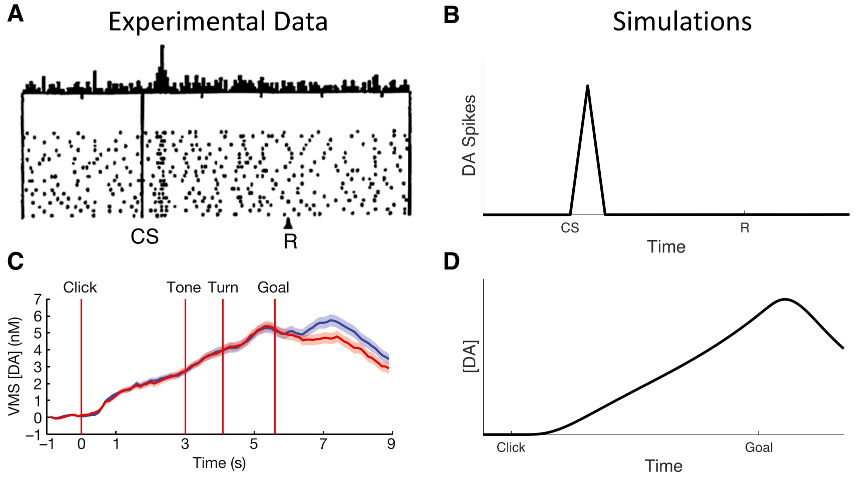

Figure 3. Differences in feedback result in different RPE behaviors

(A) Schultz et al.[1] 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.[1]

(B) Our derivations recapitulate this result. In the absence of sensory feedback, RPEs converge to zero. Note here the absence of an RPE at reward time in the experimental data. This is predicted by the model because the CS-R duration is very small (under 1.5 s) in the experimental paradigm, so temporal uncertainty is also small. Longer durations are predicted to result in an irreducible RPE response, as has been experimentally observed,[18] a point we return to in the discussion.

(C) Howe et al.[7] found that the DA signal ramps during a well-learned navigation task over the course of a single trial. Figure from Howe et al.[7]

(D) Our derivations recapitulate this result. In the presence of sensory feedback, RPEs track the shape of the estimated value function. See STAR Methods for simulation details.

changes on average, i.e., $\mathbb{E}[\widehat{V}_t^{(n+1)}] = \mathbb{E}[\widehat{V}_t^{(n)}]$. By Equation 10, the RPE will therefore converge to the following:

$$\delta_\tau = \frac{\beta}{\alpha}\widehat{V}_\tau. \qquad \text{(Equation 13)}$$

Therefore, with sensory feedback, the RPE ramps and tracks $\widehat{V}_\tau$ in shape (Figure 2D). In the absence of feedback, $\beta = 0$; thus, there is no ramp. Note here that the RPE is not a function of the learning rate α, as β itself is directly proportional to α (Equation 12).

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 (Figure 3A).[1,6,18–25] 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 (Figure 3C).[7] DA ramping is also present in classical conditioning tasks that do not involve locomotion but that include either spatial or non-spatial feedback,[11] as well as in two-armed bandit tasks,[8] in the timing of movement initiation,[10] and when executing self-initiated action sequences.[9,26]

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, there are 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). Intuitively, information received after an RPE has already been computed (and hence, after a DA response has already occurred) biases the learning of value. To offset this bias, the RPE converges to be non-zero at the equilibrium state (when the value is well learned). Furthermore, because of the convexity of the value function, this non-zero RPE must increase as the reward is approached.

In a direct test of the competing views of DA, we recently devised a series of experiments to disentangle the value and RPE interpretations (Figure 4, top panels).[11] We trained mice on a virtual reality paradigm, in which the animals experience virtual spatial navigation toward a reward. Visual stimuli on the (virtual) walls on either side of the path afforded the animals information about their location at any given moment. We then introduced a number of experimental manipulations—changing the speed of virtual motion, introducing a forward "teleportation" at various start and end points along the path, and pausing the navigation for 5 s before resuming virtual motion. We showed that the value interpretation of DA made starkly different predictions from the RPE hypothesis and then demonstrated that DA behavior was consistent with RPEs and not values.

To show this difference, we noted that RPEs can be approximated as the derivative of value (Equation 4, where $r_t = 0$ leading up to reward time, and γ is close to 1; note that this view ignores any contribution of state uncertainty). We then assumed that value is "sufficiently convex" (STAR Methods) to produce a derivative that increases monotonically. The task, then, was to simply examine the expected effect of each experimental manipulation on the value versus its derivative.

This view is limited in a number of ways. Perhaps most importantly, the presented model—that RPEs are the approximate derivative of the value—fails to capture the recursive effect of RPEs on the value: not only does a value estimate generate an RPE but also the RPE modifies the value estimate. If RPEs ramp, then they are always positive. But how, then, can the agent settle on a single value estimate if the RPE is always causing the estimate to increase? A second limitation of this model is that it had to *assume* a sufficiently convex value function to achieve a

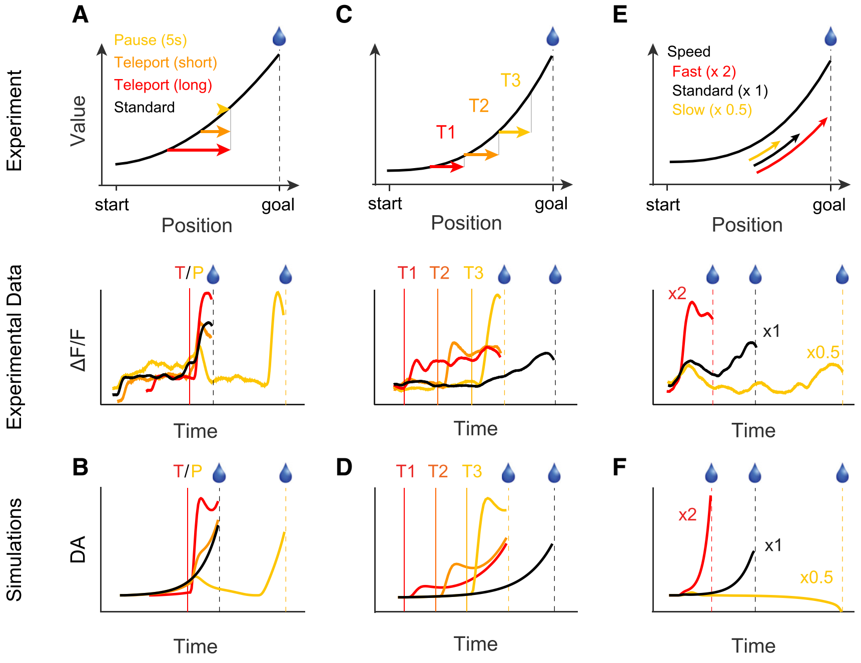

Figure 4. RPE behaviors match DA responses under various task manipulations

We trained head-fixed mice on a visual virtual reality task, in which they virtually navigated a scene with a reward at the end.[11] We then manipulated various aspects of the task.

(A) When the mice were teleported from different locations to the same end point, a large DA response resulted and scaled with the size of the teleport. When the navigation was paused for 5 s, the DA response dropped to baseline, with a large response occurring upon resuming navigation.

(B) Our derivations recapitulate this result. With an instantaneous jump toward the reward, the RPE is very large and increases with larger jumps. During a pause, the RPE drops to zero but rapidly increases when navigation resumes.

(C) When the mice were teleported from different locations but with the same magnitude, large DA responses resulted and increased in size closer to the reward.

(D) Our derivations recapitulate this result. Because of the convexity of the value function, an instantaneous teleportation of fixed magnitude will result in a larger RPE when it occurs closer to the reward.

(E) When the scene was navigated through more quickly, the ramp was steeper.

(F) Our derivations recapitulate this result. Faster navigation results in denser visual feedback per time point, i.e., the uncertainty kernels, defined by visual landmarks, become tighter with respect to true time. By Equations 12 and 13, this results in a greater reduction in uncertainty and thus a steeper ramp.

(A), (C), and (E) are from Kim et al.[11] See STAR Methods for simulation details.

monotonically increasing derivative (and hence a ramping RPE), leaving open the question of where this convexity originates from. Finally, this view cannot accommodate experiments where ramps are not observed. Instead, the model would seemingly predict ramping in all tasks, even though, as amply discussed above, this is not the case (e.g., Schultz et al.[1] and Kobayashi and Schultz[18]). In Figure 4, we show that our uncertainty-based model, which is not subject to these limitations, predicts the entire range of experimental results in Kim et al.[11]

Manipulation of sensory feedback and DA bumps

We have shown that our framework captures an array of DA behaviors. However, the manipulations considered above do not isolate sensory feedback as the key contributor to ramping. We, therefore, sought to develop an experimental paradigm that can distinguish our uncertainty-based model from the conventional models.

By describing a relationship between sensory feedback and DA ramps, our model predicts that a wide variety of DA responses can be elicited under the appropriate uncertainty profiles. In particular, our model makes an interesting prediction about a third type of behavior that to the best of our knowledge has not been previously observed: if state uncertainty rapidly increases over the course of a trial, then rather than a ramp, DA responses should exhibit a bump (Figure 5D). To see this intuitively, we can examine the RPE behaviors early and late in a trial in which the visual scene is gradually darkened, putatively decreasing the sensory feedback over the course of the trial. Initially, when the brightness is still high, the RPE should behave as in the constant brightness condition (i.e., ramps). As the scene darkens, wider uncertainty kernels “blur” the convex value function more. Thus, the early ramp in the darkening condition will be higher than that of the constant condition. However, later in the trial, as the animal approaches the reward, wider uncertainty kernels serve to flatten the estimated value function (near the maximum value, averaging over a larger window decreases the value estimate). Thus, the RPE will begin to decrease. Taken together, this results in an RPE bump that increases early on and decreases later. Furthermore, because of the lack of feedback near the reward time, the flatter estimated value function will result in a larger reward response than in the constant condition.

To test these predictions explicitly, we dynamically modulated the reliability of sensory evidence by changing the brightness of the visual scene over the course of a single trial (“darkening” condition; Figures 5 and S3; Video S1). The darkening condition (25% of trials) was randomly interleaved with the constant brightness condition (75% of trials). We independently interleaved the standard speed and fast conditions (on 25% of trials, the scene moved 1.7 times faster than the standard speed condition). Including a small portion of fast conditions appeared to help animals pay attention to the task. We monitored DA activity in the ventral striatum using fiber fluorometry (Figures 5B and 5C). Note that animals showed anticipatory licking in the darkening

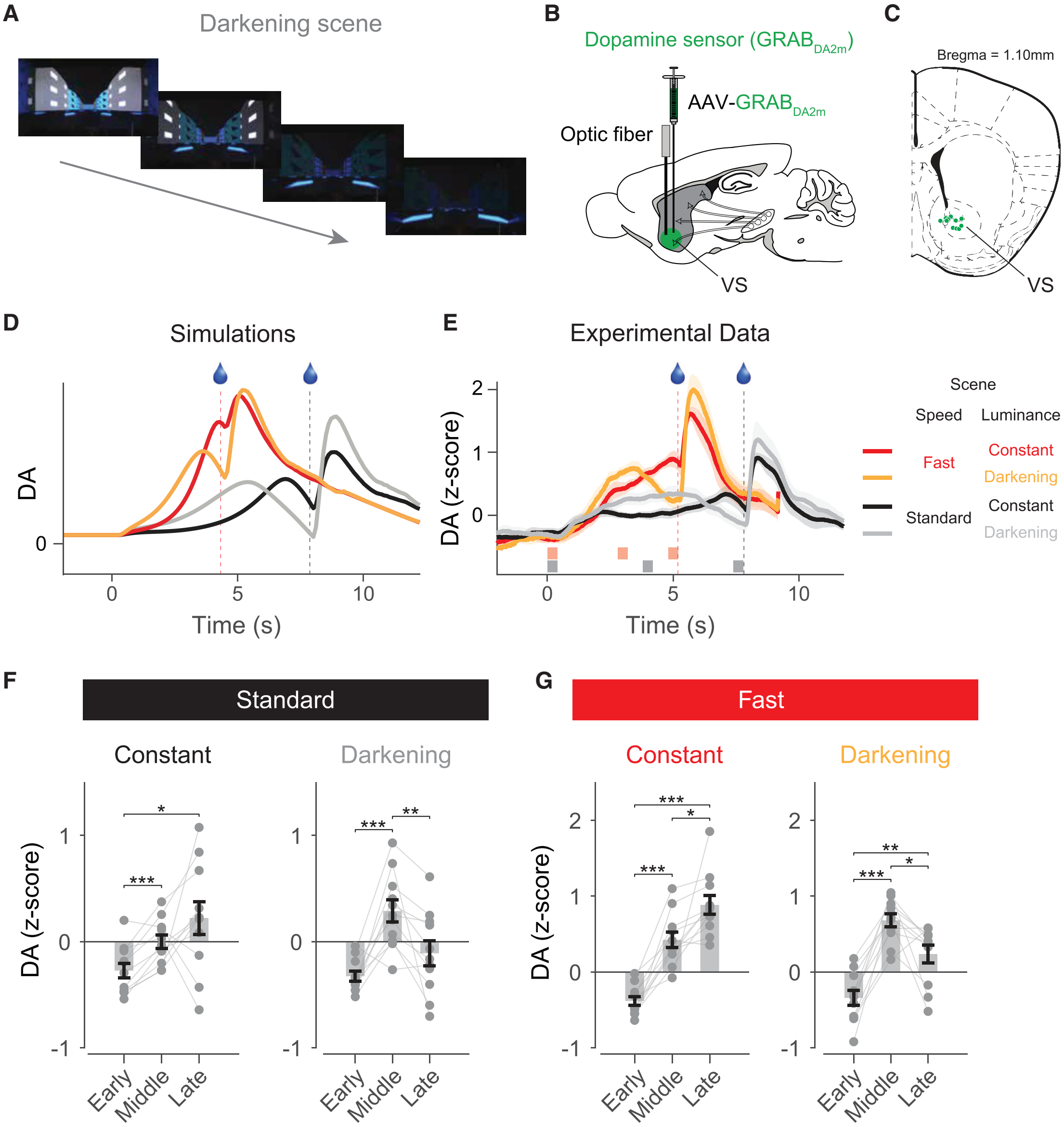

Figure 5. The state uncertainty model predicts DA responses in the darkening experiments

(A) Images of the visual scene captured at four different locations. The floor patterns were intact to prevent animals from inferring that the trial was aborted.

(B) Experimental design for fiber fluorometry. Adeno-associated virus (AAV) expressing a DA sensor (GRAB$_{DA2m}$) was injected into the ventral striatum (VS). DA signals were monitored through an optical fiber implanted into the VS.

(C) Recording locations. A coronal section of the brain at bregma, 1.10 mm.

(D) Model predictions. Note the three properties of the DA response in the darkening condition: the DA bump, the greater initial ramp compared with the constant condition, and the stronger reward response compared with the constant condition. Black, constant condition with standard speed; gray, darkening condition with standard speed; red, constant condition with fast speed (x1.7); yellow, darkening condition with fast speed.

(E) DA responses. Shaded areas at the bottom depict time windows for the three epochs used in (F) and (G).

(F) Average DA responses in the standard conditions. Three dots connected with lines represent individual animals (n = 11 mice).

(G) Average DA responses in the fast conditions (n = 11 mice). Shadings and error bars represent standard errors of the mean.

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, t test. See also Figure S3.

conditions (Figure S3B), suggesting that the animals did not think that the trials were aborted.

As predicted, our manipulations of scene brightness—putatively, manipulations of the sensory feedback—caused a DA bump, a signal that increases early on and decreases later (Figure 5E, gray and yellow curves). When the scene moved at standard speed, DA activity modestly ramped up in the constant condition (Figure 5F, left), whereas DA activity displayed a bump in the darkening condition (Figure 5F, right). The average responses in the middle epoch were significantly greater than those of either the start or end epoch ($p < 0.01$, t test, n = 11 mice). Ramping in the constant condition became more evident when the scene moved fast (Figure 5G, left). Nevertheless, we still observed a bump in the middle when the visual scene was darkened (Figure 5G, right). Furthermore, because of the lack of feedback near the reward time, our model predicts that the flatter estimated value function will result in a larger (phasic) response to the reward, compared with the constant condition, for both the standard and fast conditions, as indeed observed (Figure S3C, left and right, respectively; $p < 0.01$, t test, $n = 11$ mice).

DISCUSSION

While a large body of work has established phasic DA as an error signal,[1,3–6] more recent work has questioned this view.[7–9,26] 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.[8] 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[7]), but it can also be elicited in virtual reality tasks, in which animals do not need to run for reward (classical conditioning tasks[11]). Within classical conditioning, DA ramps can occur in the presence of navigational or non-navigational stimuli indicating time to reward.[11] Within operant tasks, ramps can be observed in the period preceding the action,[27] as well as during the action itself.[7] These ramps are, furthermore, not specific to experimental techniques and measurements and can be observed in cell body activities, axonal calcium signals, and DA concentrations.[11]

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 persistent 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. Moreover, a curious lemma of this result is that a measured DA signal whose shape tracks with the estimated value need not be evidence against the RPE hypothesis of DA, contrary to some claims.[8,28] Indeed, in the presence of persistent sensory feedback, δ_τ and $\widehat{V}_\tau$ 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 (see below for further discussion).

Our model implies that a variety of peculiar DA responses can be attained under the appropriate sensory feedback profiles. In particular, knowing that the value increases monotonically over the course of a trial, our results imply that a rapidly decreasing sensory feedback profile will result in a previously unobserved DA bump. By testing animals on conditions in which the visual scenes gradually darkened over the course of a single trial, we found exactly this result: a DA response that ramps up early on and ramps down later.

Our work takes inspiration from previous studies that examined the role of state uncertainty in DA responses.[18,29–34] For instance, temporal uncertainty increases with longer durations.[15–17] 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[18] and Fiorillo et al.[30] have observed. Similarly, Starkweather et al.[33] have found that in tasks with uncertainty both in *whether* reward will be delivered and *when* it is delivered, DA exhibits a prolonged dip (i.e., a negative ramp) leading up to reward delivery. Here, the value initially increases as expected reward time is approached, but then it 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 Babayan et al.[25] and Starkweather et al.[35]). As the authors of these studies note, both the 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.

Alternative hypotheses

One might argue that state uncertainty is not necessary to explain the results in the darkening experiments. To address this issue, we considered the possibilities that the DA responses can be explained either by the value interpretation of DA or by an RPE hypothesis that does not account for state uncertainty (STAR Methods). Briefly, the non-monotonic behavior of the DA response is incompatible with the value interpretation of DA, as darkening the visual scene should not decrease the value. Indeed, the animals' lick rates continued to increase in both the constant and darkening conditions (Figure S3). Second, the DA patterns are incompatible with the conventional, uncertainty-independent RPE view. To show this, we recovered the value functions from the putative RPE signals and found that the value in the darkening condition would have to be globally greater than that in the constant condition. However, under the uncertainty-free RPE hypothesis, the value in the darkening condition should either be the same as in the constant condition (value estimates unaffected by brightness) or smaller (if an inability to see the reward at the end of the trial leads to an assumed reward probability that is less than 1). We expand on these points in STAR Methods.

Finally, we note that our results are based on the assumption that animals maintain the same value function across experimental conditions. Said differently, we have assumed here that animals learn the value function in the constant condition and subsequently apply this previously learned value function to probe trials in which the scene is gradually darkened. It is possible, however, that animals learn a separate value function for the darkening conditions. Because RPEs in our model increase with larger values and decrease with lower feedback, it remains possible that such an alternative model will still capture the observed effects (STAR Methods).

While we have derived RPE ramping from normative principles, it is important to note that a complete correction is not necessary to produce ramping. Furthermore, biases in value learning may also produce ramping. For instance, one earlier proposal by Gershman[12] was that the 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 STAR Methods for a general derivation of the conditions for a ramp). Morita and Kato,[36] on the other hand, posited that value updating involves a decay term, which is qualitatively similar to that in Equation 10, and thus RPE ramping (see also implementations in Mikhael and Bogacz[37] and Cinotti et al.[38]). 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 (STAR Methods). 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 (STAR Methods).

Alternative accounts of DA ramping that deviate more significantly from our framework have also been proposed. In particular, Lloyd and Dayan[39] 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[40]). 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.[41] 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. Ramps may also result from *perceived* control, i.e., they may only occur if the animal *thinks* it can control the outcome of the task. While the virtual reality experiments of Kim et al.[11] strongly argue against this possibility, as the head-fixed animals who did not display running behavior during the task still exhibited ramps, it remains possible that these animals adopted some other, unmeasured superstitious behavior, thus resulting in perceived control. Finally, and relatedly, Howe et al.[7] have proposed that 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.[42–46] 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.

Lingering questions

A number of questions arise from our analysis. First, while our work examines learning with sensory feedback at the normative and algorithmic levels, how this uncertainty-guided update is implemented neurobiologically remains an open question. Our model predicts that RPEs depend on both the reduction in uncertainty and the estimated value. As the latter term develops with exposure to multiple trials, presumably via strengthening of synaptic weights,[47,48] so too will the ramps. However, how the signal noise and resulting reduction in uncertainty are encoded and how they evolve in parallel during the first few trials remain a subject of active debate.[49]

Second, 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, whose exact shape has been the subject of much interest.[50–53] 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. Beyond learning a suboptimal value function, the agent will furthermore be biased *across* options, as two equally rewarding options will generate different value functions if one was learned with feedback and the other was not (see STAR Methods for a similar case in which this bias is costly). Note also that, while we have chosen the exponential shape in Equation 2 after the conventional TD models, our ramping results extend to any convex value function.

Third, due to the presumed exponential shape, the ramping behaviors resulting from our analysis may also at times look exponential, rather than linear. 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. It remains an open question of how to best bring abstract TD models into alignment with biophysically realistic assumptions about the signal-generating process.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Temporal difference learning and its neural correlates
  - Value learning under state uncertainty
  - Acute changes in state uncertainty result in biased value learning
  - RPEs are approximately the derivative of value
  - Sensory feedback in continuous time
  - RPE ramps result from sufficiently convex value functions
  - Biased value estimates and reward forfeiture
  - Alternative hypotheses and DA bumps
  - DA bumps as a consequence of learning
  - Alternative causes of ramping
  - Simulation details
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical analysis
  - Fluorometry (photometry)
  - Licking and locomotion

- ○ Session-averaged time course
- ○ Population-averaged time course
- ○ Quantification for the darkening experiments

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2022.01.025.

ACKNOWLEDGMENTS

The project described was supported by the National Institutes of Health grants T32GM007753 and T32MH020017 (J.G.M.), R01 MH110404 and MH095953 (N.U.), and U19 NS113201-01 (S.J.G. and N.U.); the Air Force Office of Scientific Research grant FA9550-20-1-0413 (S.J.G. and N.U.); the Simons Collaboration on the Global Brain (N.U.); and a research fellowship from the Alfred P. Sloan Foundation (S.J.G.). 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. designed and conducted the experiments. 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. analyzed and simulated the model. J.G.M., H.R.K., N.U., and S.J.G. contributed to the writing of the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 20, 2021
Revised: November 22, 2021
Accepted: January 10, 2022
Published: February 2, 2022

REFERENCES

1. Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.
2. Schultz, W. (2007). Behavioral dopamine signals. Trends Neurosci. *30*, 203–210.
3. Glimcher, P.W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc. Natl. Acad. Sci. U. S. A *108* (supplement 3), 15647–15654.
4. Niv, Y., and Schoenbaum, G. (2008). Dialogues on prediction errors. Trends Cogn. Sci. *12*, 265–272.
5. 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. Nat. Neurosci. *16*, 966–973.
6. 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.
7. 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*, 575–579.
8. 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. Nat. Neurosci. *19*, 117–126.
9. 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. Sci. Rep. *6*, 20231.
10. Hamilos, A.E., Spedicato, G., Hong, Y., Sun, Fangmiao, Li, Y., and Assad, J.A. (2020). Dynamic dopaminergic activity controls the timing of self-timed movement. bioRxiv. https://doi.org/10.1101/2020.05.13.094904.
11. 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. (2020). A unified framework for dopamine signals across timescales. Cell *183*, 1600–1616.e25.
12. Gershman, S.J. (2014). Dopamine ramps are a consequence of reward prediction errors. Neural Comput. *26*, 467–471.
13. Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. Mach. Learn. *3*, 9–44.
14. Bellman, R. (1957). Dynamic Programming (Princeton University Press).
15. Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. Psychol. Rev. *84*, 279–325.
16. Church, R.M. (2003). A concise introduction to scalar timing theory. In Functional and Neural Mechanisms of Interval Timing, W.H. Meck, ed. (Taylor & Francis), pp. 3–22.
17. Staddon, J.E. (1965). Some properties of spaced responding in pigeons. J. Exp. Anal. Behav. *8*, 19–27.
18. Kobayashi, S., and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. J. Neurosci. *28*, 7837–7846.
19. 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*, 1690–1692.
20. 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*, 53–57.
21. 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*, 85–88.
22. 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. J. Neurosci. *34*, 698–704.
23. Menegas, W., Bergan, J.F., Ogawa, S.K., Isogai, Y., Umadevi Venkataraju, KannanU., Osten, P., Uchida, N., and Watabe-Uchida, M. (2015). Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. eLife *4*, e10032.
24. 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.
25. Babayan, B.M., Uchida, N., and Gershman, S.J. (2018). Belief state representation in the dopamine system. Nat. Commun. *9*, 1891.
26. 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. Biol. Psychiatry *71*, 846–854.
27. Totah, N.K.B., Kim, Y., and Moghaddam, B. (2013). Distinct prestimulus and poststimulus activation of VTA neurons correlates with stimulus detection. J. Neurophysiol. *110*, 75–85.
28. Berke, J.D. (2018). What does dopamine mean? Nat. Neurosci. *21*, 787–793.
29. Kakade, S., and Dayan, P. (2002). Dopamine: generalization and bonuses. Neural Netw. *15*, 549–559.
30. Fiorillo, C.D., Newsome, W.T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. Nat. Neurosci. *11*, 966–973.
31. Rao, R.P.N. (2010). Decision making under uncertainty: a neural model based on partially observable Markov decision processes. Front. Comput. Neurosci. *4*, 146.
32. de Lafuente, V., and Romo, R. (2011). Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. Proc. Natl. Acad. Sci. USA *108*, 19767–19771.

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

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

35. 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.e6.

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

37. Mikhael, J.G., and Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. PLoS Comput. Biol. *12*, e1005062.

38. 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. Sci. Rep. *9*, 6770.

39. Lloyd, K., and Dayan, P. (2015). Tamping ramping: algorithmic, implementational, and computational explanations of phasic dopamine signals in the accumbens. PLoS Comput. Biol. *11*, e1004622.

40. Nicola, S.M., Surmeier, J., and Malenka, R.C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. Annu. Rev. Neurosci. *23*, 185–215.

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

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

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

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

45. 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. Proc. Natl. Acad. Sci. USA *104*, 16311–16316.

46. Gardner, M.P.H., Schoenbaum, G., and Gershman, S.J. (2018). Rethinking dopamine as generalized prediction error. Proc. Biol. Sci. *285*, 20181645.

47. Houk, J.C., Adams, J.L., and Barto, A.G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Models of Information Processing in the Basal Ganglia, J.C. Houk, J.L. Davis, and D.G. Beiser, eds. (MIT Press).

48. Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. *16*, 1936–1947.

49. Deneve, S. (2012). Making decisions with unknown sensory reliability. Front. Neurosci. *6*, 75.

50. Rachlin, H., and Green, L. (1972). Commitment, choice and self-control 1. J. Exp. Anal. Behav. *17*, 15–22.

51. Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. Psychol. Bull. *82*, 463–496.

52. Tobin, H., and Logue, A.W. (1994). Self-control across species (Columba livia, Homo sapiens, and Rattus norvegicus). J. Comp. Psychol. *108*, 126–133.

53. Rachlin, H. (2000). The Science of Self-Control (Harvard University Press).

54. Ludvig, E.A., Sutton, R.S., and Kehoe, E.J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Comput. *20*, 3034–3054.

55. Ludvig, E.A., Sutton, R.S., and Kehoe, E.J. (2012). Evaluating the TD model of classical conditioning. Learn. Behav. *40*, 305–319.

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

57. Moore, J.W., Desmond, J.E., and Berthier, N.E. (1989). Adaptively timed conditioned responses and the cerebellum: a neural network approach. Biol. Cybern. *62*, 17–28.

58. Sutton, R.S., and Barto, A.G. (1990). Time-derivative models of Pavlovian reinforcement. In Learning and Computational Neuroscience: Foundations of Adaptive Networks, M. Gabriel, and J. Moore, eds. (MIT Press), pp. 497–537.

59. Allan, L.G. (2002). The location and interpretation of the bisection point. Q. J. Exp. Psychol. B *55*, 43–60.

60. Wearden, J.H. (2002). Traveling in time: a time-left analogue for humans. J. Exp. Psychol. Anim. Behav. Process. *28*, 200–208.

61. Wearden, J.H., and Jones, L.A. (2007). Is the growth of subjective time in humans a linear or nonlinear function of real time? Q. J. Exp. Psychol. (Hove) *60*, 1289–1302.

62. Jozefowiez, J., Gaudichon, C., Mekkass, F., and Machado, A. (2018). Log versus linear timing in human temporal bisection: a signal detection theory study. J. Exp. Psychol. Anim. Learn. Cogn. *44*, 396–408.

63. Ren, Y., Müller, H.J., and Shi, Zhuanghua (2020). Ensemble perception in the time domain: evidence in favor of logarithmic encoding of time intervals. bioRxiv. https://doi.org/10.1101/2020.01.25.919407.

64. Larsen, T., Leslie, D.S., Collins, E.J., and Bogacz, R. (2010). Posterior weighted reinforcement learning with state uncertainty. Neural Comput. *22*, 1149–1179.

65. Gershman, S.J., and Uchida, N. (2019). Believing in dopamine. Nat. Rev. Neurosci. *20*, 703–714.

66. Lustig, C., Matell, M.S., and Meck, W.H. (2005). Not "just" a coincidence: frontal-striatal interactions in working memory and interval timing. Memory *13*, 441–448.

67. O'Keefe, J., and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. Nature *381*, 425–428.

68. Gallistel, C.R., King, A., and McDonald, R. (2004). Sources of variability and systematic error in mouse timing behavior. J. Exp. Psychol. Anim. Behav. Process. *30*, 3–16.

69. Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129–141.

70. Morris, G., Arkadir, D., Nevet, A., Vaadia, E., and Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron *43*, 133–143.

71. Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science *299*, 1898–1902.

72. Daw, N.D., Courville, A.C., and Touretzky, D.S. (2006). Representation and timing in theories of the dopamine system. Neural Comput. *18*, 1637–1677.

73. Daw, N.D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. Neural Netw. *15*, 603–616.

74. Niv, Y., Duff, M.O., and Dayan, P. (2005). Dopamine, uncertainty and TD learning. Behav. Brain Funct. *1*, 6.

75. Aronov, D., and Tank, D.W. (2014). Engagement of neural circuits underlying 2D spatial navigation in a rodent virtual reality system. Neuron *84*, 442–456.

76. Franklin, K.B., and Paxinos, G. (2019). Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates, Compact: The Coronal Plates and Diagrams (Academic Press).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
AAV9-hSyn-DA2m	Vigene Biosciences	N/A
Experimental models: Organisms/strains		
Mouse: C57BL/6J	The Jackson Laboratory	Jax # 000664; RRID: IMSR JAX:000664
Software and algorithms		
VirMEn	Dmitriy Aronov	https://pni.princeton.edu/pni-softwaretools/virmen
MATLAB	MathWorks	https://www.mathworks.com/
Other		
Isosol (Isourane, USP)	Vedco	N/A
LRS-0473 DPSS Laser System	LaserGlow Technologies	Cat #R471003FX
Mono Fiber-optic Cannulas	Doric Lenses	MFC 200/245-0.53 5mm MF1.25 FLT

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, John G. Mikhael (john_mikhael@hms.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

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

Data for Figures 5 and S3 can be found at and https://doi.org/10.6084/m9.figshare.16706788.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

In addition to the fifteen GCaMP mice used in the previous study,[11] eleven adult C57/BL6J wild-type male mice were used for the scene darkening experiments using the DA sensor (DA2m). All mice were backcrossed for more than 5 generations with C57/BL6J mice. Animals were singly housed on a 12 hr dark/12 hr light cycle (dark from 07:00 to 19:00). All procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Animal Care and Use Committee.

METHOD DETAILS

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 a subjective 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[1,54,55]:

$$\widehat{V}_t = \sum_d w_d x_{d,t}, \qquad \text{(Equation 14)}$$

where $\widehat{V}_t$ denotes the estimated value at time t, and $x_{d,t}$ denotes the d^{th} 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:

$$w_d^{(n+1)} = w_d^{(n)} + \alpha\delta_t^{(n)} x_{d,t}, \qquad \text{(Equation 15)}$$

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 $\widehat{V}_t$ 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 $\widehat{V}_t$ 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,[47,48] phasic activity of midbrain DA neurons encodes the error signal δ_t,[1,3–6] and the output nuclei of the basal ganglia (substantia nigra pars reticulata and internal globus pallidus) encode estimated value $\widehat{V}_t$.[56]

We have implicitly assumed in the Results a maximally flexible feature set, the complete serial compound representation,[1,48,57,58] 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 $\widehat{V}_t = w_t$, and we can write Equation 15 directly in terms of $\widehat{V}_t$, as in Equation 5.

Value learning under state uncertainty

The animal has access to subjective time τ, from which it forms a belief state $p(t|\tau)$, or, in Bayesian terms, a posterior distribution over true time. For simplicity, we have taken this distribution to be Gaussian, and we assume weak priors so that temporal estimates, though noisy, are accurate. In this case, the subjective time estimate is $\mathbb{E}[t|\tau]$ and is equal to the posterior mean. Note here that we are only concerned with capturing the noisy property of internal clocks. While a large literature has sought to establish the exact relationship between internal ('psychological') time and true time with varying degrees of success (e.g., linear vs. logarithmic relationship[59–63]), our work is invariant to this exact relationship, and only depends on animals' ability to reproduce time veridically on average, with some noise,[15–17] which we take here to be Gaussian. Intuitively, animals only have access to subjective time, and compute values and RPEs with respect to subjective time. Because the mapping between subjective and objective time is monotonic, a ramp in subjective time will also be a ramp in objective time.

Given the subjective time τ, the RPE is then:

$$\delta_\tau = r_\tau + \gamma\widehat{V}_{\tau+1} - \widehat{V}_\tau, \qquad \text{(Equation 16)}$$

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

$$\widehat{V}_t^{(n+1)} = \widehat{V}_t^{(n)} + \alpha\delta_\tau^{(n)} p(t|\tau). \qquad \text{(Equation 17)}$$

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.[25,64]

Note here that, in the absence of uncertainty, our task structure obeys the Markov property: State transitions and rewards are independent of the animal's history given its current state. An appeal of using belief states is that the task remains Markovian, but in the posterior distributions rather than in the signals, and the TD algorithm can be applied directly to our learning problem, as in Equations 16 and 17. This problem is a type of partially observable Markov decision process.[65]

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 time domain by computing the convolution of an exponential value function with a Gaussian kernel:

$$\int_t \gamma^{T-t}\mathcal{N}(t;\tau,\sigma_t^2)dt = \int_t \frac{\gamma^{T-t}\exp\left(-\frac{1}{2}\left(\frac{t-\tau}{\sigma_t}\right)^2\right)}{\sigma_t\sqrt{2\pi}}dt \qquad \text{(Equation 18)}$$

$$= \gamma^{T-\tau}\exp\left[\frac{(\ln\gamma)^2\sigma_t^2}{2}\right]\left[\frac{1}{2}\mathrm{erf}\left(\frac{\sigma_t^2\ln\gamma + t - \tau}{\sigma_t\sqrt{2}}\right)\right]_t, \qquad \text{(Equation 19)}$$

where σ_t is the standard deviation of the uncertainty kernel at t. The integral is evaluated over the entire temporal interval (i.e., the duration of a trial leading up to reward), but the contribution of distant timepoints is negligible when the Gaussian kernel width is small compared to the total temporal interval. Thus we can evaluate the integral from $-\infty$ to $+\infty$ for analytical convenience, representing points that far precede and far exceed τ relative to the kernel width, respectively:

$$\int_t \gamma^{T-t}\mathcal{N}\left(t;\tau,\sigma_t^2\right)dt = \gamma^{T-\tau}\exp\left[\frac{(\ln\gamma)^2\sigma_t^2}{2}\right]\left[\frac{1}{2}\operatorname{erf}\left(\frac{\sigma_t^2\ln\gamma + t - \tau}{\sigma_t\sqrt{2}}\right)\right]_{-\infty}^{+\infty} \tag{Equation 20}$$

$$= \gamma^{T-\tau}\exp\left[\frac{(\ln\gamma)^2\sigma_t^2}{2}\right]\left[\frac{1}{2}((+1)-(-1))\right] \tag{Equation 21}$$

$$= \gamma^{T-\tau}\exp\left[\frac{(\ln\gamma)^2\sigma_t^2}{2}\right]. \tag{Equation 22}$$

The second term on the right-hand side is greater than one, so value is overestimated. 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 right-hand side of Equation 22; note also that while we have assumed a Gaussian distribution, our qualitative 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, at τ, the learning rules are:

$$\widehat{V}_\tau = \sum_t p(t|\tau, \sigma_t = s)\ \widehat{V}_t \tag{Equation 23}$$

$$\widehat{V}_{\tau+1} = \sum_t p(t|\tau + 1, \sigma_{t+1} = l)\ \widehat{V}_t \tag{Equation 24}$$

$$\delta_\tau = r_\tau + \gamma\widehat{V}_{\tau+1} - \widehat{V}_\tau \tag{Equation 25}$$

$$\widehat{V}_t^{(n+1)} = \widehat{V}_t^{(n)} + \alpha\delta_\tau^{(n)}p(t|\tau, \sigma_t = s). \tag{Equation 26}$$

Notice that $\widehat{V}_{\tau+1}$ takes different values depending on the state: At τ, the agent computes $\widehat{V}_{\tau+1}$ according to Equation 25, whereas at $\tau+1$, it computes $\widehat{V}_{\tau+1}$ as:

$$\widehat{V}_{\tau+1} = \sum_t p(t|\tau + 1, \sigma_{t+1} = s)\ \widehat{V}_t. \tag{Equation 27}$$

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

$$\delta_\tau = r_\tau + \gamma(a\widehat{V}_{\tau+1,s}) - \widehat{V}_{\tau,s}, \tag{Equation 28}$$

where we use the comma notation in the subscripts 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 $\widehat{V}_\tau = (a\gamma)^{T-\tau}r$. Here, a>1, so value is systematically overestimated. By the learning rules in Equations 23, 24, 25, and 26, this is because δ_τ is inflated by

$$\sum_t p(t|\tau + 1, \sigma_{t+1} = l)\ \widehat{V}_t - \sum_t p(t|\tau + 1, \sigma_{t+1} = s)\widehat{V}_t = \exp\left[\frac{(\ln\gamma)^2 l^2}{2}\right]\widehat{V}_\tau - \exp\left[\frac{(\ln\gamma)^2 s^2}{2}\right]\widehat{V}_\tau \tag{Equation 29}$$

$$= \left(\exp\left[\frac{(\ln\gamma)^2\left(l^2 - s^2\right)}{2}\right] - 1\right)\widehat{V}_\tau \tag{Equation 30}$$

$$= \frac{\beta}{\alpha}\widehat{V}_\tau. \tag{Equation 31}$$

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.

In extension of the Temporal difference learning and its neural correlates section, we can posit an implementation of uncertainty kernels in which sensory information is relayed from cortical areas[47,48] and the uncertainty due to Weber's law is based in fronto-striatal circuitry.[66]

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

$$\delta_t \simeq \frac{\widehat{V}_{t+1} - \widehat{V}_t}{(t+1) - t}, \qquad \text{(Equation 32)}$$

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:

$$\delta(t) = \lim_{\Delta t \to 0} \frac{\gamma^{\Delta t}\widehat{V}(t+\Delta t) - \widehat{V}(t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{\gamma^{\Delta t}\widehat{V}(t+\Delta t) - \gamma^{\Delta t}\widehat{V}(t) + (\gamma^{\Delta t} - 1)\widehat{V}(t)}{\Delta t} \qquad \text{(Equation 33)}$$

$$= \lim_{\Delta t \to 0} \frac{\widehat{V}(t+\Delta t) - \widehat{V}(t)}{\Delta t} \lim_{\Delta t \to 0} \gamma^{\Delta t} + \lim_{\Delta t \to 0} \frac{(\gamma^{\Delta t} - 1)\widehat{V}(t)}{\Delta t} \qquad \text{(Equation 34)}$$

$$= \dot{\widehat{V}}(t) \lim_{\Delta t \to 0} \gamma^{\Delta t} + \widehat{V}(t) \lim_{\Delta t \to 0} \frac{\gamma^{\Delta t} - 1}{\Delta t}$$

$$= \dot{\widehat{V}}(t) \lim_{\Delta t \to 0} \gamma^{\Delta t} + \widehat{V}(t) \lim_{\Delta t \to 0} \frac{\gamma^{\Delta t} \ln \gamma}{1} \qquad \text{(Equation 35)}$$

$$= \dot{\widehat{V}}(t) \lim_{\Delta t \to 0} \gamma^{\Delta t} + \widehat{V}(t)(\ln \gamma) \lim_{\Delta t \to 0} \gamma^{\Delta t}$$

$$= \dot{\widehat{V}}(t) + \widehat{V}(t) \ln \gamma, \qquad \text{(Equation 36)}$$

where $\dot{\widehat{V}}(t)$ is the time derivative of $\widehat{V}(t)$, and the fifth 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.[15–17] 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

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

$$\begin{aligned} \delta(\tau) &= \lim_{\Delta\tau \to 0} \frac{\gamma^{\Delta\tau} e^{\frac{(\ln\ \gamma)^2}{2} w^2 \left((\tau+\Delta\tau)^2 - \tau^2\right)} \widehat{V}(\tau+\Delta\tau) - \widehat{V}(\tau)}{\Delta\tau} \\ &= \dot{\widehat{V}}(\tau) + \widehat{V}(\tau)\ln\gamma + \widehat{V}(\tau)(\ln\gamma)^2 w^2 \tau \\ &> \dot{\widehat{V}}(\tau) + \widehat{V}(\tau)\ln\gamma. \end{aligned} \qquad \text{(Equation 37)}$$

RPE ramps result from sufficiently convex value functions

By Equation 36, 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

$$\ddot{\widehat{V}}(t)+\dot{\widehat{V}}(t)\ln\gamma>0, \qquad \text{(Equation 38)}$$

where $\ddot{\widehat{V}}(t)$ is the second derivative of $\widehat{V}(t)$ with respect to time. For an intuition of this relation, note that when $\gamma\simeq1$, the inequality can be approximated as $\ddot{\widehat{V}}(t)>0$, which denotes any convex function. The exact inequality, however, has a tighter requirement on $\widehat{V}(t)$: Since $\dot{\widehat{V}}(t)\ln\gamma<0$ for all t, ramping will only be observed if the contribution from $\ddot{\widehat{V}}(t)$ (i.e., the convexity) outweighs the quantity $\dot{\widehat{V}}(t)\ln\gamma$ (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 $\widehat{V}(t)$, the RPE does not ramp. In other words, to produce an RPE ramp, $\widehat{V}(t)$ has to be 'sufficiently' convex.

Biased value estimates and reward forfeiture

Let us illustrate here how a biased value function can lead to suboptimal choices. Imagine a two-armed bandit task in which the animal chooses between two options, A and B, yielding rewards r_A and r_B, respectively, after a fixed delay T.

Assume $r_A=1$ is learned under conditions with rich sensory feedback, and $r_B=1.5$ is learned without feedback. Assume, also, that the animal learns according to the TD algorithm without a correction term. Using the simulation parameters for Figure 2A, with a delay of $T=20$, it follows that the values at the time of choice are $\widehat{V}_A(0)=0.2$ (Figure 2A, black curve at $t=28$) and $\widehat{V}_B(0)=r_B\gamma^T=(1.5)(0.9^{20})=0.18$ (Figure 2A, approximated as blue curve at $t=28$, scaled by r_B). After learning, the animal will be more likely to select A. (Furthermore, a greedy animal will asymptotically only select A.) With each selection of A, the animal forfeits an additional $\frac{r_B-r_A}{r_A}=50\%$ of reward potential.

Alternative hypotheses and DA bumps

We have argued in the main text that DA bumps can be captured by an uncertainty-driven view of RPEs but not by the value interpretation or the standard, uncertainty-free RPE hypothesis. To rule out the alternative hypotheses, we begin by deconvolving the DA2m response, yielding a signal that we interpret as either pure value or uncertainty-free RPE.

The deconvolved signal is monotonic in the constant condition but non-monotonic in the darkening condition (Figure S1B). On the other hand, the licking data—putatively reflecting the animal's estimate of value—increases monotonically in both conditions (Figure S3B, top panel). Taken together, these findings rule out the value interpretation of DA.

Next, we show that this signal is incompatible with an uncertainty-free RPE. To do so, we infer the value from the computed RPE (Figure S1C, using the derivation below). There is one free parameter, γ. We find that value is greater in the darkening condition than in the constant condition, even though under the uncertainty-free RPE hypothesis, it should either be the same (value estimate unaffected by brightness) or smaller (if an inability to see the reward location suggests a probability of receiving reward that is no longer equal to 1). Although γ is a free parameter, this result does not depend on γ, as $V_{t+1}=\frac{\delta_t+V_t}{\gamma}$, so γ simply amplifies or reduces existing differences, but does not reverse them.

To derive value from RPEs and γ, we use the relation:

$$V_t=\sum_{t'=0}^{t-1}\frac{\delta_{t'}}{\gamma^{t-t'}} \quad \text{for } t>0. \qquad \text{(Equation 39)}$$

To show that Equation 39 solves for V_t using Equation 4 leading up to reward (i.e., when $r_t=0$), we use proof by induction. First, for $t=1$,

$$V_1=\sum_{t'=0}^{0}\frac{\delta_{t'}}{\gamma^{t-t'}}=\frac{\delta_0}{\gamma}. \qquad \text{(Equation 40)}$$

Thus Equation 39 holds for $t=1$. Now assume it holds for t; let us show it also holds for $t+1$: as required.

$$\begin{aligned} V_{t+1} &= \frac{\delta_t}{\gamma}+\frac{V_t}{\gamma} \\ &= \frac{\delta_t}{\gamma}+\frac{1}{\gamma}\sum_{t'=0}^{t\ 1}\frac{\delta_{t'}}{\gamma^{t-t'}} \\ &= \frac{1}{\gamma}\left(\delta_t+\sum_{t'=0}^{t\ 1}\frac{\delta_{t'}}{\gamma^{t-t'}}\right) \\ &= \frac{1}{\gamma}\sum_{t'=0}^{t}\frac{\delta_{t'}}{\gamma^{t-t'}}, \end{aligned} \qquad \text{(Equation 41)}$$

DA bumps as a consequence of learning

In modeling the darkening manipulation, we have assumed that animals do not learn a separate value function for the probe trials in the darkening condition. We noted, however, that because of the opposite effects of the uncertainty profile and value on the RPE signal, bumps should still be observed when the manipulation occurs during learning (rather than only during performance). We show this analytically here.

Consider a manipulation in which the scene is gradually darkened, transitioning from perfect brightness to complete darkness over the course of a single trial. Using the terminology in the main text, the reduction in standard deviation $(l - s)$ decreases monotonically over the course of the trial (less sensory feedback), eventually reaching zero. But value increases monotonically over the trial, starting at zero. By Equation 13, the RPE reflects a product of $\widehat{V}$ and β, which itself depends on $(l^2 - s^2) = (l - s)(l + s)$. This means that the RPE should be zero at the beginning of the task and the end, but be positive in the middle. Because both V and β are continuous and differentiable, so is their product. Thus we predict that the RPE will gradually increase, reach some maximum, and subsequently decrease back to zero within a single trial (Figure S2).

Alternative causes of ramping

We have argued that ramping follows from normative principles. Here we illustrate that various types of biases (‘bugs’ in the implementation) may also lead to RPE ramps.

Ramping due to bias in state estimation

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,[67] 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

$$\widehat{V}(t) = \gamma^{-\eta t} r, \tag{Equation 42}$$

where $\eta > 1$ is our overestimation factor. Therefore, by Equation 36,

$$\begin{aligned} \delta(t) &= \dot{\widehat{V}}(t) + \widehat{V}(t)\ln\gamma \\ &= (\ln\gamma)(1-\eta)\gamma^{-\eta t} r, \end{aligned} \tag{Equation 43}$$

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

$$\begin{aligned} \delta_t &= \gamma \widehat{V}_{t+1} - \widehat{V}_t \\ &= \gamma\gamma^{-\eta(t+1)} r - \gamma^{-\eta t} r \\ &= \gamma^{-\eta t}\left(\gamma^{1-\eta} - 1\right) r. \end{aligned} \tag{Equation 44}$$

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

Ramping due to state-dependent discounting of estimated value

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

$$\begin{aligned} \delta(t) &= \dot{\widehat{V}}(t) + \widehat{V}(t)\ln\gamma \\ &= (-\ln\eta)(\eta\gamma)^{T-t} r, \end{aligned} \tag{Equation 45}$$

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

$$\delta_t = (\eta\gamma)^{T-t}\left(\frac{1}{\eta} - 1\right) r, \tag{Equation 46}$$

and

$$\delta_{t+1} = (\eta\gamma)^{-1}\delta_t. \tag{Equation 47}$$

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

Simulation details

In all simulations, the agent updated its estimate of a value function according to the TD algorithm, implemented by Equations 7, 8, 10, and 12. Task-specific details and choices of parameters are described below.

Impulse response function

To model experiments involving Ca^{2+} and DA2m signals, we used the GCaMP impulse response function obtained in Kim et al.,[11] and the DA2m impulse response function was obtained in the same manner, by averaging responses to unexpected reward. These functions were convolved with the computed RPEs to obtain simulated Ca^{2+} signals (Figure 4) and DA2m signals (Figures 5D, S1, and S2D).

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, set to $w = 0.15$.[68] 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 agent 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 agent iterated through all 50 states on every trial (three red curves on figure only to visually illustrate value overestimation). The agent experienced 1000 successive trials.

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 agent experienced 1000 successive trials.

Relationship with experimental data (Figures 3 and 4)

For convolutions over negative RPEs, it is important to account for the low baseline firing rates of DA neurons, i.e., that negative RPEs cannot elicit phasic responses that equal those elicited by positive RPEs of similar magnitude. Thus, following previous experimental[69–71] and theoretical[72–74] work, we account for an asymmetry between positive and negative RPEs in the DA signal. We do so by scaling the RPEs by the maximum change in spiking activity in either the positive or negative direction. After Kim et al.,[11] resting state spiking activity is approximately 5 spikes/second, the maximum spiking is 30 spikes/second, and the minimum spiking is 0 spikes/second. Thus one unit of positive RPE influences the DA response $\frac{30-5}{5-0} = 5$ times as strongly as one unit of negative RPE.

Figure 3: For our TD learning model, we have chosen $\gamma = 0.98$, $\alpha = 0.1$, and Weber fraction $w = 0.15$. For the navigation task, kernels have standard deviation $l = 3$ before feedback and $s = 0.1$ after feedback. For (B) and (D), we have set $n =$ 10 and 70 states, respectively, between trial start and reward. RPEs were convolved with the GCaMP kernel, as described above, to produce simulated DA behaviors. The agent experienced 2000 successive trials.

Figure 4: For our TD learning model, we have chosen $\gamma = 0.93$, $\alpha = 0.1$, and $w = 0.15$. The locomotion manipulations in the pause, teleport, and speed conditions all matched those in the experiments of Kim et al.[11] In particular, standard trials had length 7.6s from the CS to reward, and we set 10 states per second in our simulations. The agent was trained on the standard task and subsequently experienced either an unexpected pause, an unexpected teleport, or an unexpected change in navigation speed. In the pause condition, the agent experienced a 5-s pause at the 53rd state (i.e., after navigating 70% of states between the CS and reward). In the short and long teleport conditions, states 59-62 and 40-62 were omitted, respectively, corresponding to 5% and 30% of states between the CS and reward. In the teleport conditions of equal magnitude, 25 states (30% of states between the CS and reward) were omitted, beginning at state 5, 25, or 45. Kernels have standard deviation $l = 1$ before feedback and $s = 0.5$ after feedback for the teleport and pause manipulations. In the speed conditions, the task was experienced at either 20 (fast), 10 (normal), or 5 (slow) states per second. Kernels have standard deviation $l = 3$ before feedback and $s = 1$ after feedback for the standard-speed manipulation. Experiencing the trial twice as fast corresponds to the kernels being stretched by a factor of 2, resulting in a greater reduction in uncertainty and a steeper ramp. Intuitively, navigating a track very quickly leads to lower precision about one's exact location at any given moment. Similarly, experiencing the trial in the slow condition corresponds to a smaller reduction in uncertainty and a shallower ramp. In our simulation, the reduction in uncertainty is sufficiently weak that the shape of the value function dominates the RPE (see black curve in Figure 1D, corresponding to estimated value without feedback). Near reward time, the estimated value function may not be sufficiently convex (and may even be concave) with weak or absent feedback, so the RPE becomes negative. RPEs were convolved with the GCaMP kernel, as described above, to produce simulated DA behaviors. The agent experienced 2000 successive trials.

Manipulation of sensory feedback and DA bumps

Figure 5: The TD model is identical to that in Figure 4. For both the constant and darkening conditions, we have chosen $\gamma = 0.93$, $\alpha = 0.1$, $w = 0.15$, and $n = 200$ states. For the constant condition, the small kernel width is a constant, $s = a$. For the darkening condition, the width resembles that of the constant condition early on and resembles one without feedback later, $(s - a)(s - wt - b) = c$. The shape of this function is controlled by two parameters, c and b. The first determines how smoothly s transitions from resembling that of the constant condition to behaving according to Weber's law, and the second determines when this occurs. The large uncertainty kernel width is $l = s + z$, where z is a constant in the constant condition, and z decreases smoothly to zero over the course of the trial in the darkening condition, which we model as $z = \frac{d}{1+\exp(et)}$. We set $a = 8$, $b = 0.3$, $c = 3$, $d = 1$, and $e = 1$. Because the reduction in uncertainty $(l^2 - s^2)$ is constant in the constant condition and decreases in the darkening condition, it follows that β is constant in the constant condition and decreases in the darkening condition, as well. RPEs were convolved with the DA2m kernel, as described above, to produce simulated DA behaviors. The agent experienced 2000 successive trials.

Surgery and virus injections. *Surgery for fiber fluorometry of DA sensor signals*. To prepare animals for recording, we performed a single surgery with three key components: (1) injection of a DA sensor into the ventral striatum, (2) head-plate installation, and (3) implantation of an optical fiber into the striatum.[24,25] At the time of surgery, all mice were 2–4 months old. All surgeries were performed under aseptic conditions with animals anesthetized with isoflurane (1-2% at 0.5-1.0 L/min). Analgesia (ketoprofen for post-surgery treatment, 5 mg/kg I.P.; buprenorphine for pre-operative treatment, 0.1 mg/kg, I.P.) was administered for 3 days following each surgery. We removed the skin above the surface of the brain and dried the skull using air. We injected 400 nL of

AAV9-hSyn-DA2m (Vigene Biosciences) into the ventral striatum (bregma 1.0, lateral 1.1, depths 4.2 and 4.1 mm). Virus injection lasted several minutes, and then the injection pipette was slowly removed over the course of several minutes.

We then installed a head-plate for head-fixation by gluing a head-plate onto the top of the skull (C&B Metabond, Parkell). We used ring-shaped head-plates to ensure that the skull above the striatum would be accessible for fiber implants. Finally, during the same surgery, we also implanted optical fibers into the ventral striatum. To do this, we first slowly lowered optical fibers (200 μm diameter, Doric Lenses) into the striatum using a fiber holder (SCH_1.25, Doric Lenses). The coordinates we used for targeting were bregma 1.0, lateral 1.1, depth 4.1 mm. Once fibers were lowered, we first attached them to the skull with UV-curing epoxy (Thorlabs, NOA81), and then a layer of black Ortho-Jet dental adhesive (Lang Dental, IL). After waiting for fifteen minutes for this glue to dry, we applied a small amount of rapid-curing epoxy (A00254, Devcon) to attach the fiber cannulas to the underlying glue and head-plate. After waiting for fifteen minutes for the epoxy to cure, the surgery was completed.

Surgery for fiber fluorometry of GCaMP signals in the ventral striatum. To examine axonal calcium signals of dopaminergic neurons in the ventral striatum, we injected AAV-FLEX-GCaMP into the midbrain of DAT-Cre mice.[11] Surgical procedures up to virus injection were the same as the DA sensor injections described above. We unilaterally injected 250 nL of AAV5-CAG-FLEX-GCaMP6m (1×10^{12} particles/mL, Penn Vector Core) into both the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) (500 nL total). To target the VTA, we made a small craniotomy and injected the virus at bregma 3.1, lateral 0.6, depths 4.4 and 4.1 mm. To target SNc, we injected the virus at bregma 3.3, lateral 1.6, depths 3.8 and 3.6 mm.

Virtual reality setup. Virtual environments were displayed on three liquid crystal display (LCD) monitors with thin frames.[11] VirMEn software[75] was used to generate virtual objects and render visual images using perspective projection. Mice were head-restrained at the center of three monitors. Mice were placed on a cylindrical styrofoam treadmill (diameter 20.3 cm, width 10.4 cm). The rotational velocity of the treadmill was encoded using a rotary encoder. The output pulses of the encoder were converted into continuous voltage signals using a customized Arduino program running on a microprocessor (Teensy 3.2). Water reward was given through a water spout located in front of the animal's mouth. Licking tongue movements were monitored using an infrared sensor (OPB819Z, TT Electronics). Voltage signals from the rotary encoder and the lick sensor were digitized into a PCI-based data-acquisition system (PCIe-6323, National Instruments) installed on the visual stimulation computer. Timing and amount of water were controlled through a micro-solenoid valve (LHDA 1221111H, The Lee Company) and switch (2N7000, On Semiconductor). Analog output TTL pulse was generated from the visual stimulation computer to deliver reward to the animals.

Virtual linear track experiments. Animals were trained in a virtual linear track (see Kim et al.[11] for details). The maze was composed of a starting platform and a corridor with walls on both sides. We first trained animals on the standard approach-to-target task to learn the association between target location and reward. Once the animals learned the task, we ran a series of tasks with test trials to examine the nature of DA signals. In this paper, we simulated three main experiments in the previous study (Figure 4).[11] We typically ran each task for two consecutive days (with a zero- or one-day break). Unless otherwise noted, unexpected reward (5 μL) was given during the inter-trial interval on 3-6% of trials.

Scene darkening manipulation. We dynamically modulated the reliability of sensory evidence by changing the brightness of the visual scene (Video S1). The brightness of the visual scene at each time point was determined by multiplying the original RGB color values with a time-varying multiplier. The multiplier $k(t)$ is a function of the animal's position as defined below (Figure S3A).

$$P_{\text{norm}}(t) = \frac{P(t)}{91}, \text{ if } P(t) \le 91 \tag{Equation 48}$$

$$P_{\text{norm}}(t) = 1, \text{ if } P(t) > 91 \tag{Equation 49}$$

$$k(P_{\text{norm}}(t)) = k_{\text{start}} + (k_{\text{end}} - k_{\text{start}})(1 - P_{\text{norm}}(t))^3, \tag{Equation 50}$$

where $k_{\text{start}} = 1.0, k_{\text{end}} = 0.05$, and $P(t)$ is animal's position at time t. The brightness of the floor pattern was intact to provide the animals a clue that trials were not aborted. We randomly interleaved four experimental conditions. On 25% of trials, the visual scene was darkened as described above. Brightness was kept constant ($k(t) = 1$) for the rest of the trials. Independent of the brightness manipulation, the speed of visual scene progression was increased by 1.7 times on 25% of trials. Since the darkening depends on the position of the animal, for each darkening condition, the brightness of the scene at the reward location is identical between the standard and fast conditions.

Fiber fluorometry (photometry). Fluorescent signals from the brain were recorded using a custom-made fiber fluorometry (photometry) system as described in our previous studies.[11,24,25] The blue light (473 nm) from a diode-pumped solid-state laser (DPSSL; 80–500 μW; Opto Engine LLC, UT, USA) was attenuated through a neutral density filter (4.0 optical density, Thorlabs, NJ, USA) and coupled into an optical fiber patchcord (400 μm, Doric Lenses) using a 0.65 NA microscope objective (Olympus). The patchcord connected to the implanted fiber was used to deliver excitation light to the brain and to collect the fluorescence emission signals from the brain. The fluorescent signal from the brain was spectrally separated from the excitation light using a dichroic mirror (T556lpxr, Chroma), passed through a bandpass filter (ET500/50, Chroma), focused onto a photodetector (FDS100, Thorlabs), and amplified using a current preamplifier (SR570, Stanford Research Systems). Acquisition from the red fluorophore (tdTomato) was simultaneously acquired (bandpass filter ET605/70 nm, Chroma) but was not used for further analyses. The voltage signal from the preamplifier was digitized through a data acquisition board (PCI-e6321, National Instruments) at 1 kHz and stored in a computer using a custom software written in LabVIEW (National Instruments).

Histology. Mice were perfused with phosphate buffered saline (PBS) followed by 4% paraformaldehyde in PBS. The brains were cut in 100-μm coronal sections using a vibratome (Leica). Brain sections were loaded on glass slides and stained with DAPI (Vectashield). The locations of fiber and tetrode tips were determined using the standard mouse brain atlas.[76]

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

We used a t-test to compare between conditions (Figures 5 and S1). Kolmogorov-Smirnov test was used to check the normality assumption.

Fluorometry (photometry)

Power line noise in the raw voltage signals was removed by notch filter (MATLAB, Natick, MA, USA). A baseline of the voltage signal was defined by the lowest 10% of signals using a 2-min window. The baseline was subtracted from the raw signal, and the results were z-scored by a session-wide mean and standard deviation.

Licking and locomotion

Lick timing was defined as deflection points (peaks) of the output signals above a threshold. To plot the time course of licks, instantaneous lick rate was computed by a moving average using a 200-ms window.

Session-averaged time course

Licks, locomotion speed, and z-scored DA responses for individual trials were aligned by external events (e.g., trial start or teleport onset), and then smoothed using a moving average method. We did not smooth locomotion speed and fluorometry signals. The results were then averaged across trials for each experimental condition to generate a session-averaged time course.

Population-averaged time course

For calcium recording experiments, we computed the mean of session-averaged time courses from the second session dataset (as the average of all session averages) along with the standard error (the total number of sessions being the sample size) for each experimental condition. Population-average time courses are used to summarize behavior and DA responses.

Quantification for the darkening experiments

We quantified the z-scored DA sensor responses in the darkening experiment using three time windows (Figure 5E, shaded areas at the bottom). For the standard conditions, we used [0 s 0.4 s] from the trial start, [3.8 s 4.2 s] from the trial start, and [-0.4 s 0 s] from the reward onset. For the fast conditions, we used [0 s 0.4 s] from the trial start, [2.8 s 3.2 s] from the trial start, and [-0.4 s 0 s] from the reward onset.