Rethinking dopamine prediction errors

Matthew P.H. Gardner[1], Geoffrey Schoenbaum[1,2,3], and Samuel J. Gershman[4]
[1]Intramural Research Program of the National Institute on Drug Abuse, NIH
[2]Department of Anatomy and Neurobiology, University of Maryland School of Medicine
[3]Department of Neuroscience, Johns Hopkins School of Medicine [4]Department of Psychology and Center for Brain Science, Harvard University

Address for correspondence:
Samuel Gershman
Department of Psychology
Harvard University
52 Oxford St., room 295.05
Cambridge, MA 02138
Phone: 773-607-9817
E-mail: gershman@fas.harvard.edu

Abstract

Midbrain dopamine neurons are commonly thought to report a reward prediction error, as hypothesized by reinforcement learning theory. While this theory has been highly successful, several lines of evidence suggest that dopamine activity also encodes sensory prediction errors unrelated to reward. Here we develop a new theory of dopamine function that embraces a broader conceptualization of prediction errors. By signaling errors in both sensory and reward predictions, dopamine supports a form of reinforcement learning that lies between model-based and model-free algorithms. This account remains consistent with current canon regarding the correspondence between dopamine transients and reward prediction errors, while also accounting for new data suggesting a role for these signals in phenomena such as sensory preconditioning and identity unblocking, which ostensibly draw upon knowledge beyond reward predictions.

Introduction

The hypothesis that midbrain dopamine neurons report a reward prediction error (RPE, the discrepancy between observed and expected reward) enjoys a seemingly unassailable accumulation of support from electrophysiology (Bayer and Glimcher, 2005; Eshel et al., 2015,1; Roesch et al., 2007; Waelti et al., 2001), calcium imaging (Menegas et al., 2017; Parker et al., 2016), optogenetics (Chang et al., 2016; Steinberg et al., 2013; Tsai et al., 2009), voltammetry (Day et al., 2007; Hart et al., 2014), and human brain imaging (D'ardenne et al., 2008; Pessiglione et al., 2006). The success of the RPE hypothesis is exciting because the RPE is precisely the signal a reinforcement learning (RL) system would need to update reward expectations (Montague et al., 1996; Schultz et al., 1997). Support for this RL interpretation of dopamine comes from findings that dopamine complies with basic postulates of RL theory (Waelti et al., 2001), shapes the activity of downstream reward-predictive neurons in the striatum (Cheer et al., 2007; Day et al., 2007), and plays a causal role in the control of learning (Chang et al., 2016; Pessiglione et al., 2006; Steinberg et al., 2013; Tsai et al., 2009).

Despite these successes, however, there are a number of signs that this is not the whole story. First, it has long been known that dopamine neurons respond to novel or unexpected stimuli, even in the absence of changes in value (Horvitz, 2000; Ljungberg et al., 1992; Menegas et al., 2017; Strecker and Jacobs, 1985). While some theorists have tried to reconcile this observation with the RPE hypothesis by positing that value is affected by novelty (Kakade and Dayan, 2002) or uncertainty (Gershman, 2017), others have argued that this response constitutes a distinct function of dopamine (Bromberg-Martin et al., 2010a; Redgrave et al., 2008; Schultz, 2016), possibly mediated by an anatomically segregated projection from midbrain to striatum (Menegas et al., 2017). A second challenge is that some dopamine neurons respond to aversive stimuli. If dopamine responses reflect RPEs, then one would expect aversive stimuli to *reduce* responses (as observed in some studies; Mirenowicz and Schultz, 1996; Ungless et al., 2004). A third challenge is that dopamine activity (Bromberg-Martin et al., 2010b) and its putative hemodynamic correlates (Daw et al., 2011) are influenced by information, such as changes in stimulus contingencies, that should in principle be invisible to a pure "model-free" RL system that updates reward expectations using RPEs. This has led to elaborations of the RPE hypothesis according to which dopamine has access to some "model-based" information, for examples in terms of probabilistic beliefs or samples from a model-based simulator (Daw et al., 2006; Gershman, 2017; Gershman et al., 2014; Nakahara and Hikosaka, 2012; Nakahara et al., 2004; Starkweather et al., 2017).

While some of these puzzles can be resolved within the RPE framework by modifying assumptions about the inputs to and modulators of the RPE signal, recent findings have proven more unyielding. In this paper we focus on three of these findings: (1) dopamine transients are necessary for learning induced by unexpected changes in the sensory features of expected rewards (Chang et al., 2017); (2) dopamine neurons respond to unexpected changes in sensory features of expected rewards (Takahashi et al., 2017); and (3) dopamine transients are both sufficient and necessary for learning stimulus-stimulus associations (Sharpe et al., 2017). Taken together, these findings seem to contradict the RPE framework supported by so much other data.

Here we will suggest one possible way to reconcile the new and old findings, based on the idea that dopamine computes prediction errors over sensory features, much as was previously hypothesized for

rewards. This sensory prediction error (SPE) hypothesis is motivated by normative considerations: SPEs can be used to estimate a predictive feature map known as the *successor representation* (SR; Dayan, 1993). The key advantage of the SR is that it simplifies the computation of future rewards, combining the efficiency of model-free RL with some of the flexibility of model-based RL. Neural and behavioral evidence suggests that the SR is part of the brain's computational repertoire (Momennejad et al., 2017; Russek et al., 2017), possibly subserved by the hippocampus (Garvert et al., 2017; Stachenfeld et al., 2017). Here, building on the pioneering work of Suri (2001), we argue that dopamine transients previously understood to signal RPEs may instead constitute the SPE signal used to update the SR.

Theoretical framework

The reinforcement learning problem

RL theories posit an environment in which an animal accumulates rewards as it traverses a sequence of "states" governed by a transition function $T(s'|s)$, the probability of moving from state s to state s', and a reward function $R(s)$, the expected reward in state s. The RL problem is to predict and optimize *value*, defined as the expected discounted future return (cumulative reward):

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

where r_t is the reward received at time t in state s_t, and $\gamma \in [0, 1]$ is a discount factor that determines the weight of temporally distal rewards. Because the environment is assumed to obey the Markov property (transitions and rewards depend only on the current state), the value function can be written in a recursive form known as the *Bellman equation* (Sutton and Barto, 1998):

$$V(s_t) = \mathbb{E}[r_t + \gamma V(s_{t+1})]. \tag{2}$$

The Bellman equation allows us to define efficient RL algorithms for estimating values, as we explain next.

Model-free and model-based learning

Model-free algorithms solve the RL problem by directly estimating V from interactions with the environment. The Bellman equation specifies a recursive consistency condition that the value estimate $\hat{V}(s_t)$ must satisfy in order to be accurate. By taking the difference between the two sides of the Bellman equation, $\mathbb{E}[r_t + \gamma \hat{V}(s_{t+1})] - \hat{V}(s_t)$, we can obtain a measure of expected error; the direction and degree of the error is informative about how to correct $\hat{V}(s_t)$.

Because model-free algorithms do not have access to the underlying environment model (R and T) necessary to compute the expected error analytically, they typically rely on a stochastic sample of the error based on experienced transitions and rewards:

$$\delta_t = r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t). \tag{3}$$

This quantity, commonly known as the *temporal difference (TD) error*, will on average be 0 when the value function has been perfectly estimated. The TD error is the basis of the classic TD learning algorithm (Sutton and Barto, 1998), which in its simplest form updates the value estimate according to $\Delta\hat{V}(s_t) \propto \delta_t$. The RPE hypothesis states that dopamine reports the TD error (Montague et al., 1996; Schultz et al., 1997).

Model-free algorithms like TD learning are efficient because they *cache* value estimates, which means that state evaluation (and by extension action selection) can be accomplished by simply inspecting the values cached in the relevant states. This efficiency comes at the cost of flexibility: if the reward function changes at a particular state, the entire value function must be re-estimated, since the Bellman equation implies a coupling of values between different states. For this reason, it has been proposed that the brain also makes use of model-based algorithms (Daw and Dayan, 2014; Daw et al., 2005), which occupy the opposite end of the efficiency-flexibility spectrum. Model-based algorithms learn a model of the environment (R and T) and use this model to evaluate states, typically through some form of forward simulation or dynamic programming. This approach is flexible, because local changes in the reward or transition functions will instantly propagate across the entire value function, but at the cost of relying on comparatively inefficient simulation.

Some of the phenomena that we discuss in the Results have been ascribed to model-based computations supported by dopamine (Langdon et al., 2018), thus transgressing the clean boundary between the model-free function of dopamine and putatively non-dopaminergic model-based computations. The problem with this reformulation is that it is unclear what exactly dopamine is contributing to model-based learning. Although prediction errors are useful for updating estimates of the reward and transition functions used in model-based algorithms, these do not require a TD error. A distinctive feature of the TD error is that it bootstraps a future value estimate (the $\gamma\hat{V}(s_{t+1})$ term); this is necessary because of the Bellman recursion. But learning reward and transition functions in model-based algorithms can avoid bootstrapping estimates because the updates are local thanks to the Markov property.

To make this concrete, a simple learning algorithm (guaranteed to converge to the maximum likelihood solution under some assumptions about the learning rate) is to update the model parameters according to:

$$\Delta R(s) \propto r_t - R(s_t) \tag{4}$$

$$\Delta T(s'|s_t) \propto \mathbb{I}(s_{t+1} = s') - T(s'|s_t), \tag{5}$$

where $\mathbb{I}(\cdot) = 1$ if its argument is true, and 0 otherwise (cf. Gläscher et al., 2010). These updates can be understood in terms of prediction errors, but *not* TD errors (they do not bootstrap future value estimates). The TD interpretation is important for explaining phenomena like the shift in signaling to earlier reward-predicting cues (Schultz et al., 1997), the temporal specificity of dopamine responses (Hollerman and Schultz, 1998; Takahashi et al., 2016), and the sensitivity to long-term values (Enomoto et al., 2011). Thus, it remains mysterious how to retain the TD error interpretation of dopamine, which has been highly successful as an empirical hypothesis, while simultaneously accounting for the sensitivity of dopamine to SPEs.

The successor representation

To reconcile these data, we will develop the argument that dopamine reflects sensory TD errors, encompassing both reward and non-reward features of a stimulus. In order to introduce some context to this idea, let us revisit the fundamental efficiency-flexibility trade-off. One way to find a middle-ground between the extremes occupied by model-free and model-based algorithms is to think about different ways to *compile* a model of the environment. Model-based algorithms are maximally uncompiled: they explicitly represent the parameters of the model. Model-free algorithms are maximally compiled: they only represent the summary statistics (state values) that are needed for reward prediction. A third possibility is a partially compiled model. (Dayan, 1993) presented one such scheme, based on the following mathematical identity:

$$V(s_t) = \sum_{s'} M(s_t, s')R(s'), \tag{6}$$

where M denotes the successor representation (SR), the expected discounted future state occupancy:

$$M(s_t, s') = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \mathbb{I}(s_{t+k} = s')\right]. \tag{7}$$

Intuitively, the SR represents states in terms of the frequency of their successor states. From a computational perspective, the SR is appealing for two reasons. First, it renders value computation a linear operation, yielding efficiency comparable to model-free evaluation. Second, it retains some of the flexibility of model-based evaluation. Specifically, changes in rewards will instantly affect values because the reward function is represented separately from the SR. On the other hand, the SR will be relatively insensitive to changes in transition structure, because it does not explicitly represent transitions—these have been compiled into a convenient but inflexible format. Behavior reliant upon such a partially-compiled model of the environment should be more sensitive to reward changes than transition changes, a prediction recently confirmed in humans (Momennejad et al., 2017).

The SR obeys a recursion analogous to the Bellman equation:

$$M(s_t, s') = \mathbb{E}[\mathbb{I}(s_t = s') + \gamma M(s_{t+1}, s')]. \tag{8}$$

Following the logic of the previous section, this implies that a TD learning algorithm can be used to estimate the SR:

$$\Delta\hat{M}(s_t, s') \propto \delta_t^M(s') = \mathbb{I}(s_t = s') + \gamma\hat{M}(s_{t+1}, s') - \hat{M}(s_t, s'), \tag{9}$$

where $\hat{M}$ denotes the approximation of M.

One challenge facing this formulation is the *curse of dimensionality*: in large state spaces it is impossible to accurately estimate the SR for all states. Generalization across states can be achieved by defining the SR over state features and modeling this feature-based SR with linear function approximation:

$$\hat{M}(s_t, j) = \sum_i f_i(s_t)W_{ij}, \tag{10}$$

where $f_i(s)$ denotes the ith feature of state s and W is a weight matrix that parametrizes the approximation. In general the features can be arbitrary, but for the purposes of this paper, we will assume that the features correspond to distinct stimulus identities; thus $f_i(s) = 1$ if stimulus i is present in state s, and 0 otherwise. Linear function approximation leads to the following learning rule for the weights:

$$\Delta W_{ij} \propto \delta_t^M(j) f_i(s_t), \tag{11}$$

where

$$\delta_t^M(j) = f_j(s_t) + \gamma \hat{M}(s_{t+1}, j) - \hat{M}(s_t, j) \tag{12}$$

is the TD error under linear function approximation. We will argue that dopamine encodes this TD error.

One issue with comparing this vector-valued TD error to experimental data is that we don't yet know how particular dopamine neurons map onto particular features. In order to make minimal assumptions, we will assume that each neuron has a uniform prior probability of encoding any given feature. Under ignorance about feature tuning, the expected TD error is then proportional to the superposition of feature-specific TD errors, $\sum_j \delta_t^M(j)$. In our simulations of dopamine, we take this superposition to be the "dopamine signal" (see also (Daw et al., 2006)), but we wish to make clear that this is a provisional assumption that we ultimately hope to replace once the feature tuning of dopamine neurons is better understood.

There are several notable aspects of this new model of dopamine. First, it naturally captures SPEs, as we will illustrate shortly. Second, it also captures RPEs if reward is one of the features. Specifically, if $f_j(s_t) = r_t$, then the correspond column of the SR is equivalent to the value function, $M(s, j) = V(s)$, and the corresponding TD error is the classical RPE, $\delta_t^M(j) = \delta_t$. Third, the TD error is now vector-valued, which means that dopamine neurons may be heterogeneously tuned to particular features (as hypothesized by some authors, Lau et al., 2017), or they multiplex several features (Tian et al., 2016), or both. Notably, although the RPE correlate has famously been evident in single-units, representation of these more complex or subtle prediction errors may be an ensemble property.

Simulations

Some of the most direct evidence for our hypothesis comes from a recent study by Chang et al. (2017), who examined whether dopamine is necessary for learning about changes in reward identity (Figure 1A). Animals first learned to associate two stimuli (X_B and X_{UB}) with different reward flavors. These stimuli were then reinforced in compound with other stimuli (A_B and A_{UB}). Critically, the $X_{UB}A_{UB}$ trials were accompanied by a change in reward flavor, a procedure known as "identity unblocking" that attenuates the blocking effect (Blaisdell et al., 1997; McDannald et al., 2011; Rescorla, 1999). This effect eludes explanation in terms of model-free mechanisms, but is naturally accommodated by the SR since changes in reward identity induce sensory prediction errors. Chang et al. (2017) showed that optogenetic inhibition of dopamine at the time of the flavor

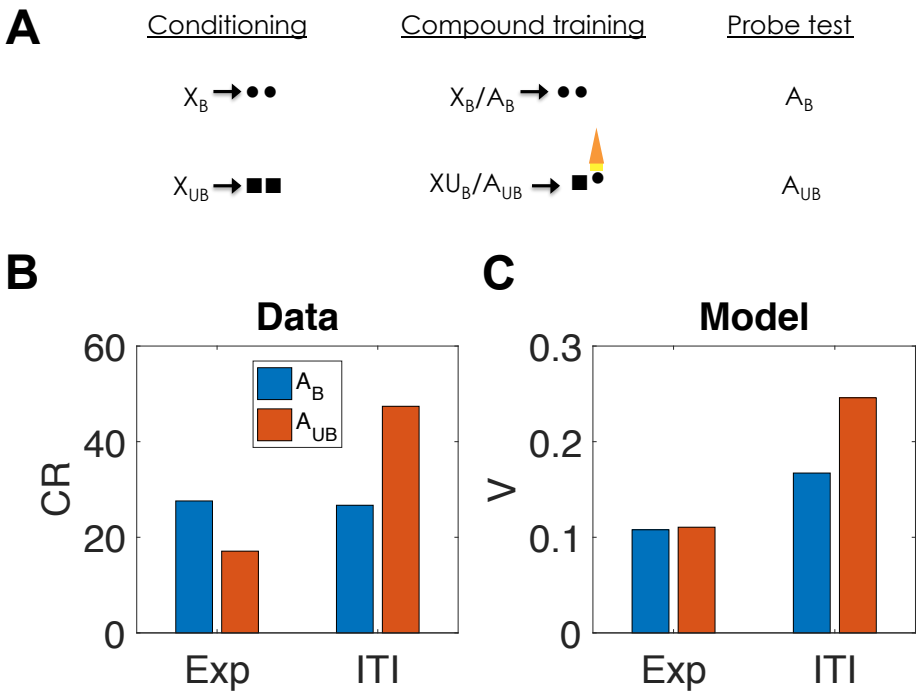

Figure 1: **Inhibition of dopamine neurons prevents learning induced by changes in reward identity**. (*A*) Identity unblocking paradigm. Circles and squares denote distinct reward flavors. Orange light symbol indicates when dopamine neurons were suppressed optogenetically to disrupt any positive SPE; this spanned a 5s period beginning 500ms prior to delivery of the second reward. (*B*) Conditioned responding on the probe test. Exp: experimental group, receiving inhibition during reward outcome. ITI: control group, receiving inhibition during the intertrial interval. Data replotted from Chang et al. (2017). (*C*) Model simulation.

change prevents this unblocking effect (Figure 1B). Our model accounts for this finding (Figure 1C), because inhibition suppresses SPEs that are necessary for driving learning.

Electrophysiological experiments have confirmed that dopamine neurons respond to changes in identity, demonstrating a neural signal that is capable of explaining the data from Chang et al. (2017). We have already mentioned the sizable literature on novelty responses, but the significance of this activity is open to question, because the animal's prior value expectation is typically unclear. A study reported by Takahashi et al. (2017) provides more direct evidence for an SPE signal, using a task (Figure 2A) in which animals experience both shifts in value (amount of reward) and identity (reward flavor). Takahashi and colleagues found that individual dopamine neurons exhibited the expected changes in firing to shifts in value (Figure 2B, reward addition and omission) and also showed a stronger response following a value-neutral change in reward identity (Figure 2B, identity switch), changes in firing similar to those predicted by the model under these conditions (Figure 2C).

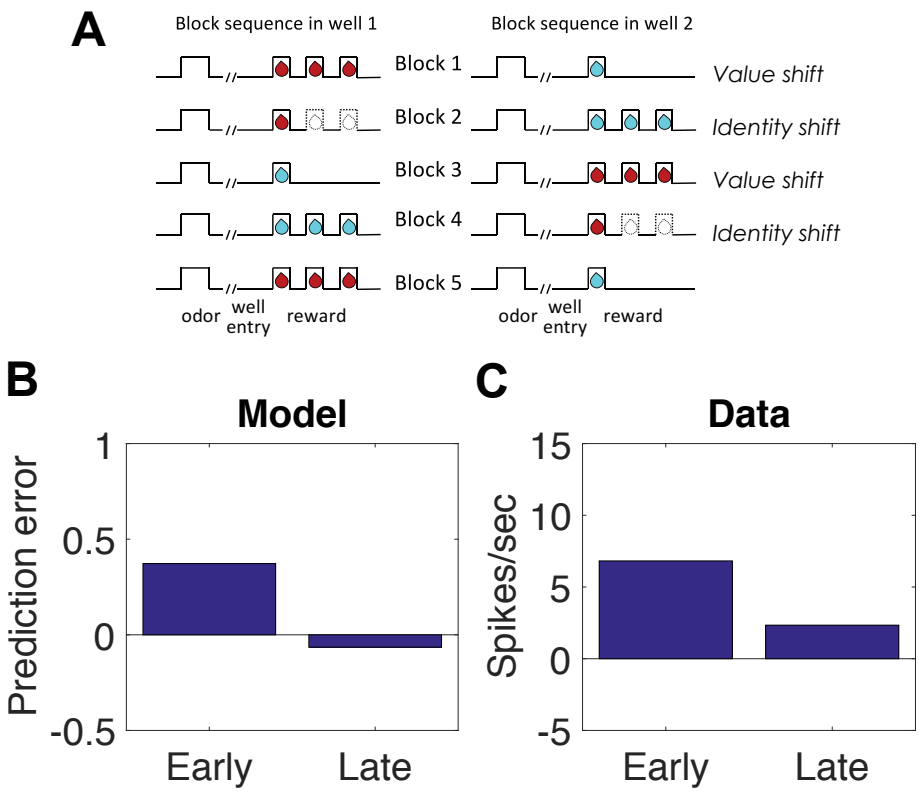

Figure 2: **Dopamine neurons respond to changes in reward identity**. (A) Time course of stimuli presented to the animal on each trial. Dashed indicate reward omission, solid lines indicate reward delivery. (B) Firing rate of dopamine neurons following first delivery of reward. Data replotted from Takahashi et al. (2017). (C) Model simulation of TD error.

A strong form of our proposal is that dopamine transients are both sufficient and necessary for learning stimulus-stimulus associations. Recent experiments using a sensory preconditioning paradigm (Sharpe et al., 2017) have tested this using sensory preconditioning. In this paradigm (Figure 3A), various stimuli and stimulus compounds (denoted A, EF, AD, AC) are associated with another stimulus X through repeated pairing in an initial preconditioning phase. In a subsequent conditioning phase, X is associated with reward (sucrose pellets). In a final probe test, conditioned responding to a subset of the individual stimuli (F, D, C) is measured in terms of the number of food cup entries elicited by the presentation of the stimuli. During the preconditioning phase, one group of animals received optogenetic activation of dopamine neurons via channelrhodopsin (ChR2) expressed in the ventral tegmental area of the midbrain. In particular, optogenetic activation was applied either coincident with the onset of X on AC→X trials, or (as a temporal control) 120-180 seconds after X on AD→X trials. Another control group of animals received the same training and optogenetic activation, but expressed light-insensitive enhanced yellow fluorescent protein (eYFP).

A blocking effect was discernible in the control (eYFP) group, whereby A reduced acquisition

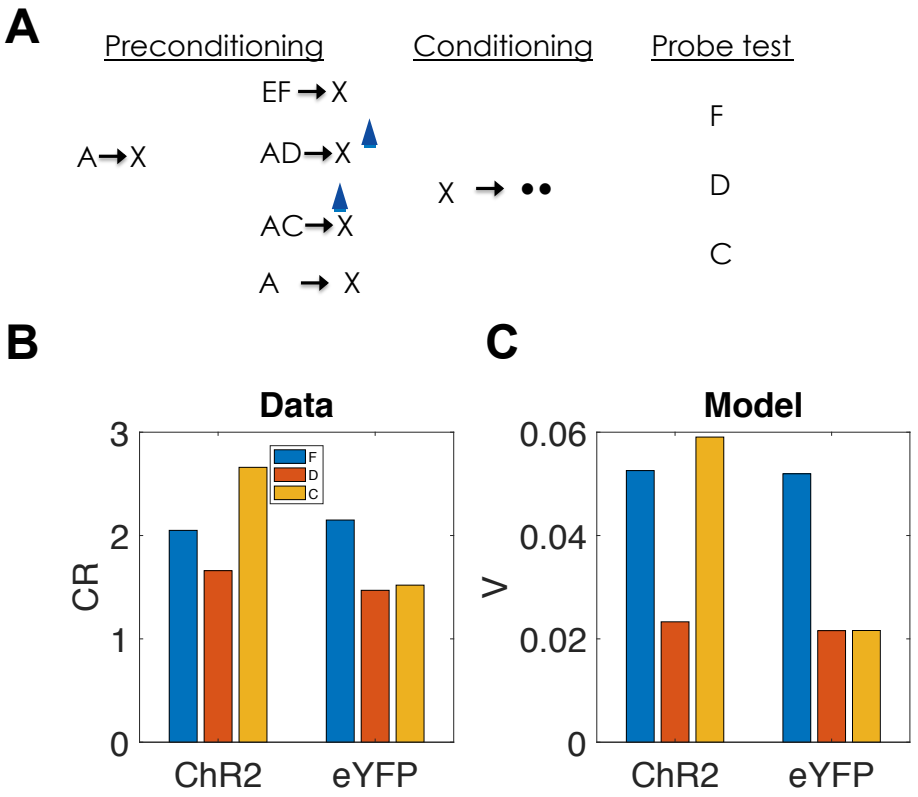

Figure 3: **Dopamine transients are sufficient for learning stimulus-stimulus associations**. (*A*) Sensory preconditioning paradigm. The initial preconditioning phase is broken down into two sub-phases. Letters denote stimuli, arrows denote temporal contingencies, and circles denote rewards. Blue light symbol indicates when dopamine neurons were activated optogenetically to mimic a positive SPE; this spanned a 1s period beginning at the start of X. (*B*) Number of food cup entries occurring during the probe test for experimental (ChR2) and control (eYFP) groups. Data replotted from Sharpe et al. (2017). (*C*) Model simulation, using the value estimate as a proxy for conditioned responding.

of conditioned responding to C and D, compared to F, which was trained in compound with a novel stimulus (Figure 3B). The blocking effect was eliminated by optogenetic activation in the experimental (ChR2) group, specifically for C, which received activation coincident with X. Thus, activation of dopamine neurons was sufficient to drive stimulus-stimulus learning in a temporally specific manner.

These findings raise a number of questions. First, how does one explain blocking of stimulus-stimulus associations? Second, how does one explain why dopamine affects this learning in the apparent absence of new reward information?

In answer to the first question, we can appeal to an analogy with blocking of stimulus-reward associations. The classic approach to modeling this phenomenon is to assume that each stimulus

acquires an independent association and that these associations summate when the stimuli are presented in compound (Rescorla and Wagner, 1972). While there are boundary conditions on this assumption (Soto et al., 2014), it has proven remarkably successful at capturing a broad range of learning phenomenon, and is inherited by TD models with linear function approximation (e.g., Gershman, 2017; Ludvig et al., 2008; Schultz et al., 1997). Summation implies that if one stimulus (A) perfectly predicts reward, then a second stimulus (C) with no pre-existing association will fail to acquire an association when presented in compound with A, because the sum of the two associations will perfectly predict reward and hence generate an RPE of 0. The same logic can be applied to stimulus-stimulus learning by using linear function approximation of the successor representation, which implies that stimulus-stimulus associations will summate and hence produce blocking, as observed in Sharpe et al. (2017).

In answer to the second question, we argue that dopamine is involved in stimulus-stimulus learning because it reflects a multifaceted SPE, as described in the previous section. By assuming that optogenetic activation adds a constant to the SPE (see Methods), we can capture the unblocking findings reported by Sharpe and colleagues (Figure 3C). The mechanism by which optogenetic activation induces unblocking is essentially the same as the one suggested by the results of Steinberg et al. (2013) for conventional stimulus-reward blocking: by elevating the prediction error, a learning signal is engendered where none would exist otherwise. However, while the results of Steinberg and colleagues are consistent with the original RPE hypothesis of dopamine, the results of Sharpe et al. (2017) cannot be explained by this model and instead require the analogous dopamine-mediated mechanism for driving learning with SPEs.

In addition to establishing the sufficiency of dopamine transients for learning, (Sharpe et al., 2017) also established their necessity, using optogenetic inactivation. In a variation of the sensory preconditioning paradigm (Figure 4A), two pairs of stimulus-stimulus associations were learned (A→X and B→Y). Subsequently, X and Y were paired with different reward flavors, and finally conditioned responding to A and B was evaluated in a probe test. In one group of animals expressing halorhodopsin in dopamine neurons (NpHR), optogenetic inhibition was applied coincident with the transition between the stimuli on B→Y trials. A control group expressing light-insensitive eYFP was exposed to the same stimulation protocol. Sharpe and colleagues found that inhibition of dopamine selectively reduced responding to B (Figure 4B), consistent with our model prediction that disrupting dopamine transients (a negative prediction error signal) should attenuate stimulus-stimulus learning (Figure 4C).

Limitations and extensions

One way to drive a wedge between model-based and model-free algorithms is to devalue rewards (e.g., through pairing the reward with illness or selective satiation) and show effects on previously acquired conditioned responses to stimuli that predict those rewards. Because model-free algorithms like TD learning need to experience unbroken stimulus-reward sequences to update stimulus values, the behaviors they support are insensitive to such reward devaluation. Model-based algorithms, in contrast, are able to propagate the devaluation to the stimulus without direct experience, and hence allow behavior to be devaluation-sensitive. Because of this, devaluation-sensitivity has frequently been viewed as an assay of model-based RL (Daw et al., 2005).

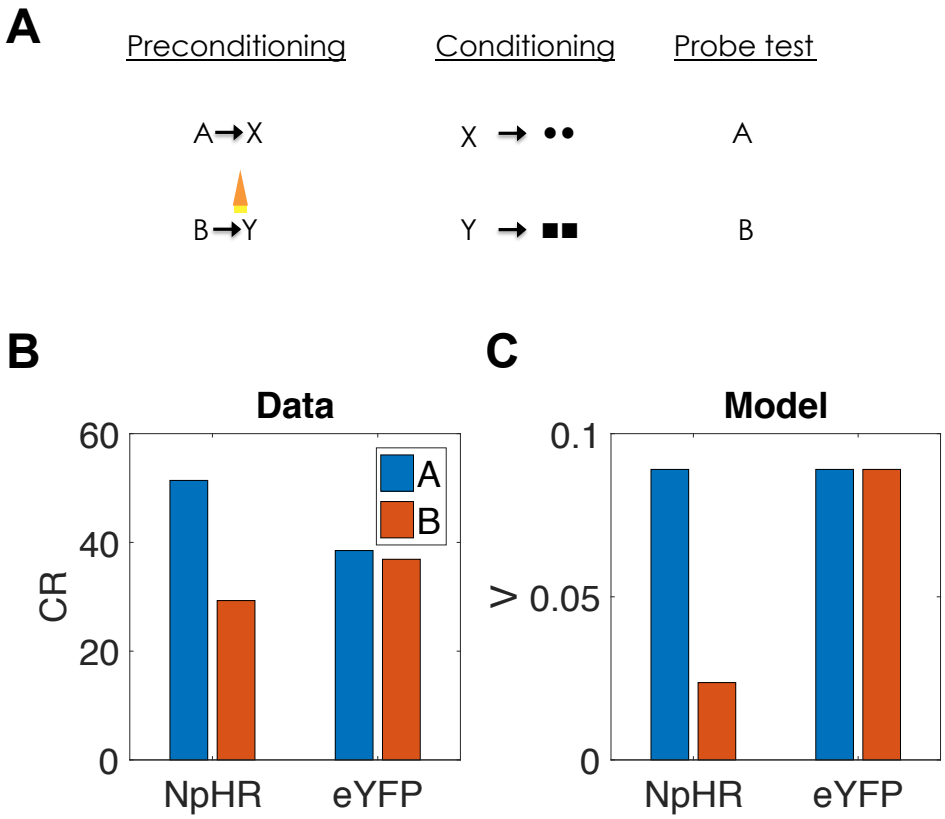

Figure 4: **Dopamine transients are necessary for learning stimulus-stimulus associations**. (*A*) Sensory preconditioning paradigm. Circles and squares denote distinct reward flavors. Orange light symbol indicates when dopamine neurons were suppressed optogenetically to disrupt any positive SPE; this spanned a 2.5s period beginning 500ms prior to the end of B. (*B*) Number of food cup entries occurring during the probe test for experimental (NpHR) and control (eYFP) groups. Data replotted from Sharpe et al. (2017). (*C*) Model simulation.

However, such sensitivity can also be a property of SR-based RL, since the SR represents the association between the stimulus and food and is also able to update the reward function of the food as a result of devaluation. Thus, like model-based accounts, an SR model can account for changes in previously learned behavior to reward-predicting stimuli after devaluation, both in normal situations (Momennejad et al., 2017; Russek et al., 2017) and when learning about those stimuli is unblocked by dopamine activation (Keiflin et al., 2017). However, the SR model cannot spontaneously acquire transitions between states that are not directly experienced (Momennejad et al., 2017; Russek et al., 2017). With this in mind, we consider the finding that reward devaluation alters the learning induced by activation of dopamine neurons in the sensory preconditioning paradigm of Sharpe et al. (2017).

A key aspect of the reward devaluation procedure is that the food was paired with illness after the end of the entire preconditioning procedure and in the absence of any of the stimuli (and in fact

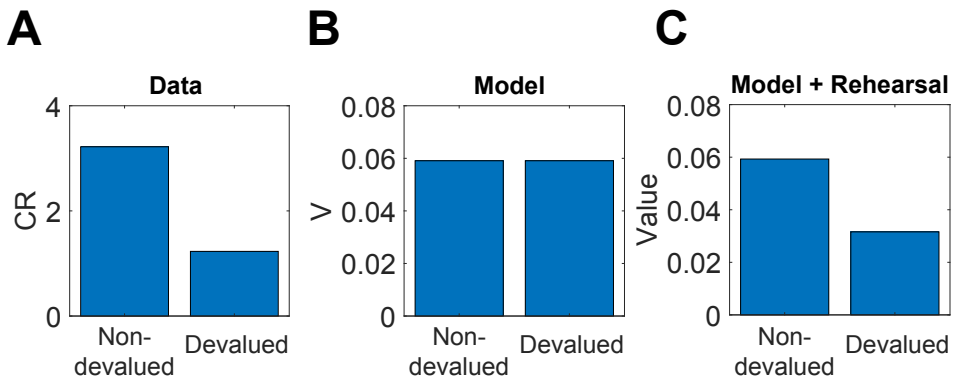

Figure 5: **Behavior to preconditioned cue that is unblocked by activation of dopamine neurons is sensitive to devaluation of the predicted reward**. Data (A, replotted from Sharpe et al., 2017) and model simulation (B) for conditioned responding to stimulus C in the probe test. Animals in the devalued group were injected with lithium chloride in conjunction with ingestion of the reward (sucrose pellets), causing a strong aversion to the reward. Animals in the nondevalued group were injected with lithium chloride approximately 6 hours after ingestion of the reward. (C) A version of the model with rehearsal of stimulus X during reward devaluation was able to capture the devaluation-sensitivity of animals.

not in the training chamber). In the SR model, only stimuli already predictive of the food can change their values after devaluation. In the paradigm of Sharpe and colleagues, X was associated with food but C was not. Moreover, C was associated with X before any association with food was established. Because of this, C is not updated in the SR model to incorporate an association with food. It follows that, unlike the animals in Sharpe et al. (2017), the model will not be devaluation-sensitive when probed with C (Figure 5B).

It is possible to address this failure within our theoretical framework in a number of different ways. One way we considered was to allow optogenetic activation to increment predictions for *all* possible features, instead of being restricted to recently active features by a *feature eligibility trace* (see Methods), as in the simulations thus far. With such a promiscuous artificial error signal, the model can recapitulate the devaluation effect, because C would then become associated with food (along with everything else) in the preconditioning phase itself. The problem with this work-around is that it also predicts that animals should develop a conditioned response to the food cup for all the cues during preconditioning, since food cup shaping prior to preconditioning seeds the food state with reward value. As a result, any cue paired with the food state immediately begins to induce responding at the food cup. Such behavior is not observed, suggesting that the artificial update caused by optogenetic activation of the dopamine neurons is locally restricted.

A second more conventional way to address this failure within our theoretical framework is to assume that there is some form of offline rehearsal or simulation that is used to update cached predictions (Gershman et al., 2014; Johnson and Redish, 2005; Pezzulo et al., 2014). Russek et al. (2017) have shown that such a mechanism is able to endow SR-based learning with the ability to retrospectively update predictions even in the absence of direct experience. A minimal implementation of such a mechanism in our model, simply by "confabulating" the presence of X during reward devaluation, is

sufficient to capture the effects of devaluation following optogenetic activation of dopamine neurons (Figure 5C). This solution makes the experimental prediction that the devaluation-sensitivity of this artificially unblocked cue should be time-dependent, under the assumption that the amount of offline rehearsal is proportional to the retention interval.

Discussion

The RPE hypothesis of dopamine has been one of theoretical neuroscience's signature success stories. This paper has set forth a significant generalization of the RPE hypothesis that enables it to account for a number of anomalous phenomena, without discarding the core ideas that motivated the original hypothesis. The proposal that dopamine reports an SPE is grounded in a normative theory of reinforcement learning (Dayan, 1993), motivated independently by a number of computational (Barreto et al., 2017; Russek et al., 2017), behavioral (Gershman et al., 2012; Momennejad et al., 2017; Smith et al., 2013) and neural (Brea et al., 2016; Garvert et al., 2017; Howard and Kahnt, 2018; Stachenfeld et al., 2017) considerations.

An important strength of the proposal is that it extends the functional role of dopamine beyond RPEs, while still accounting for the data that motivated the original RPE hypothesis. This is because, if reward is treated as a sensory feature, then one dimension of the vector-valued SPE will be the RPE. Indeed, dopamine SPEs should behave systematically like RPEs, except that they respond to features: they should pause when expected features are unexpectedly omitted, they should shift back to the earliest feature-predicting cue, and they should exhibit signatures of cue competition, such as overexpectation. SPEs are used to update cached predictions, analogous to the RPE in model-free algorithms. However these cached predictions extend beyond value to include information about the occupancy of future states (the SR). The SR can be used in a semi-flexible manner that allows behavior to be sensitive to changes in the reward structure, such as devaluation by pairing a reward with illness. As a result, even if dopamine is constrained by the model proposed here, it would support significantly more flexible behavior than supposed by classical model-free accounts (Montague et al., 1996; Schultz et al., 1997), even without moving completely to an account of model-based computation in the dopamine system (Langdon et al., 2018).

Nevertheless, the theory proposed here—particularly if it incorporates off-line rehearsal in order to fully explain the results of Sharpe et al. (2017)—strains the dichotomy between model-based and model-free algorithms that has been at the heart of modern RL theories (Daw et al., 2005). As noted earlier, SR requires offline rehearsal to incorporate the effects of devaluation after preconditioning in Sharpe et al or manipulations of the transition structures of tasks (Momennejad et al., 2017). If the effects of dopamine SPEs are mediated by some form of offline rehearsal or simulation, then we should be able to control the effects of dopamine by manipulating retention intervals or attention (see Gershman et al., 2014). For example, the strength of devaluation sensitivity in Sharpe et al should vary with the retention interval prior to the probe test. Another testable implication of the theory is that we should see heterogeneous tuning of dopamine neurons, reflecting the vector-valued nature of the SPE. These predictions set an exciting new agenda for dopamine research by embracing a broader conception of dopamine function.

Finally, while we have focused on dopamine in this paper, a complete account obviously needs to integrate the computational functions of several different brain regions. Two are particularly relevant: the hippocampus and the orbitofrontal cortex. Many lines of evidence are consistent with the idea that the hippocampus encodes a "predictive map" resembling the SR (Stachenfeld et al., 2017). For example, hippocampal place cells alter their tuning with repeated experience to fire in anticipation of future locations (Mehta et al., 2000), and fMRI studies have found predictive coding of non-spatial states (Garvert et al., 2017; Schapiro et al., 2016). The orbitofrontal cortex has also been repeatedly implicated in predictive coding, particularly of reward outcomes (e.g., Gottfried et al., 2003; Schoenbaum et al., 1998), but also of sensory events (Chaumon et al., 2013), and the orbitofrontal cortex is critical for sensory-specific outcome expectations in Pavlovian conditioning (Ostlund and Balleine, 2007). Wilson et al. (2014) have proposed that the orbitofrontal cortex encodes a "cognitive map" of state space, which presumably underpins this diversity of stimulus expectations. Thus, evidence suggests that both hippocampus and orbitofrontal cortex encode some form of predictive representation (see Wikenheiser and Schoenbaum, 2016, for a review), and dopaminergic modulation of these regions is well-established (Aou et al., 1983; Lisman and Grace, 2005). Whether these representations and their modulation by dopamine complies with the theoretical framework elaborated here or goes beyond it remains to be seen.

Methods

Linear value function approximation

Under the linear function approximation scheme described in the Results, the value function estimate is given by:

$$\hat{V}(s_t) = \sum_i f_i(s_t) \sum_j U(j) W_{ij}, \tag{13}$$

where $U(j)$ is the reward expectation for feature j, updated according to a delta rule:

$$\Delta U(j) = \alpha_U f_j(s_t)[r_t - \hat{V}(s_t)] \tag{14}$$

with learning rate α_U.

Excitatory and inhibitory asymmetry in the TD error term

There is a large body of evidence in associative learning suggesting an imbalance between excitatory and inhibitory learning (Bathellier et al., 2013; Konorski, 1948). Mirroring this imbalance is an asymmetry in the dynamic range of the firing rate of single dopaminergic neurons in the midbrain (Bayer and Glimcher, 2005). In accordance with these observations, we assume that the error terms (ΔW_{ij} and ΔU_j) are rescaled by a factor of 1/4 for negative prediction errors. This is equivalent to assuming separate learning rates for positive and negative prediction errors (see Collins and Frank, 2014).

Simulation parameters

We used the following parameters in the simulations: $\gamma = 0.95, \alpha_W = 0.06, \alpha_U = 0.03$, where α_W is the learning rate for the weight matrix W, α_U is the learning rate for the reward function, and γ is the discount rate. We used the same set of parameters across all simulations. However, our results are largely robust to variations in these parameters.

Modeling optogenetic activation and inhibition

Optogenetic intervention was modeled by modifying the TD error as follows:

$$\delta_t^M(j) = \begin{cases} (1+\eta)\delta_t^M(j) & \eta < 0 \\ [f_j(s_t)\eta + \delta_t^M(j) & \eta > 0 \end{cases} \tag{15}$$

where $\eta = 1.0$ for optogenetic activation and -0.8 for inhibition. The asymmetry between the functions for activation and inactivation was chosen to better match the the hypothesized function of optogenetic stimulation based on empirical findings. For positive stimulation of dopamine, it is thought that the increased dopamine activity should enhance learning with the currently active features, which in the SR model is the $f_j(s_t)$ term. For optogenetic inhibition of dopamine, we have found that punctate versus prolonged inhibition causes differential effects, with punctate inhibition resulting in negative prediction errors and prolonged inhibition resulting in shunting of the error signal citepchang16. Our inhibition in the experiments included in this paper were prolonged, necessitating a different model of the inhibitory optogenetic manipulation.

Acknowledgments

We are grateful to Brian Sadacca and Andrew Wikenheiser for helpful discussions. This work was supported by the National Institutes of Health (CRCNS 1R01MH109177 to S.J.G.) and the Intramural Research Program at NIDA ZIA-DA000587 (to G.S.). The opinions expressed in this article are the authors' own and do not reflect the view of the NIH/DHHS.

References

Aou, S., Oomura, Y., Nishino, H., Inokuchi, A., and Mizuno, Y. (1983). Influence of catecholamines on reward-related neuronal activity in monkey orbitofrontal cortex. *Brain Research*, 267:165–170.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., Silver, D., and van Hasselt, H. P. (2017). Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4056–4066.

Bathellier, B., Tee, S. P., Hrovat, C., and Rumpel, S. (2013). A multiplicative reinforcement learning model capturing learning dynamics and interindividual variability in mice. *Proceedings of the National Academy of Sciences*, 110:19950–19955.

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

Blaisdell, A. P., Denniston, J. C., and Miller, R. R. (1997). Unblocking with qualitative change of unconditioned stimulus. *Learning and Motivation*, 28:268–279.

Brea, J., Gaál, A. T., Urbanczik, R., and Senn, W. (2016). Prospective coding by spiking neurons. *PLoS Computational Biology*, 12(6):e1005003.

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010a). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, 68:815–834.

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., and Hikosaka, O. (2010b). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*, 104:1068–1076.

Chang, C. Y., Esber, G. R., Marrero-Garcia, Y., Yau, H.-J., Bonci, A., and Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nature Neuroscience*, 19:111–116.

Chang, C. Y., Gardner, M., Di Tillio, M. G., and Schoenbaum, G. (2017). Optogenetic blockade of dopamine transients prevents learning induced by changes in reward features. *Current Biology*, 27:3480–3486.

Chaumon, M., Kveraga, K., Barrett, L. F., and Bar, M. (2013). Visual predictions in the orbitofrontal cortex rely on associative content. *Cerebral Cortex*, 24:2899–2907.

Cheer, J. F., Aragona, B. J., Heien, M. L., Seipel, A. T., Carelli, R. M., and Wightman, R. M. (2007). Coordinated accumbal dopamine release and neural activity drive goal-directed behavior. *Neuron*, 54:237–244.

Collins, A. G. and Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, 121:337–366.

D'ardenne, K., McClure, S. M., Nystrom, L. E., and Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, 319:1264–1267.

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

Daw, N. D. and Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Phil. Trans. R. Soc. B*, 369:20130478.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69:1204–1215.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8:1704–1711.

Day, J. J., Roitman, M. F., Wightman, R. M., and Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature Neuroscience*, 10:1020–1028.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5:613–624.

Enomoto, K., Matsumoto, N., Nakai, S., Satoh, T., Sato, T. K., Ueda, Y., Inokawa, H., Haruno, M., and Kimura, M. (2011). Dopamine neurons learn to encode the long-term value of multiple future rewards. *Proceedings of the National Academy of Sciences*, 108:15462–15467.

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.

Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience*, 19:479–486.

Garvert, M. M., Dolan, R. J., and Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, 6:e17086.

Gershman, S. J. (2017). Dopamine, inference, and uncertainty. *Neural Computation*, 29:3311–3326.

Gershman, S. J., Markman, A. B., and Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, 143:182–194.

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., and Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, 24:1553–1568.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66:585–595.

Gottfried, J. A., O'Doherty, J., and Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, 301:1104–1107.

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

Hollerman, J. R. and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, 96:651–656.

Howard, J. D. and Kahnt, T. (2018). Identity prediction errors in the human midbrain update reward-identity expectations in the orbitofrontal cortex. *Nature Communications*, 9:1611.

Johnson, A. and Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks*, 18:1163–1171.

Kakade, S. and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, 15:549–559.

Keiflin, R., Pribut, H. J., Shah, N. B., and Janak, P. H. (2017). Phasic activation of ventral tegmental, but not substantia nigra, dopamine neurons promotes model-based pavlovian reward learning. *bioRxiv*.

Konorski, J. (1948). Conditioned reflexes and neuron organization.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., and Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology*, 49:1–7.

Lau, B., Monteiro, T., and Paton, J. J. (2017). The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia. *Current Opinion in Neurobiology*, 46:241–247.

Lisman, J. E. and Grace, A. A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron*, 46:703–713.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67:145–163.

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 Computation*, 20:3034–3054.

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., and Schoenbaum, G. (2011). Ventral

striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience*, 31:2700–2705.

Mehta, M. R., Quirk, M. C., and Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, 25:707–715.

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.

Mirenowicz, J. and Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, 379:449–451.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1:680–692.

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

Nakahara, H. and Hikosaka, O. (2012). Learning to represent reward structure: A key to adapting to complex environments. *Neuroscience Research*, 74:177–183.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron*, 41:269–280.

Ostlund, S. B. and Balleine, B. W. (2007). Orbitofrontal cortex mediates outcome encoding in pavlovian but not instrumental conditioning. *Journal of Neuroscience*, 27:4819–4825.

Parker, N. F., Cameron, C. M., Taliaferro, J. P., Lee, J., Choi, J. Y., Davidson, T. J., Daw, N. D., and Witten, I. B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature Neuroscience*, 19:845–854.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442:1042–1045.

Pezzulo, G., van der Meer, M. A., Lansink, C. S., and Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Sciences*, 18:647–657.

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews*, 58:322–339.

Rescorla, R. A. (1999). Learning about qualitatively different outcomes during a blocking procedure. *Learning & Behavior*, 27:140–151.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. and Prokasy, W., editors, *Classical Conditioning II: Current Research and theory*, pages 64–99. Appleton-Century-Crofts, New York, NY.

Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, 10:1615–1624.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., and Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, 13:e1005768.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., and Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, 26:3–8.

Schoenbaum, G., Chiba, A. A., and Gallagher, M. (1998). Orbitofrontal cortex and basolateral

amygdala encode expected outcomes during learning. *Nature Neuroscience*, 1:155–159.

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*, 17:183–195.

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

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., Niv, Y., and Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, 20.

Smith, T. A., Hasinski, A. E., and Sederberg, P. B. (2013). The context repetition effect: Predicted events are remembered better, even when they dont happen. *Journal of Experimental Psychology: General*, 142:1298–1308.

Soto, F. A., Gershman, S. J., and Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review*, 121:526–558.

Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, 20:1643–1653.

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

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

Strecker, R. E. and Jacobs, B. L. (1985). Substantia nigra dopaminergic unit activity in behaving cats: effect of arousal on spontaneous discharge and sensory evoked activity. *Brain Research*, 361:339–350.

Suri, R. E. (2001). Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. *Experimental Brain Research*, 140:234–240.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., and Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron*, 95:1395–1405.

Takahashi, Y. K., Langdon, A. J., Niv, Y., and Schoenbaum, G. (2016). Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat vta depends on ventral striatum. *Neuron*, 91:182–193.

Tian, J., Huang, R., Cohen, J. Y., Osakada, F., Kobak, D., Machens, C. K., Callaway, E. M., Uchida, N., and Watabe-Uchida, M. (2016). Distributed and mixed information in monosynaptic inputs to dopamine neurons. *Neuron*, 91:1374–1389.

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., De Lecea, L., and Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science*, 324:1080–1084.

Ungless, M. A., Magill, P. J., and Bolam, J. P. (2004). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, 303:2040–2042.

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, 412:43–48.

Wikenheiser, A. M. and Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, 17:513–523.

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81:267–279.