Learning and Expression of Dopaminergic Reward Prediction Error via Plastic Representations of Time

Ian Cone[1,2,3], Claudia Clopath[1], Harel Z. Shouval[2,4]

[1]Department of Bioengineering, Imperial College London, London, United Kingdom; [2]Department of Neurobiology and Anatomy, University of Texas Medical School at Houston, Houston, TX; [3]Applied Physics Program, Rice University, Houston, TX; [4]Department of Electrical and Computer Engineering, Rice University, Houston, TX

Abstract

Dopamine (DA) releasing neurons in the midbrain learn response patterns that represent reward prediction error (RPE). Typically, models proposing a mechanistic explanation for how dopamine neurons learn to exhibit RPE are based on temporal difference (TD) learning, a machine learning algorithm. However, mechanistic models motivated by TD learning face two significant hurdles. First, TD based models typically require rather unrealistic components, such as long and robust temporal chains of feature specific neurons which tile, a priori, each interval from a given stimulus to a given reward. Secondly, various predictions of TD clash with experimental observations of how RPE evolves over learning. Here, we present a biophysically plausible plastic network model of spiking neurons, that learns RPEs and can replicate results observed in multiple experiments. This model, coined FLEX (**F**lexibly **L**earned **E**rrors in E**x**pected Reward), learns feature specific representations of time, allowing for neural representations of stimuli to adjust their timing and relation to rewards in an online manner. Following learning, model dopamine neurons in FLEX report a distribution of response types, as observed experimentally and as used in machine learning. Dopamine firing in our model reflects an RPE before and after learning but not necessarily during

learning, allowing the model to reconcile seemingly inconsistent experiments and make unique predictions that contrast those of TD.

Introduction

The discovery that the activity of midbrain dopamine releasing neurons (DA) can reflect something akin to an algorithmic reward prediction error (RPE) has provided fundamental insight into how the brain learns, and has formed a link between the "teaching signals" of biological and algorithmic reinforcement learning[1–9]. However, a biologically realistic mechanism that allows dopamine neurons to learn to represent RPE remains elusive. Moreover, experimentally observed dynamics of DA neurons during learning are at odds with the dynamics predicted by the most prominent algorithm for learning RPEs, Temporal Difference (TD) learning.

The main hurdle of implementing TD in a biophysically realistic manner is figuring out how to represent the temporal basis upon which the association between cue and reward occurs. Previous attempts have commonly utilized tag delay lines[1,10,11] (or the so-called complete serial compound or CSC), or "microstimuli", [12,13], to act as a temporal basis. In the CSC conception, each timestep is represented by one neuron, so a delay period between cue and reward is encoded by a sequential chain of neurons, each firing for a given duration. In the case of microstimuli, increasingly wide Gaussians act as temporal basis functions to tile the delay. Though networks constructed from these techniques are capable of learning RPEs, they can encounter catastrophic failures in the presence of noise and/or spiking neurons. While there is some evidence that striatal populations could act as neural correlates of the "microstimuli"[14] there is limited experimental evidence to support the pre-existence of task-specific, CSC-like temporal

bases in the brain prior to training. Additionally, for such theoretical models the size of the network N scales poorly with the time to be encoded T ($N \alpha T$ in the case of the CSC).

Beyond issues with potential biophysical implementations, different variants of TD learning make contrasting predictions about the dynamics of learning RPE. For traditional TD(0) methods, each trial consists of the RPE iteratively traveling backwards in time, starting at the time of the US and eventually settling at the time of the CS[1,9]. In intermediate trials, then, an RPE "bump" is located somewhere in the delay period. Such "backpropagation" (not to be confused with backpropagation learning) of the RPE through the delay period has been frequently contradicted by experimental results[5,15–18]. Experiments instead suggest that the RPE signals in basic trace conditioning are constrained to the time of the CS and/or US. A recent study has observed what appears to be backpropagation of the RPE during early training trials[19], suggesting that more investigation is needed. For now, however, the preponderance of evidence supports the idea that there is no backpropagating bump of RPE during learning. A modification of TD learning, TD(λ)[20], utilizes memory traces that decay with rate λ in order to assign credit to distal rewards. In turn, this variant of TD does not exhibit a significant backpropagating "bump" of RPE during intermediate trials[5]. However, even TD(λ) has certain restrictions (such as the total RPE never surpassing the initial RPE to the unexpected reward) that seem to contradict experimental results[3,5,21].

To address these issues, we designed a spiking network model (**F**lexibly **L**earned **E**rrors in E**x**pected Reward, or FLEX) that learns dopaminergic reward prediction error via local plasticity rules and internally generates representations of time, bypassing the TD method entirely. In FLEX, the temporal basis is not assumed to exist *a priori*. Instead, the

network develops so-called "Timer" cells that, following trace conditioning, encode feature-specific cue-reward delays via their persistent activity. These flexible Timer cells are consistent with experimental observations of neurons in V1[22–24], mPFC[25] and OFC[26]. The activity of the Timers themselves is shaped by learning, enabling a flexible encoding of the delay period(s). This allows a network of a fixed size to encode times on the order of hundreds of milliseconds to seconds, bypassing the N α T scaling inherent in the CSC. This flexible encoding of the delay period also facilitates temporally specific suppression of expected rewards. Additionally, since the dynamics of learning in FLEX are fundamentally different from those utilizing TD-like methods (including TD(0) and TD(λ)) , there is no "backpropagating" bump of RPE, and CS-evoked and US-evoked DA cell responses evolve on different timescales. The learning rule reaches fixed points on the order of tens to hundreds of trials and the model can encode delays on the order of seconds. We demonstrate that FLEX can replicate the classical profile of dopaminergic RPE observed from trace conditioning and sequential conditioning experiments, while also accounting for more recent observations of distributed "optimistic" and "pessimistic" RPEs[6,27]. Finally, FLEX makes unique testable predictions regarding the evolution of DA neuron firing during learning that contrast those made by TD.

Results

Biophysically Inspired Architecture Allows for Flexible Encoding of Time

Rather than building its temporal basis from a tag-delay line, external clock, or an artificial neural network (such as an LSTM[28]), the model we present here internally and flexibly learns durations by emulating distinct classes of cell responses that have been observed in trace conditioning experiments in medial prefrontal cortex (mPFC) and

primary visual cortex (V1)[22–26]. The first class of neurons observed ("Timers") are feature-specific and learn to maintain persistently elevated activity that spans the delay period between cue and reward, eventually decaying at the time of reward (real or expected). "Sustained Decrease" cells (not shown) are observed to be essentially the inverse of the Timers, maintaining decreased activity throughout the delay period. Finally, the "Messengers" have an activity profile that peaks at the time of real or expected reward. This set of cells form what has been coined a "Core Neural Architecture" (CNA)[29], a potentially canonical package of neural temporal representations. A slew of previous studies have shown the CNA to be a robust phenomenon experimentally[22–24], and computational work has demonstrated that the CNA can be used to learn and recall single temporal intervals[29], Markovian and non-Markovian sequences[30]. For simplicity, our model omits the Sustained Decrease cells and treats connections between populations within the CNA as fixed (previous work has shown that such a construction is robust to perturbation of these weights[29,30]).

We use a previously established learning rule (Two-trace learning or TTL[24,30–32]) based on competition between two Hebbian activated eligibility traces, one associated with LTD and one associated with LTP (see Methods). We use this rule because it solves the temporal credit assignment problem inherent in trace conditioning, reaches stable fixed points, and two traces have been experimentally observed in trace conditioning tasks[24]. In theory, other methods capable of solving the temporal credit assignment problem (such as a rule with a single eligibility trace[33]) could also be used to facilitate learning in FLEX, but owing to its functionality and experimental support, we choose to utilize TTL for this work.

The network's full architecture (visualized in **Figure 1a**) consists of two separate regions - the VTA and a cortical region such as the mPFC or OFC which houses feature-specific CNAs and thereby provides input to the VTA (we hereafter will call this cortical region PFC for simplicity. See Discussion for more details). Both regions (PFC and VTA) consist of leaky-integrate-and-fire (LIF) neurons that are separated into distinct populations, within and between which there are both fixed and learned synaptic connections.

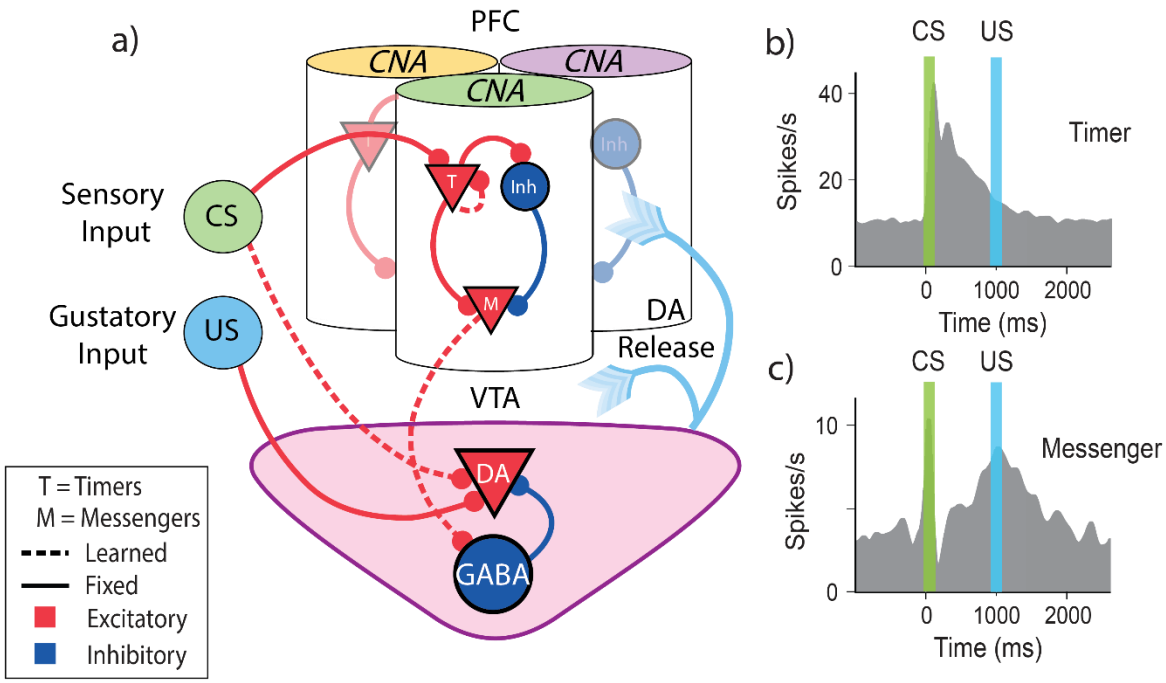


Figure 1 – Biophysically Inspired Architecture Allows for Flexible Encoding of Time
a) Diagram of model architecture. CNAs (visualized here as columns) located in PFC are selective to certain sensory stimuli (indicated here by color atop the column) via fixed excitatory inputs (CS). VTA DA neurons receive fixed input from naturally positive valence stimuli, such as food or water reward (US). DA neuron firing releases dopamine, which acts as a learning signal for both PFC and VTA. Solid lines indicate fixed connections, while dotted lines indicate learned connections. **b,c)** Schematic representation of data adapted from Liu et al. 2015**. b)** Timers learn to characteristically decay at the time of cue-predicted reward. **c)** Messengers learn to have a firing peak at the time of cue-predicted reward.

We begin by modelling VTA, within which we include both dopaminergic and GABAergic neurons. The DA neurons receive fixed gustatory inputs from naturally appetitive stimuli such as food and water, while also potentially receiving learned inputs from reward-predictive neutral external sensory cues. The GABAergic neurons provide inhibition to the DA neurons through fixed connections to the DA neurons. In the absence of external stimuli, both the DAergic and GABAergic neurons exhibit spontaneous firing, with the mean DA baseline rate hovering around ~5 Hz. DA neuron firing releases dopamine, which acts to mediate a universal reinforcement signal D(t) for all learned connections in the model.

D(t) is a simple piecewise linear function of the DA neuron firing rate $f_D(t)$ with a threshold θ around the baseline firing rate f_0:

Equation 1 – Dopamine reinforcement

$$D(t) = \begin{cases} f_D(t) - (f_0 - \theta), & f_D(t) \leq f_0 - \theta \\ 0, & f_0 - \theta < f_D(t) < f_0 + \theta \\ f_D(t) - (f_0 + \theta), & f_D(t) \geq f_0 + \theta \end{cases}$$

Note that the dopaminergic reinforcement can be both positive and negative, even though actual DA neuron firing (and the subsequent release of dopamine neurotransmitter) can itself only be positive. The bipolar nature of D(t) implicitly assumes that background tonic levels of dopamine do not modify weights, and that changes in synaptic efficacies are a result of a departure (positive or negative) from this background level. The neutral region around f_0 provides robustness to small fluctuations in firing rates which are inherent in spiking networks.

We integrate this VTA architecture with a model of PFC (**Figure 1a**), which contains feature-specific "columns", each housing a CNA microcircuit. A given microcircuit

consists of a pool of excitatory and inhibitory neurons. Fixed connections between these neurons divide them into three functional populations: two excitatory (Timers (T) and Messengers (M)) and one inhibitory (Inh). Though the connections between these populations are fixed, previous work has shown that they, along with the populations they mediate (T, M, Inh), can be learned from random initial weights[29]. Alongside these fixed connections are learned recurrent synaptic weights within the Timer population, which flexibly adjust the persistent firing time of the Timers to match the given cue-reward delay (Methods, Supplemental Methods). As the T population modifies its timing to span the delay period, the M population "follows along" owing to its T and Inh inputs, resulting in the observed activity profiles that match those seen in experiment (**Figure 1b,c**) [22–24]. A more detailed discussion and analysis of the requirements, functionality, and behavior of the CNA microcircuit is explored in previous work[29,30].

FLEX combines these models of PFC and VTA, allowing them to provide each other with mutual feedback. In short, rewarding stimuli trigger excess dopamine release in VTA, which acts as a reinforcement signal D(t). PFC uses this reinforcement to learn feature specific representations which are responsible for suppressing the excess dopamine produced by "expected" rewards back to baseline. This crosstalk is the crux of FLEX and provides a framework for how phasic reward prediction errors and representations of task information can co-evolve. Our model's construction implies a unique potential interpretation of RPE-like dopamine neuron responses. In FLEX, these RPE-like dopaminergic signals are a necessary component of learning novel cue-reward associations but are not themselves associations, as is the case with purely model-free methods such as TD. Instead, cue-reward associations emerge in higher order areas like PFC/OFC with the guidance of RPE-like dopaminergic reinforcement.

To understand the mechanisms of FLEX in further detail and demonstrate their efficacy, we replicate several different experimental results with simulations. We begin by considering the case of simple trace conditioning, where a CS is followed reliably by a US. In doing so, we also examine the distributional nature of RPEs resulting from FLEX. Then, we test the model's ability to learn multiple reward-predictive cues with a sequential conditioning protocol. Finally, we examine how our FLEX's dynamics during learning diverge from classic results of the TD class of methods.

CS-evoked and US-evoked Dopamine Responses Evolve on Different Timescales

First, we test the ability of FLEX to replicate basic trace conditioning results, where a single conditioned stimulus is presented, followed by an unconditioned stimulus at a fixed delay of one second (**Figure 2a**). The evolution of FLEX over training is mediated by learning (via TTL) in three sets of weights: Timer → Timer, Messenger → VTA GABA, and CS → VTA DA. These learned connections encode the feature-specific cue-reward delay, the temporally specific suppression of US-evoked dopamine, and the emergence of CS-evoked dopamine, respectively.

Upon presentation of the cue, feature-specific Timers are excited, producing Hebbian-activated eligibility traces at their recurrent synapses. When the reward is subsequently presented one second later, the excess dopamine it triggers acts as a reinforcement signal for these eligibility traces (via D(t), see **Equation 1**), causing the Timers' recurrent connections (and therefore persistent firing time) to increase (**Figure 2b**). Over repeated trials of this cue-reward pairing, the Timers' recurrent connections continue to increase until they approach their fixed points, and in turn, the Timers' persistent firing time increases until it spans the delay between CS and US

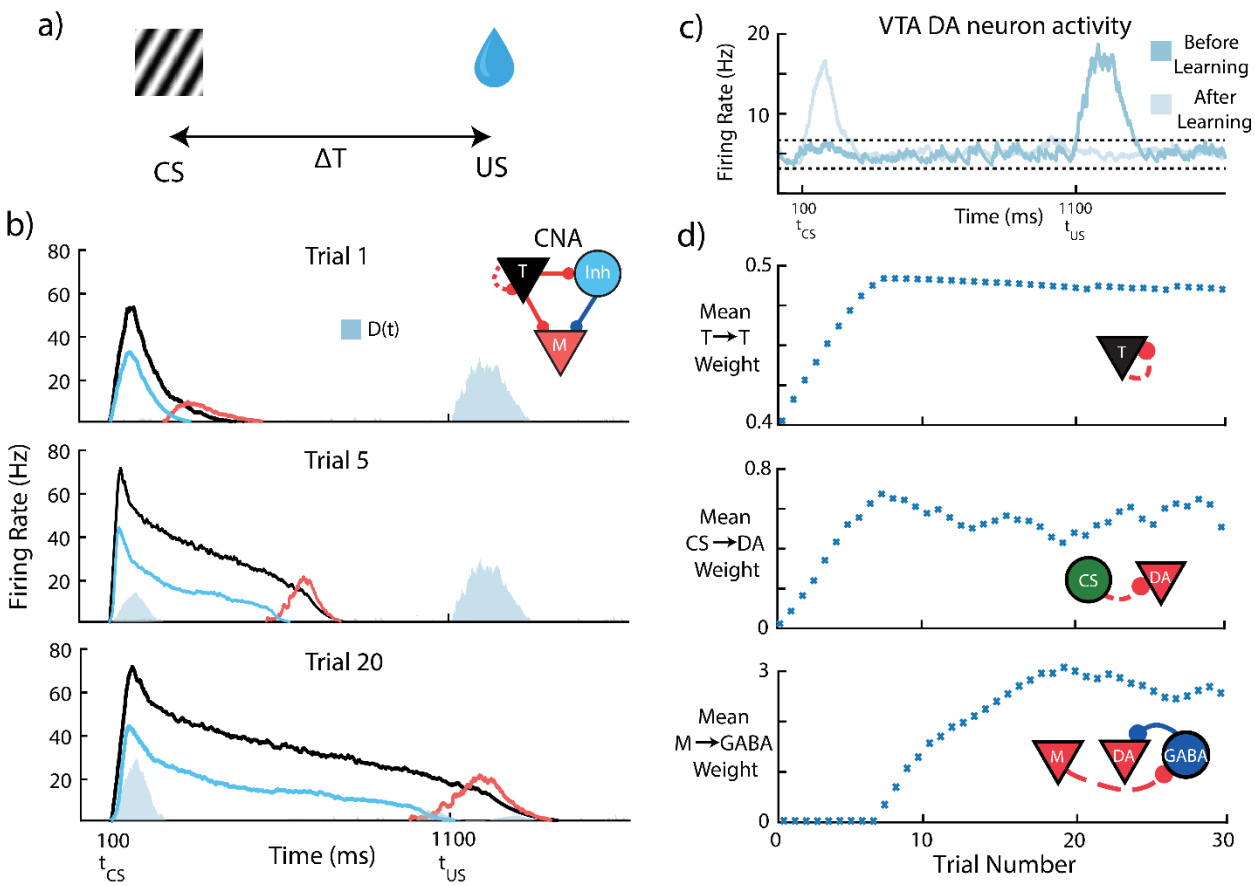


Figure 2 - CS-evoked and US-evoked Dopamine Responses Evolve on Different Timescales
a) Presentation of a given CS (such as a visual grating) is reliably followed by a reward at a time delay ΔT. Here, the CS is presented at 100ms and the reward at 1100ms. **b)** Mean firing rates for the CNA (see inset for colors), along with the dopaminergic reinforcement signal D(t)(in shaded blue), for three different stages of learning. **c)** Mean firing rate over all DA neurons taken before (blue) and after (light blue) learning. Firing above or below thresholds (dotted lines) evokes positive or negative D(t) in PFC (as shown in b)). **d)** Evolution of synaptic weights over the course of learning. Top, middle, and bottom, mean strength of T→T, CS→DA, and M→GABA synapses, respectively.

(**Supplemental Figure 1** and Supplemental Methods). These mature Timers provide a feature-specific representation of the expected cue-reward delay.

As the Timer population modifies its timing to span the delay period, the Messengers are "dragged along", since, owing to the dynamics of the Messengers' inputs (T and Inh), the Messengers themselves only fire at the end of the Timers' firing envelope. Eventually, the Messengers overlap with tonic background activity of VTA GABAergic

neurons at the time of the US (t_{US}) (**Figure 2b**). When combined with the dopamine release at t_{US}, this Hebbian overlap triggers learning at the Messenger → VTA GABA synapses (see **Figure 2c,** Methods, and Supplemental Methods). Because of the temporal specificity of the Messengers, this learned inhibition of the DA neurons (through excitation of the VTA GABAergic neurons) is effectively restricted to a short time window around the US and acts to suppress DA neural activity at t_{US} back to baseline.

Concurrent with these processes, feed-forward connections from the CS to the DA neurons increase, causing the model to develop a CS-evoked dopamine response. Again, this feed-forward learning uses dopamine release at t_{US} as the reinforcement signal to convert the Hebbian activated CS → DA eligibility traces into synaptic changes (**Supplemental Figure 2**). The emergence of excess dopamine at the time of the CS (t_{CS}) owing to these potentiated connections also acts to maintain them at a non-zero fixed point, so CS-evoked dopamine persists long after US-evoked dopamine has been suppressed to baseline (see Supplemental Methods).

In FLEX, the inhibition of the US (via M→GABA learning) occurs only after the Timers have learned the delay period (since M and GABA firing must overlap to trigger learning), giving the potentiation of the CS response time to occur first. This is in contrast to predictions of TD(0) learning where the RPE signal "travels" backwards over trials[1,9], or TD(λ) where the RPE is transferred from one time to the other[5,20]. In our model these two processes happen on two different timescales; the bump in DA neuron activity does not propagate back in time step by step, nor does the depression of the US occur in an equal and opposite manner to the excitation of the CS. This means the DA

neural firing in FLEX approximates RPE following successful learning, but DA neural firing and RPE are not necessarily equivalent during the learning period. At intermediate learning stages (e.g. trial 5, **Figure 2b,d**), the CS-evoked dopamine response (or equivalently, the CS → DA weights) has begun to potentiate to its maximum while the US-evoked dopamine response (or equivalently, inverse of the M → US weights) has only been slightly depressed, temporarily breaking from an optimal RPE signal. While this phenomenon has been seen in experiments[3,5,21,35], it has not been widely commented on – in FLEX, this is a fundamental property of the dynamics of learning. We therefore consider this transient divergence of FLEX's DA neuron firing from RPE a strong prediction of our model regarding potential learning dynamics in the dopaminergic system.

If an expected reward is omitted in FLEX, the resulting DA neuron firing will be inhibited at that time, demonstrating the characteristic dopamine "dip" seen in experiments (**Supplemental Figure 3**)[1]. This phenomenon occurs in our model because the previous balance between excitation and inhibition (from the US and GABA neurons, respectively) is disrupted when the US is not presented. The remaining input at t_{US} is therefore largely inhibitory, resulting in a transient drop in firing rates. If the CS is consistently presented without being paired with the US, the association between cue and reward is unlearned, since the consistent negative D(t) at the time of the US causes depression of CS→DA weights (**Supplemental Figure 3**).

Dopamine Responses are Sparse and Distributional

While the classical single neuron results of Schultz and others suggested that DA neurons are almost completely lacking excess firing at the time of expected

reward[1,10,36,37], more recent calcium imaging studies have revealed that a complete suppression of the US response is not universal[4,38]. Rather, many optogenetically identified dopamine neurons maintain a response to the US while developing a response to the CS. Others still never learn a response to the CS or have no response to either stimulus. Furthermore, recent work has shown that a distribution of RPEs is beneficial to reinforcement learning algorithms (so called "distributional RL")[6,27].

Interestingly, FLEX also exhibits this diversity due to a) the inherent non-linearity of its spiking neurons, and b) sparse projections from the CS to the VTA and from the US to VTA.

We examine this heterogeneity in detail (**Figure 3**), highlighting the individual neuron responses over the course of learning a simple trace conditioning task with a fixed reward value (**Figure 3a**). Most neurons in FLEX do indeed demonstrate some level of RPE, with many very closely matching the complete RPE found by Shultz et al.[1,10,36,37]. However, the degree of CS/US responsiveness of each neuron within the population falls within a distribution, akin to that found in more recent calcium imaging experiments[4,38]. To quantify this distribution, we define a measure $\beta = \frac{r_{CS} - r_{US}}{r_{CS} + r_{US}}$, where r_{CS} is the trial-averaged firing rate at the time of the CS, and r_{US} the trial-averaged firing rate of a DA neuron at the time of the US. Positive values of β indicate the neuron responds preferentially to the CS, while negative values indicate the neuron responds preferentially to the US. Before learning, most neurons have a β close to -1, almost exclusively responding to the US. Over the course of learning, however, this distribution shifts towards positive values, with most of the population being more sensitive to the CS following training (**Figure 3b**).

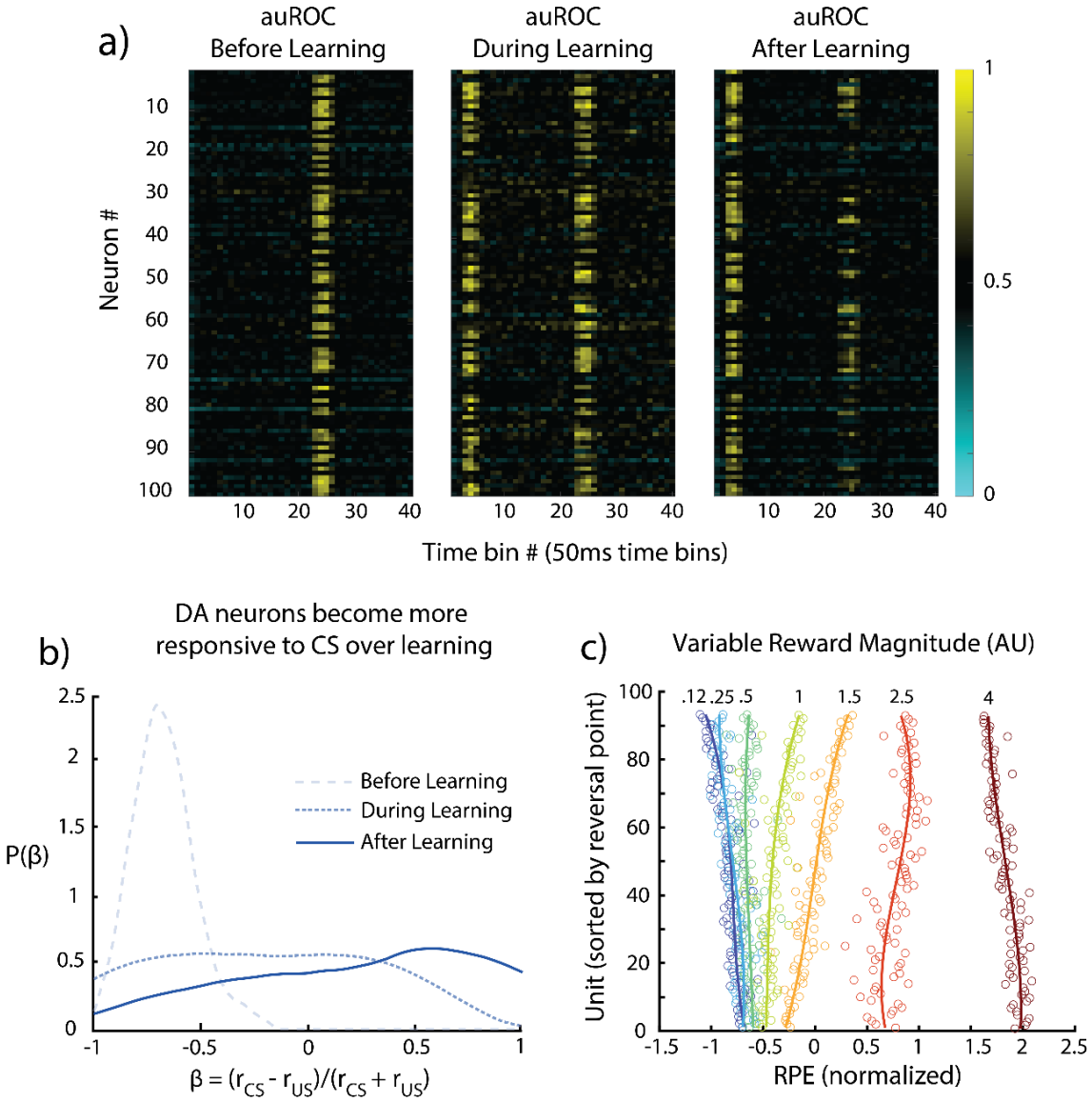


Figure 3 – Dopamine Responses are Sparse and Distributional

a) Area under receiver operating characteristic (auROC, see Supplemental Methods) for all VTA neurons in our model for 15 trials before (left, US only), 15 trials during (middle, CS+US), and 15 trials after (right, CS+US) learning. Values above and below 0.5 indicate firing rates above and below the baseline distribution. In most neurons, the CS response emerges before the US response depresses. Most, but not all, neurons display classical Schultz RPE before/after learning. The heterogeneous distribution of single neuron responses reflect experimental results[4] **b)** Distributions of CS/US responsiveness (as defined by $\beta = \frac{r_{CS} - r_{US}}{r_{CS} + r_{US}}$) before, during, and after learning. After learning the neural population is more skewed towards firing in response to the CS. **c)** Normalized RPEs measured in response to seven different reward magnitudes (indicated by color) in a trained network. Units are sorted by their interpolated reversal point (the reward for which their reported RPE would be zero). In line with previous observations[6], "optimistic" and "pessimistic" neurons differ not only in their reversal point, but also in their relative RPEs to rewards above and below their reversal point (See also **Supplemental Figure 4**).

We next detail the behavior of the network when trained and tested on a similar trace conditioning task, but this time with variable reward magnitudes following a cue at a fixed delay. Previous studies have shown that DA neurons in VTA demonstrate so-called "optimistic" and "pessimistic" value encoding[6] when exposed to a distribution of reward magnitudes. Pessimistic neurons expect less reward, so display a positive RPE (at t_{US}) to rewards above, at, or even below the mean reward. Optimistic neurons, on the other hand, will only display a positive RPE (at t_{US}) above some value larger than the mean reward. This results in a different "reversal point" for each neuron, which is the actual value of reward for which the neuron reports no RPE.

Mathematically, the measures of optimism or pessimism have previously been described via the differing slopes of linear fits to a particular neuron's positive or negative RPE responses (these slopes are defined α^+ and α^-, respectively)[6]. After training and testing FLEX with variable reward magnitudes, neurons within the population develop different levels of optimism or pessimism towards the expected reward at t_{US} (**Figure 3c**). Further, each neuron's ratio $\frac{\alpha^+}{\alpha^+ + \alpha^-}$ (defined as the "asymmetry" of the a given neuron's RPE), falls within a distribution centered around ~0.5 (**Supplemental Figure 4a**). Moreover, this asymmetry is strongly predictive of a neuron's reversal point (**Supplemental Figure 4b**), indicating that the distributional nature of FLEX is not simply a spurious artifact of noise or bias. Instead, FLEX arrives at distributional representations of RPE implicitly through its structure and the inherent non-linear F-I curve of its spiking neurons. The observed behavior of the model demonstrates that a network with non-linear units can encode the basics of

distributional RPEs, without resorting to modifications of existing RL algorithms such as TD learning that are specifically constructed in order to produce these distributions[6,27].

Model Reconciles Differing Experimental Phenomena Observed during Sequential Conditioning

Standard trace conditioning experiments with multiple cues (CS1→CS2→US) have generally reported the so-called "serial transfer of activation" – that dopamine neurons learn to fire at the time of the earliest reward-predictive cue, "transferring" their initial activation at t_{US} back to the time of the first conditioned stimulus[2,36]. However, other results have shown that the DA neural responses at the times of the first and second conditioned stimuli (t_{CS1} and t_{CS2}, respectively) evolve together, with both CS1 and CS2 becoming predictive of the reward[5]. We will show that the dynamics of learning in FLEX provides a framework which can reconcile both these results.

To account for multiple cues, FLEX's architecture makes use of multiple "columns" within the PFC, each containing a stimulus specific CNA (refer to **Figure 1a**). First, we train the model using a standard multiple cue paradigm (**Figure 4a**), always presenting the sequence CS1→CS2→US. During learning, the mean firing rates of our model DA neurons go through five distinct stages – initialization(i), acquisition(ii), reward suppression(iii), serial transfer of activation(iv), and finally, overtraining(v) (**Figure 4b,i-v**). In the initialization stage(i), an unexpected reward is presented, producing a lone RPE. During acquisition(ii), reward-predictive cues begin to trigger elevated dopamine firing upon their presentation. Following around fifty to one hundred trials of training, the dopamine of the expected reward itself is suppressed(iii). The serial transfer of activation(iv) occurs soon after the reward suppression, as CS1 becomes

predictive of CS2. Finally, after many trials following the serial transfer of activation, overtraining(v) occurs and the cue-evoked dopamine returns to baseline.

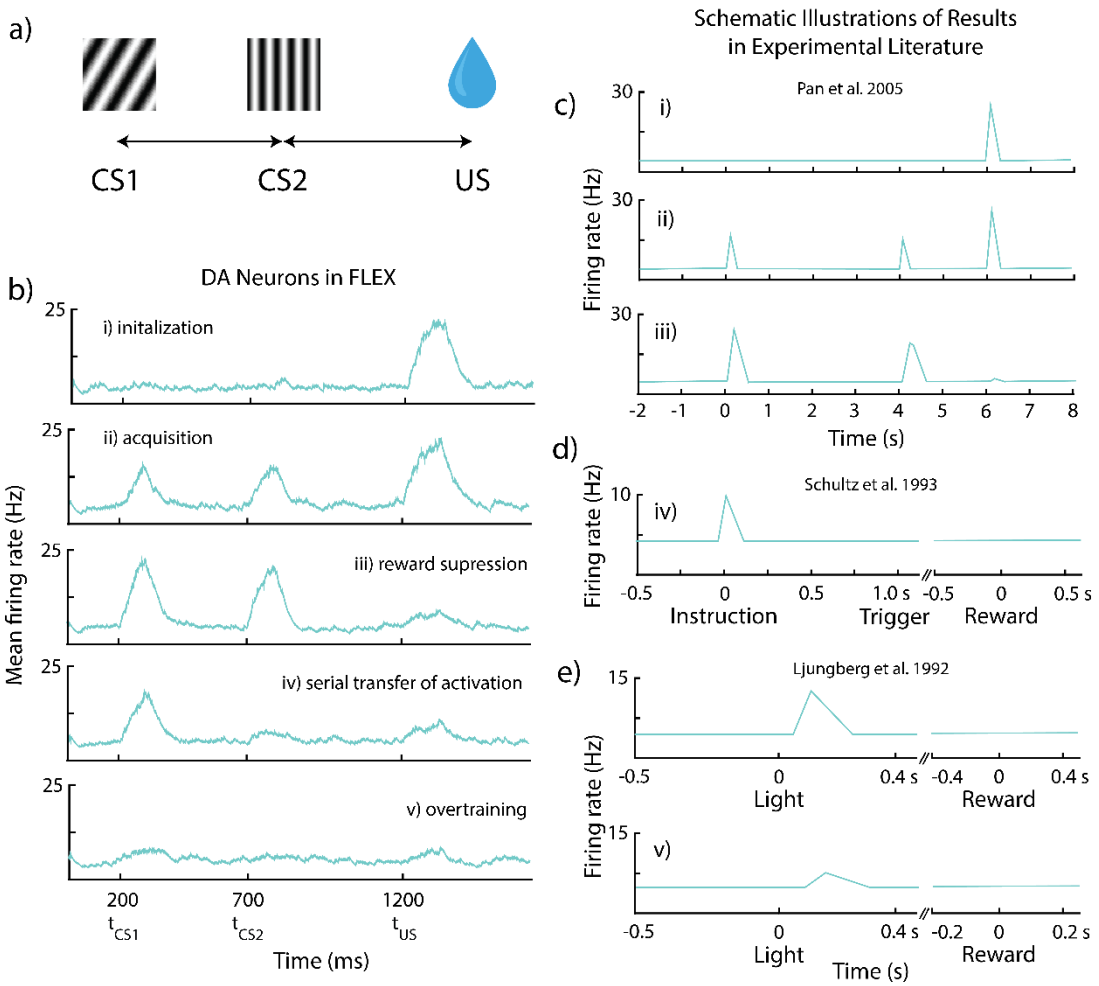


Figure 4 – Model Reconciles Differing Experimental Phenomena Observed during Sequential Conditioning

Results from "sequential conditioning", where sequential neutral stimuli CS1 and CS2 and are paired with delayed reward US. **a)** Visualization of protocol. In this example, the US is presented starting at 1500ms, with CS1 is presented starting at 100ms, and CS2 is presented starting at 800ms. **b)** Mean firing rates over all DA neurons, for five distinctive stages in learning – initialization(i), acquisition(ii), reward depression(iii), serial transfer of activation(iv), and overtraining(v). **c-e)** Schematic illustration of experimental results from recorded dopamine neurons, labeled with the matching stage of learning in our model. **c)** DA neuron firing before (top), during (middle) and after (bottom) training, wherein two cues (0s and 4s) were followed by a single reward (6s). Adapted from Pan et al. 2005. **d)** DA neuron firing after training wherein two cues (instruction, trigger) were followed by a single reward. Adapted from Schultz et al. 1993. **e)** DA neuron firing after training (top) and after overtraining (bottom). Adapted from Ljungberg et al. 1992.

The first three stages match the observations of Pan et al. 2005[5] (**Figure 4c**), wherein both neutral cues evoke dopamine release and both cues became predictive of reward. To show that both cues are indeed predictive of reward at this stage, we omit CS2 on a single trial, and observe, as in experiment[5], that the US is only partially inhibited (by CS1) (**Supplemental Figure 5**). However, after many trials (100 or so for our chosen learning rate), FLEX begins to demonstrate serial transfer of activation, and the dopamine response to the second neutral cue is slowly depressed. This occurs because once the DA neurons have developed a positive response to both cues, the dopamine release profile of the second cue is similar to the dopamine release by US. Therefore, the CS2 acts mechanistically as a US to CS1. As such, FLEX then evolves as in the standard trace conditioning case, which results in inhibition of activity to the in response to CS2. In this phase, the DA neural activity resembles that of Schultz et al. 1993[2]. Many have posited that this transfer to the earliest reward-predicting cue is a fundamental property of *all* sequential cue learning in the dopaminergic system[2,36]. Our model would agree with this hypothesis, with the caveat that for certain training protocols, intermediate stages in learning can persist for many trials, during which both cues trigger dopamine release and are predictive of reward. It is also worth noting that the order in which the sequential cues are learned in the model is important – if one performs training such that CS2→US is fully learned prior to presentation of CS1→CS2→US, FLEX's RPE signal will travel back to t_{CS1} in stepwise fashion from t_{US}→t_{CS2}→t_{CS1}, bypassing the intermediate phases observed by Pan et al (**Supplemental Figure 6**).

Many more trials after the serial transfer of activation occurs (hundreds to thousands for our chosen learning rate), we observe "overtraining" in the model, whereby the

dopamine response to the reward predicting CS is eventually depressed back down to baseline, even though both the CS and US are still being presented, in line with experimental findings[3,21]. This overtraining occurs in FLEX when the CS→DA weights are knocked out of their non-zero fixed point's attractive basin by an exceptional instance (or sequence of instances) of noise. At the levels of noise included in the model, these events occurs once every few hundred to few thousand trials.

Regarding other forms of sequential conditioning, FLEX is also capable of replicating the results of blocking and unblocking experiments (**Supplemental Figure 7**). In this protocol, the network is initially trained on a standard trace conditioning task. Once the cue-reward association is learned completely, a second cue is inserted in between the initial cue and the reward, and learning is repeated. As in experiments, this intermediate cue does not become reward predictive ("blocking")[39–41], unless additional reward or dopamine is introduced to the system ("unblocking")[42,43].

Altogether, FLEX presents unique theoretical results in sequential conditioning tasks, which, as we will examine below, diverge from the results stemming from traditional methods. Of particular interest is our model's prediction that seemingly mutually exclusive results, in which serial transfer of activation is either observed[2] or not[5], are actually compatible with each other when considering the training methods and number of trials.

Dynamics of FLEX Diverge from those of TD During Sequential Conditioning

While our results are compatible with those seen in experiment, they are incompatible with those predicted by traditional forms of TD. Most obviously, we do not observe the backpropagation of the RPE seen in TD(0). However, a variant of TD learning, TD(λ),

which uses an effective memory trace (with decay rate λ), produces a much shallower bump (or none at all) of a backpropagating RPE. We compare FLEX to TD(λ) (**Figure 5**), training on the previously described multiple cue task (see **Figure 4a**). We set the parameters of our TD(λ) model to match those in earlier work[5] and approximate the cues and reward as Gaussians centered at t_{CS1}, t_{CS2}, and t_{US}, respectively.

We observe several key defining differences between FLEX's dynamics and those of TD(λ). First, FLEX demonstrates distinct periods of stability, potentiation and depression, some of which can last for hundreds of trials (**Figure 5a**), rather than the gradual trial by trial shift of RPE seen in TD(λ) (**Figure 5b**).

Second, in TD models, the total RPE during a given intermediate trial is always less than or equal to the original RPE provided by the initial unexpected presentation of reward. In both TD(0) and TD(λ), this quantity is conserved; for versions of TD with a temporal discounting factor γ (which acts such that a reward with value V presented n timesteps in the future is only worth γ^nV, where γ ≤ 1), this quantity decreases as learning progresses. In FLEX, total dopaminergic reinforcement D(t) (FLEX's analogue of RPE) during a given trial can be greater than that evoked by the original unexpected US (see **Figure 5c**), and therefore during training FLEX diverges from a true reward prediction error. This effect is perhaps best illustrated by looking at individual trials before and during learning in both FLEX (**Figure 5d**) and TD(λ) (**Figure 5e**). This observation is a more general formulation of our earlier prediction that the evolution of the CS responses can occur independently of (and before) depression of the US response.

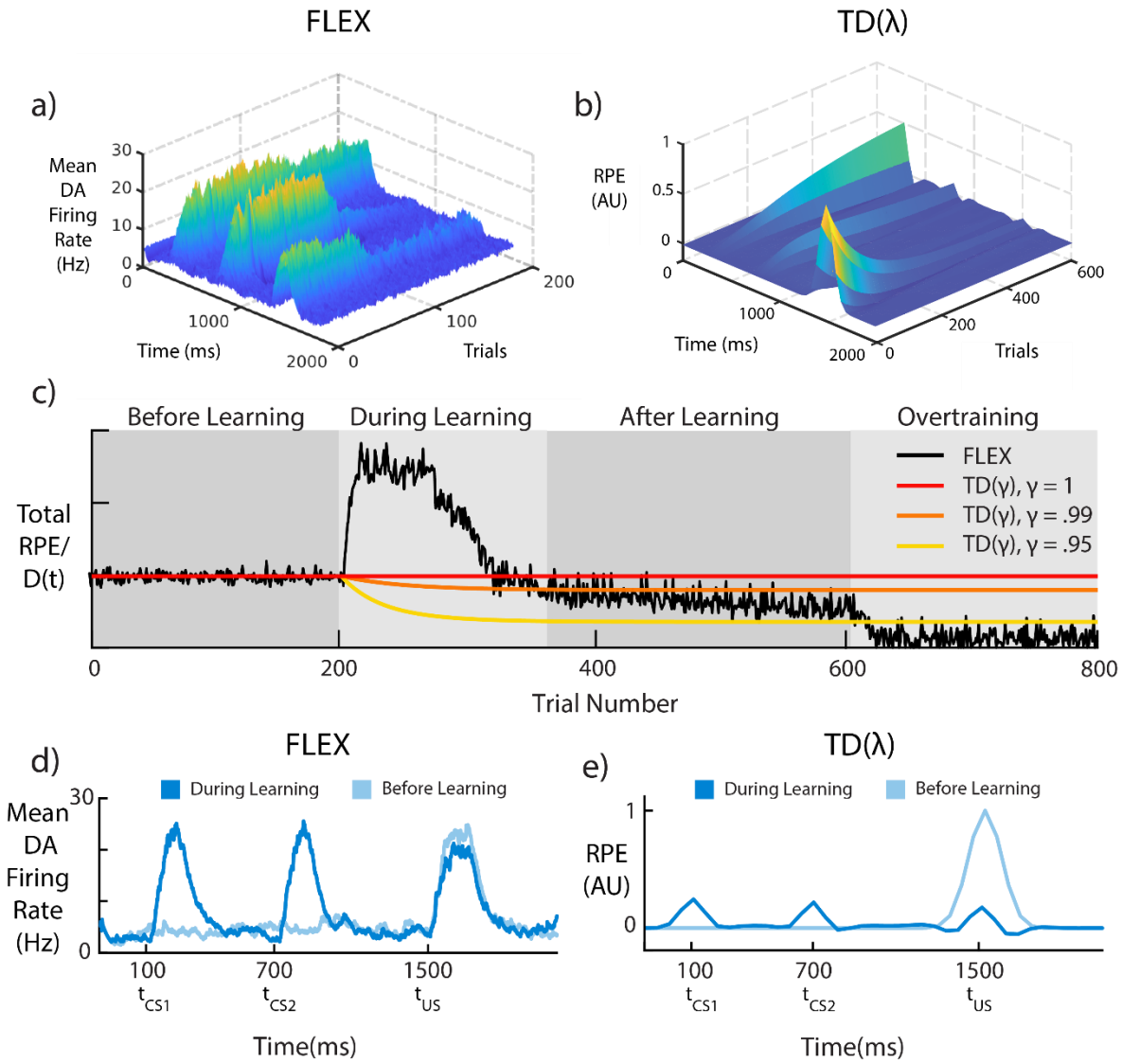


Figure 5 – FLEX Model Dynamics Diverge from those of TD During Sequential Conditioning

Dynamics of both TD(λ) and FLEX when trained with the same sequential conditioning protocol in Figure 4, where stimuli CS1 and CS2 are introduced starting on the first trial. US is presented starting at 1500ms, CS1 is presented starting at 100ms, and CS2 is presented starting at 700ms. **a)** Mean firing rate of DA neurons in FLEX over the course of the first 200 trials of training. **b)** Evolution of TD(λ) RPE over the course of training. **c)** Sum total dopaminergic reinforcement D(t) during a given trial of training in FLEX (black), and sum total RPE in three instances of TD(λ) with discounting parameters γ = 1, γ = .99, and γ = .95 (red, orange, and yellow, respectively). Shaded areas indicate functionally different stages of learning in FLEX. 200 trials of US-only presentation (before learning) are included for comparison with subsequent stages. **d)** Average firing rate of DA neurons in FLEX, and **e)** average RPE in TD(λ), during trial 25 (light blue) and trial 225 (dark blue).

Finally, in FLEX we observe overtraining after presenting the same CS1→CS2→US protocol for hundreds to thousands of trials (see **Figure 5c**, overtraining in this simulation occurs after ~600 trials, see also **Figure 4b,v**), a phenomenon observed experimentally[3,21]. In contrast, TD cannot account for overtraining because in an overtrained state, the activity of DA neurons is not representing either an accurate or a discounted reward prediction error. Furthermore, even though FLEX is no longer releasing dopamine following overtraining, it still contains an indirect prediction of reward, via the cue-reward delay encoded within the PFC.

FLEX therefore has a unique prediction that following overtraining, the omission of an expected reward will produce a negative RPE (**Supplemental Figure 8**). This occurs in FLEX because the temporally specific inhibition of DA neurons at the time of the US is mediated by the (still intact) cue-reward delay representation in PFC.

Discussion

TD has established itself as one of the most successful and descriptive models of neural function to date. However, two key factors make it sensible to consider alternatives to TD as models for how midbrain dopamine neurons could learn RPE. First, attempts for biologically plausible implementations of TD traditionally assume an absolute and fixed basis for time, presenting challenges for potential biophysical implementations. Second, different conditioning paradigms have revealed dopamine dynamics that are incompatible with classical TD conceptions of RPE[5,44,44,45].

We attempt to address these issues by introducing a novel and biophysically inspired neural architecture, which learns reward prediction errors via biophysically plausible learning rules based on the Hebbian activation of synaptic eligibility traces. We

demonstrate that our single model, FLEX, can replicate DA neuron dynamics under various conditioning paradigms with single or multiple cues. In doing so, FLEX produces neural firing patterns which approximate RPE at the endpoints of learning. However, FLEX diverges from TD-like models since DA neuron firing and RPE are not necessarily synonymous during learning. In FLEX, evolution of DA neuron firing and release is facilitated by feature specific, temporally flexible cue-reward associations housed in upstream cortical areas. We propose that the dynamics of learning in the model can account for apparent discrepancies between various experimental findings in dopaminergic midbrain neurons, while also providing novel predictions which are distinct from those of TD.

In accord with recent experimental evidence from the dopaminergic system[46–48], conditioning in FLEX follows a "model-based" rather than strictly "model-free" learning paradigm. Model-free conditioning posits that a single scalar value (e.g. RPE) is the exclusive engine for learning, forgoing any information about specific contextual features and relations of the cues, environment, etc. In contrast, model-based learning consists of an agent creating an internal model of its environment and using it, along with RPE-like signals, to perform learning. In FLEX, the interplay of the cue-reward associations in PFC and dopaminergic reinforcement D(t) from VTA constitute such a "model". Unlike model-free methods, dopamine transients in FLEX are simply reinforcement signals, and only *resemble* predictions, prediction errors, or associations. One model-based consequence of FLEX is the prediction that negative RPEs will still arise upon reward omission following overtraining. Additionally, previous work has shown that the framework of CNAs is able to learn and recall both Markovian and non-Markovian sequences of inputs[30], so in theory the PFC module within FLEX could be

generalized to learn other aspects of the environment such as sequences or relational assemblies of cues as a part of its "model".

FLEX's ability to approximate "distributional" reinforcement learning hinges on the distributional projections of reward magnitude onto FLEX's population of non-linear, spiking neurons. Another recent model considered the consequences of distributional reward *delays* to DA neurons and found that a TD-like framework containing these delays could explain why striatal targets of DA neuron axons have been observed to receive spatiotemporally patterned "waves" of dopamine[49]. In FLEX, implementation of distributional reward delays alone is not sufficient to facilitate wave-like behavior in downstream targets. It remains to be seen in future work if FLEX can incorporate more complex features of observed dopamine dynamics, including the emergence of striatal dopamine waves[49] and dopamine "ramps"[50–53].

The feasibility of FLEX can be tested by experimentally examining some of its key predictions and assumptions. The model predicts that the dynamics of DA neuron firing do not necessarily represent an RPE during learning, but only approximate it on the population level before and after learning. One simple consequence of this prediction is that the CS response potentiates before the US response depresses. In the context of conditioning with sequential cues, this separation of RPE and DA neuronal firing leads to the prediction that the seemingly contradictory results of Shultz et al.[36] and Pan et al.[5] can be reconciled by considering the stage of training. Finally, FLEX predicts that omitting an expected reward after overtraining will still result in a negative RPE.

It is also important to consider possible physical locations of our architecture. While DA neurons which display phasic RPEs are seen in (including but not exclusively) the

lateral VTA[4,7], the potential location of the CNA is a more complicated question. The exact roles of mPFC and OFC in modulating dopamine signaling (and thereby reward prediction error) are a subject of much debate[48,54–59], however, the following key results suggest OFC would be more likely to house something akin to the CNA. First, a recent study[26] used calcium imaging to track neurons in the OFC using during and after trace conditioning tasks. Unsupervised clustering of these neurons revealed a class (cluster 1) which strongly projects to VTA and has many features akin to those of our Timers. Neurons of this cluster were found to a) be feature specific, b) encode cue-reward delays via persistent activity, and c) maintain their persistent activity even after extinction of the cue-reward association. Other studies have reported classes of neurons in the OFC with a similar functionality[60]. Second, experiments which recorded from DA neurons during a delay discounting task found that inactivation of OFC differentially affected CS-evoked and US-evoked dopamine, altering reward prediction errors, while inactivation of mPFC did not seem to influence RPEs[56]. It is also a possibility that the CNA could be distributed (and/or redundant) across brain regions, whether at the level of populations or even individual cells. However, we posit that the existing experimental evidence makes it sensible to first consider CNAs (and particularly, the feature-specific Timers) as potentially residing in frontal cortical areas, likely in OFC.

One final testable assumption of FLEX is that cue-reward associations in OFC are crucial for *forming temporally specific* expectations of reward. While several studies have investigated the effects of inactivating OFC[26,56,59] or VTA-targeting OFC neurons[26] in the context of dopamine-based conditioning, the types of inactivation, the times of inactivation, and the measures used to quantify the effects have varied greatly, leaving it unclear whether these results support or challenge FLEX's specific predictions. While

we believe FLEX is consistent with this existing literature, more targeted experiments can distinguish FLEX from other models and shed light on the role of OFC in trace conditioning. Such an experiment would involve blocking LTP (such as via inhibitors of CaMKII) in OFC during learning of a temporally constrained behavioral task – FLEX would predict that this would affect the temporally specific suppression of expected reward, but not necessarily the acquisition of a dopamine response at the time of the reward-predictive cue. As a result, mean behavioral responses to the cue itself may be unaffected, but FLEX would predict that the delayed, reward-specific timing of these behaviors would be disrupted.

In this paper we propose a novel model, called FLEX, which provides a potential mechanistic and biophysically plausible implementation of RPE-like learning in midbrain dopaminergic neurons. FLEX also generates testable predictions that can distinguish it from classical methods of TD learning. FLEX is composed of two components, an analog of the VTA, and a cortical component, both of which are necessary for successful learning. Within FLEX, the cortical component learns a 'model' of spatio-temporal statistics of the environment, making FLEX not only mechanistically but also conceptually distinct from previous models, as it is no longer a purely 'model-free' theory of reinforcement learning.

Methods

Network Dynamics

The membrane dynamics for each neuron i are described by the following equations:

Equation 2 - Membrane Dynamics

$$C\frac{dv_i}{dt} = g_L(E_L - v_i) + g_{E,i}(E_E - v_i) + g_{I,i}(E_I - v_i) + \sigma$$

Equation 3 - Synaptic Activation Dynamics

$$\frac{ds_i}{dt} = -\frac{s_i}{\tau_s} + \rho(1 - s_i)\sum_k \delta(t - t_k^i)$$

The membrane potential and synaptic activation of neuron i are notated v_i and s_i. g refers to the conductance, and E to the reversal potentials indicated by the appropriate subscript, where leak, excitatory, and inhibitory are indicated by subscripts L, E, and I, respectively. σ is a mean zero noise term. The neuron spikes upon it crossing its membrane threshold potential v_{th}, after which it enters a refractory period t_{ref}. The synaptic activation s_i is updated by an amount $\rho(1 - s_i)$, at each time (t_k^i) the neuron spikes, and decays exponentially when there is no spike.

The conductance g is the product of the incoming synaptic weights and their respective presynaptic neurons:

Equation 4 - Conductance

$$g_{\alpha,i} = \sum_j W_{ij}^{\alpha} s_j$$

W_{ij}^{α} are the connection strengths from neuron j to neuron i, where the superscript α can either indicate E (excitatory) or I (inhibitory). A firing rate estimate for each neuron r_i is calculated as an exponential filter of the spikes, with a time constant τ_r.

Equation 5 - Firing Rate

$$\tau_r\frac{dr_i}{dt} = -r_i + \sum_k \delta(t - t_k^i)$$

Dopaminergic Two-Trace Learning (dTTL)

Rather than using temporal difference learning, FLEX uses a previously established learning rule based on competitive eligibility traces, known as "two-trace learning" or TTL[24,30,31]. However, we replace the general reinforcement signal R(t) of previous implementations by a dopaminergic reinforcement D(t). We repeat here the description of D(t) from the main text:

Equation 6 – Dopamine reinforcement

$$D(t) = \begin{cases} r_{DA}(t) - (r_0 - \theta), & r_{DA}(t) \leq r_0 - \theta \\ 0, & r_0 - \theta < r_{DA}(t) < r_0 + \theta \\ r_{DA}(t) - (r_0 + \theta), & r_{DA}(t) \geq r_0 + \theta \end{cases}$$

Note that the dopaminergic reinforcement can be both positive and negative, even though actual DA neuron firing (and the subsequent release of dopamine neurotransmitter) can itself only be positive. The bipolar nature of D(t) implicitly assumes that background tonic levels of dopamine do not modify weights, and that changes in synaptic efficacies are a result of a departure (positive or negative) from this background level. The neutral region around r_0 provides robustness to small fluctuations in firing rates which are inherent in spiking networks.

The eligibility traces, which are synapse specific, act as long-lasting markers of Hebbian activity. The two traces are separated into LTP- and LTD-associated varieties via distinct dynamics, which are described in the equations below.

Equation 7 - LTP Trace Dynamics

$$\tau^p \frac{dT_{ij}^p}{dt} = -T_{ij}^p + \eta^p H_{ij} \frac{\left(T_{max}^p - T_{ij}^p\right)}{T_{max}^p}$$

Equation 8 - LTD Trace Dynamics

$$\tau^{d}\frac{dT_{ij}^{d}}{dt} = -T_{ij}^{d} + \eta^{d}H_{ij}\frac{\left(T_{max}^{d} - T_{ij}^{d}\right)}{T_{max}^{d}}$$

Here, T_{ij}^{a} (where $a \in (p, d)$) is the LTP (p superscript) or LTD (d superscript) eligibility trace located at the synapse between the j-th presynaptic cell and the i-th postsynaptic cell. The Hebbian activity, H_{ij}, is a simple multiplication $r_i \cdot r_j$ for application of this rule in VTA, where r_j and r_i are the time-averaged firing rates at the pre- and post-synaptic cells. Experimentally, the "Hebbian" terms (H_{ij}) which impact LTP and LTD trace generation are complex[24], but in VTA we approximate with the simple multiplication $r_i \cdot r_j$. For synapses in PFC, we make the alteration that $H_{ij} = \frac{r_i \cdot r_j}{1+\alpha D(t)}$, acting to restrict PFC trace generation for large positive RPEs. The alteration of H_{ij} by large positive RPEs is inspired by recent experimental work showing that large positive RPEs act as event boundaries and disrupt across-boundary (but not within-boundary) associations and timing[34]. Functionally, this altered H_{ij} biases Timers in PFC towards encoding a single delay period (cue to cue or cue to reward) and disrupts their ability to encode across-boundary delays.

The LTP and LTD traces activate (via activation constant η^{a}), saturate (at a level T_{max}^{a}) and decay (with time constant τ^{a}) at different rates. D(t) binds to these eligibility traces, converting them into changes in synaptic weights. This conversion into synaptic changes is "competitive", being determined by the difference in the product:

Equation 9 - Weight Update

$$\frac{dW_{ij}}{dt} = \eta D(t)\left(T_{ij}^{p}(t) - T_{ij}^{d}(t)\right)$$

where η is the learning rate.

The interaction of the eligibility traces T_{ij}^{p} and T_{ij}^{d} with dopaminergic reinforcement D(t), as well as the fixed points of the learning rule, are examined further in **Supplemental Methods**.

The above eligibility trace learning rules, with the inclusion of dopamine, are referred to as dopaminergic "two-trace" learning or dTTL. This rule is pertinent not only because it can solve the temporal credit assignment problem, allowing the network to associate events distal in time, but also because dTTL is supported by recent experiments which have found eligibility traces for in multiple brain regions[24,61–63]. Notably, in such experiments, the eligibility traces in prefrontal cortex were found to convert into positive synaptic changes via delayed application of dopamine[24], which is the main assumption behind dTTL.

For simplicity and to reduce computational time, in the simulations shown, M→GABA connections are learned via a simple dopamine modulated Hebbian rule, $\frac{dW_{ij}}{dt} = \eta D(t) r_i r_j$. Since these connections are responsible for inhibiting the DA neurons at the time of reward, this learning rule imposes its own fixed point by suppressing D(t) down to 0. For any appropriate selection of feed-forward learning parameters in dTTL (**Equation 9**), the fixed-point $D(t) = 0$ is reached well before the fixed point $T_{ij}^{p}(t) = T_{ij}^{d}(t)$. This is because, by construction, the function of M→GABA learning is to suppress D(t) down to zero. Therefore, the fixed point $T_{ij}^{p}(t) = T_{ij}^{d}(t)$ needs to be placed (via choosing trace parameters) beyond the fixed-point $D(t) = 0$. Functionally, then, both rules act to potentiate M→GABA connections monotonically until D(t) = 0. As a result, the dopamine modulated Hebbian rule is in practice equivalent to dTTL in this case.

Network Architecture

Our network architecture consists of two regions, VTA and PFC, each of which consist of subpopulations of leaky-integrate-and-fire neurons. Both fixed and learned connections exist between certain populations to facilitate the functionality of our model.

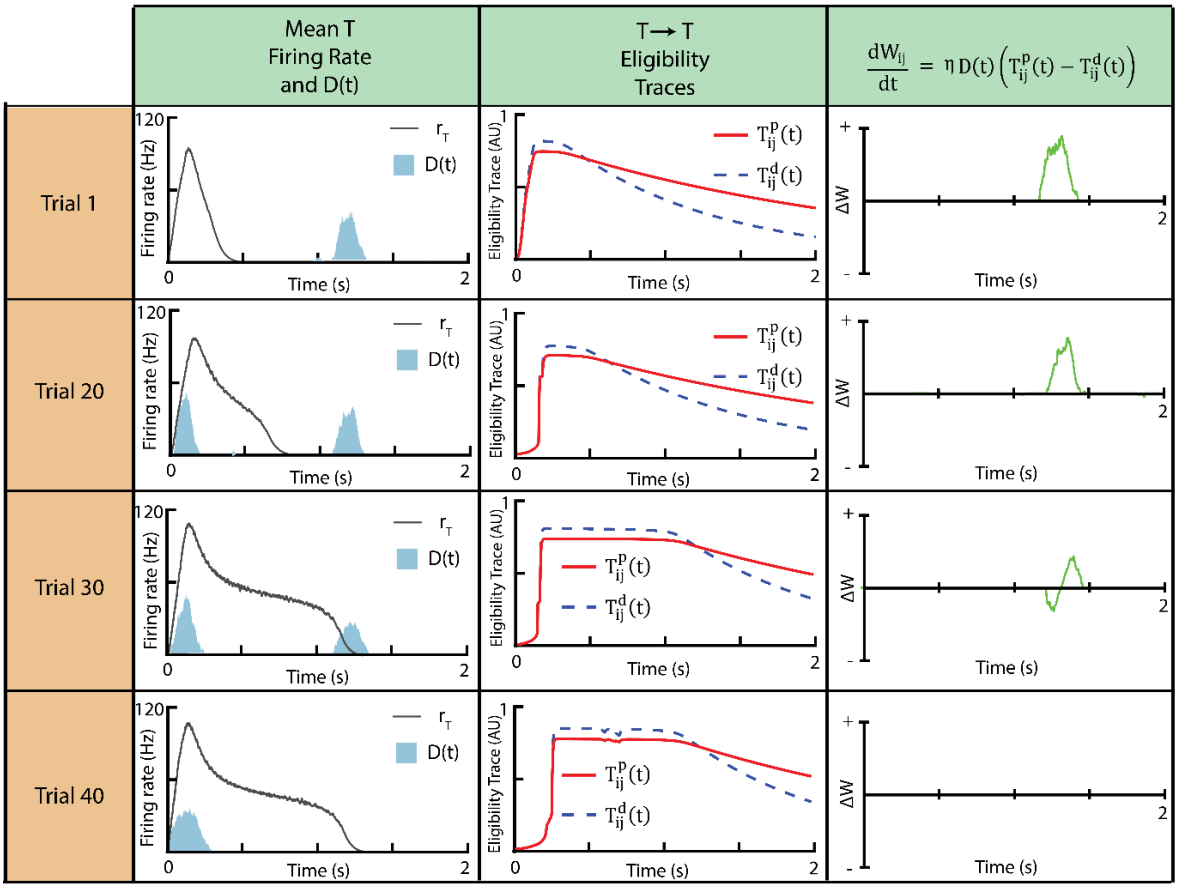
To model VTA, we include 100 dopaminergic and 100 GABAergic neurons, both of which receive tonic noisy input to establish baseline firing rates of ~5Hz. Naturally appetitive stimuli, such as food or water, are assumed to have fixed connections to DA neurons via the gustatory system. Dopamine release is determined by DA neuron firing above or below a threshold θ, and dopamine release acts as reinforcement for all learned connections in the model.

Our model of PFC is comprised of different feature-specific 'columns'. Within each column there is a CNA microcircuit, with each subpopulation (Timers, Inhibitory, Messengers) consisting of 100 LIF neurons. Previous work has shown that these subpopulations can emerge from randomly distributed connections[29], and further that a single mean field neuron can well approximate the activity of each of these subpopulations of spiking neurons[64].

Data Availability

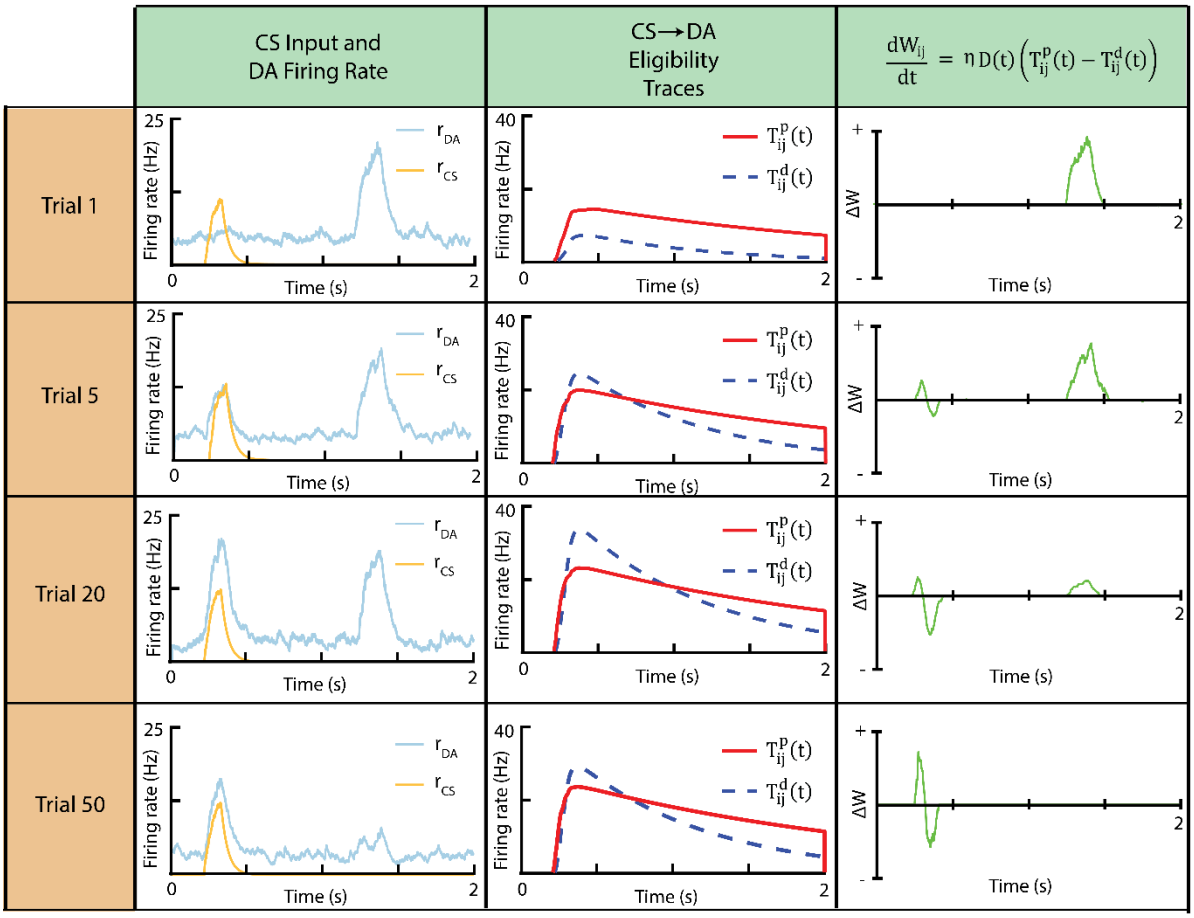
The code and parameters used to simulate FLEX are available upon request and will be made publicly accessible via modelDB upon peer-reviewed publication.

Supplemental Figures



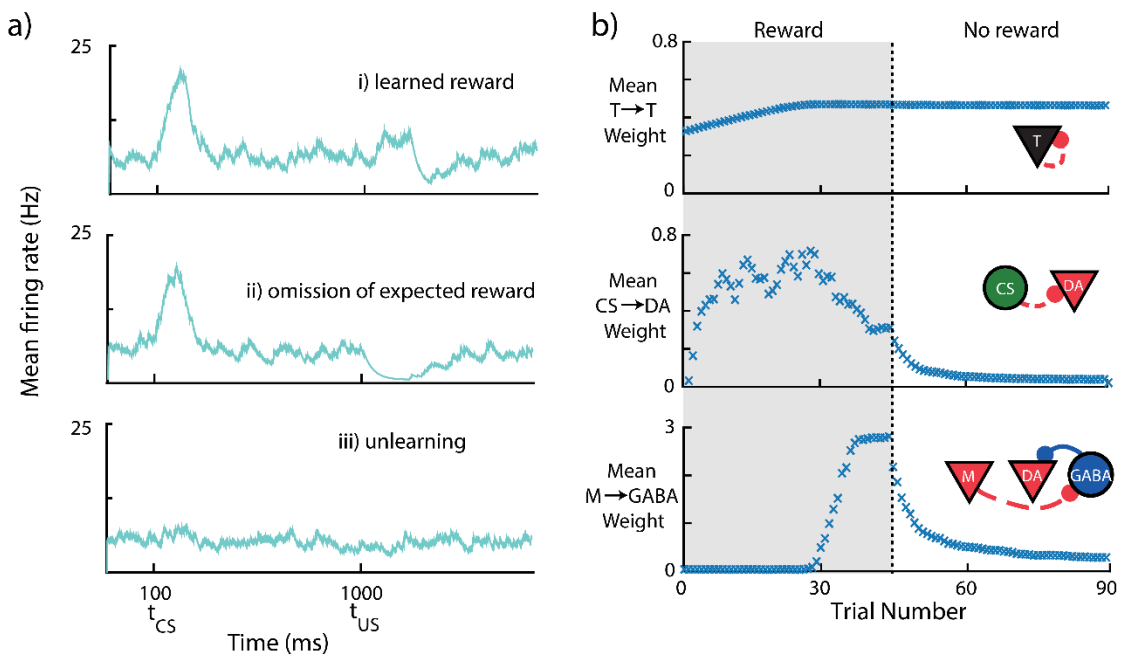
Supplemental Figure 1 – Two-Trace Learning Flexibly Encodes Cue-Reward Delay in Recurrent Connections of Timer Neurons

Demonstration of the dynamics of two-trace learning for recurrent connections. Rows consist of different trials. Left column, mean firing rate of Timer neurons (black) and dopamine reinforcement D(t) (light blue) for a given trial. Middle column, LTP (red) and LTD (blue) associated eligibility traces triggered by the Hebbian overlap $\frac{r_i \cdot r_j}{1+\alpha D(t)}$. Right column, $\frac{dW}{dt}$, calculated at a given time as the difference between the two traces ($T^p_{ij} - T^d_{ij}$) multiplied by the dopamine reinforcement, D(t). For all trials, the increase in recurrent weights is mediated by dopamine release at t_{US}. Since dopamine acts to suppress trace generation in PFC, the CS-evoked dopamine response (trial 20) has little effect on the recurrent learning. The weights increase until ΔW is zero (trial 30) and remain at their fixed point after US-evoked dopamine has been suppressed (trial 40).



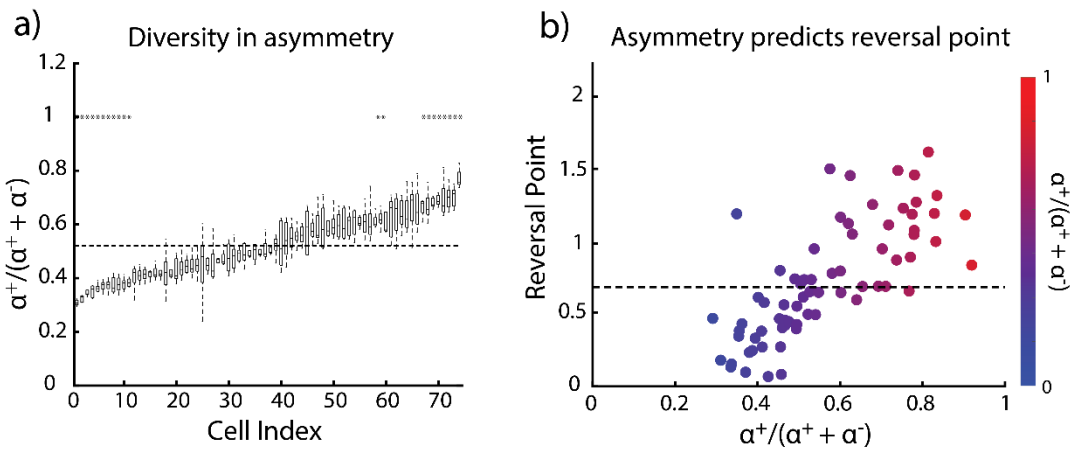
Supplemental Figure 2 – Two-Trace Learning Encodes Reward Predictive Cues via Feed-Forward Connections to DA Neurons

Demonstration of the dynamics of two-trace learning for feed-forward connections. Rows consist of different trials. Left column, input from external conditioned stimulus (orange) and mean firing rate of dopamine neurons (light blue) for a given trial. Middle column, LTP (red) and LTD (blue) associated eligibility traces triggered by the Hebbian overlap $r_{CS} * r_{DA}$. Right column, $\frac{dW}{dt}$, calculated at a given time as the difference between the two traces ($T^p_{ij} - T^d_{ij}$) multiplied by the dopamine reinforcement, D(t). Initially (trial 1), increase in feed forward weights is mediated by dopamine release at t_{US}. However, as a CS-evoked dopamine response begins to develop (trial 5), the weights are bounded at a fixed point, constrained by the relative positive (from the US) and negative (from the CS) contributions (trial 20). After the expected reward at t_{US} has been depressed, the DA neuron firing at t_{CS} decreases slightly until $\frac{dW}{dt}$ reaches a fixed point maintained by the CS dopamine alone.



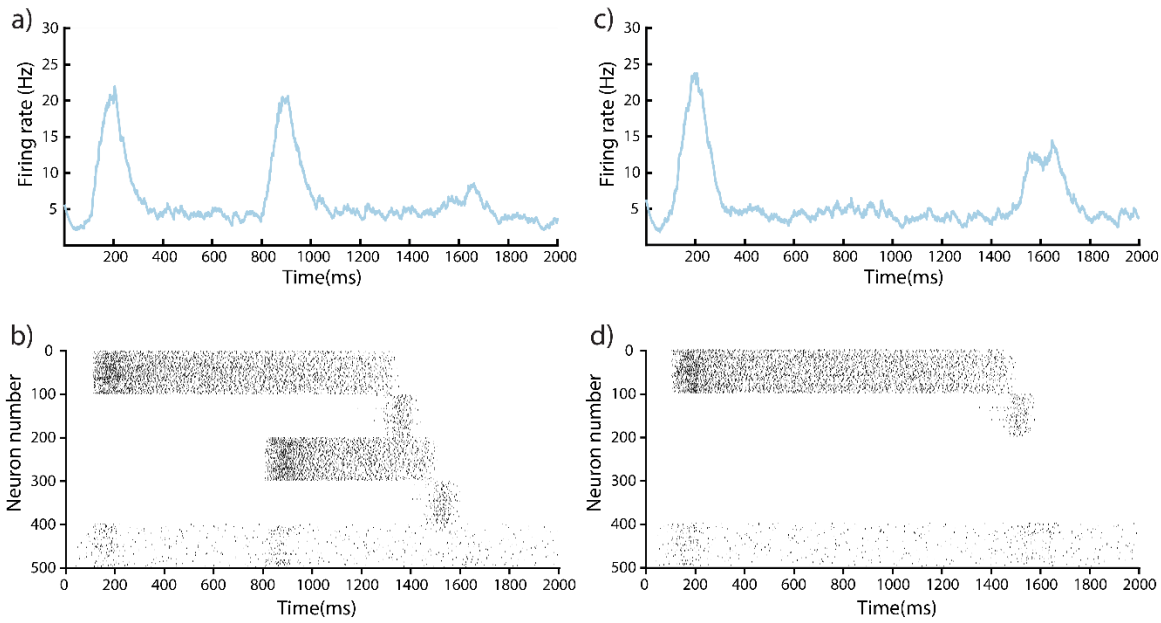
Supplemental Figure 3 – Consistent Unpairing Leads to Negative RPEs and Unlearning

a) Mean firing rate over all VTA DA neurons for a given trial after learning reward (i), upon omission of the learned reward (ii), and finally, following unlearning (iii). The omission of the expected reward produces a characteristic dopamine "dip", which in turn acts as a negative RPE and facilitates unlearning of the cue-reward association. **b)** Mean Timer→Timer (top), CS→DA (middle), and M→GABA (bottom) synaptic weights over the course of pairing and unpairing. The cue is presented for all trials, while the reward is only presented for the first 45 trials (shaded grey). For trials ~30-45, M→GABA weights increase, suppressing the amount of DA firing at the time of the reward. In response, the fixed point of CS→DA weights (which is determined, in part, by the dynamics of the dopamine reinforcement D(t)) decreases before stabilizing in trials ~40-45. Following omission, both CS→DA and M→GABA weights decrease to zero, but T→T weights are maintained, since the fixed point for these weights (see Supplemental Figure 1) is agnostic to changes in to D(t).



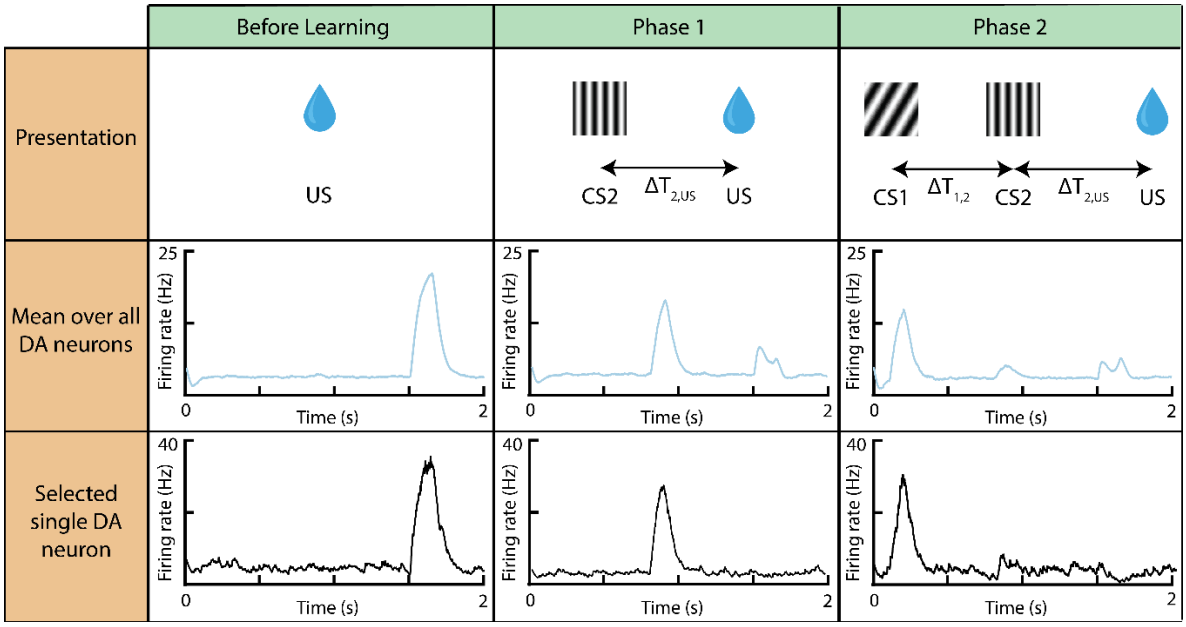
Supplemental Figure 4 – Asymmetric Responses to Positive and Negative RPEs Predict Value Reversal Point in Model DA Neurons

a) Distributions of "asymmetry" measure $\frac{\alpha^+}{\alpha^+ + \alpha^-}$ for each VTA DA neuron with an RPE response at t_{US}, calculated from 5 rounds of testing on variable reward paradigm. Trial to trial variability in each cell's asymmetry arises from the stochastic nature of the spiking network. The cell by cell heterogeneity in asymmetry is a robust phenomenon (one-way ANOVA; F(60,244) = 30.97, p = 2.25×10^{-85} vs the null hypothesis that mean asymmetry was not different across cells) **b)** Asymmetry was also positively correlated (linear regression, p = 5.87×10^{-9}) with the "reversal point" of the neuron, the value of reward for which that neuron reported zero RPE.



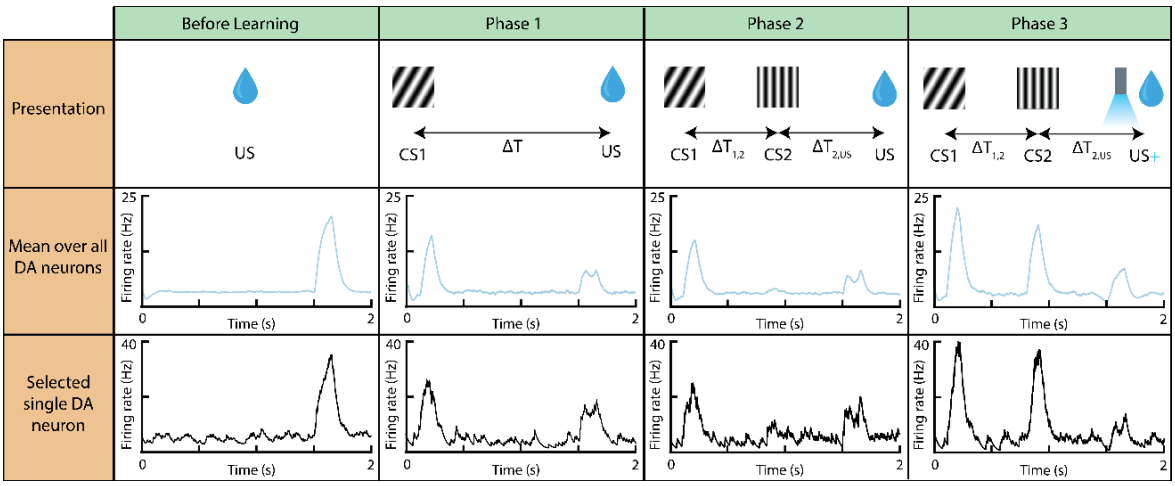
Supplemental Figure 5 – Both Cues Contribute to Reward Prediction Following Initial Training

a) Mean firing rate of all VTA neurons for presentation of CS1+CS2+US following sequential conditioning. **b)** Spike raster of Timer neurons (1-100, 201-300), Messenger neurons (101-200, 301-400), and VTA DA neurons (401-500) for the same trial. **c)** Mean firing rate of all VTA neurons for presentation of CS1+US, omitting CS2 following simultaneous conditioning. Notably, CS1 only partially predicts (and in turn partially inhibits) the dopamine response at t_{US}. d**)** Spike raster of Timer neurons (1-100, 201-300), Messenger neurons (101-200, 301-400), and VTA DA neurons (401-500) for the same trial.



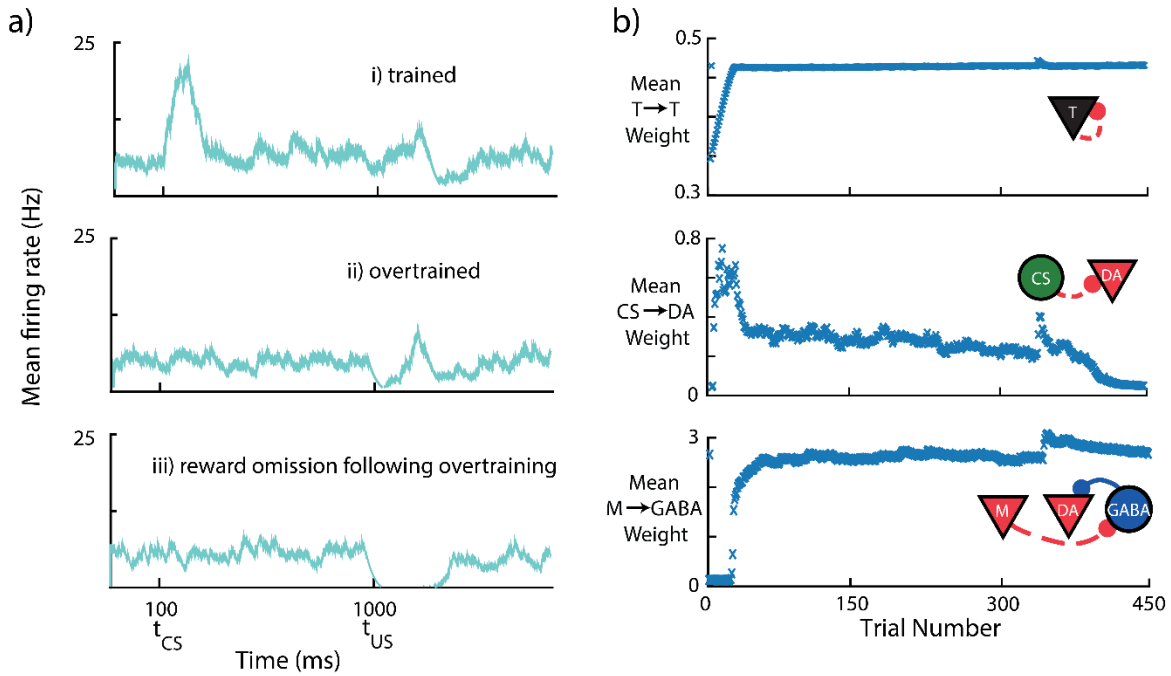
Supplemental Figure 6 – Stepwise Training Protocol Surpasses Intermediate Learning Stages and only Demonstrates Serial Transfer of Activation

Each column marks a different 25 trial phase of conditioning in a stepwise sequential paradigm. Top row, visual representation of presentation protocol in given column. Middle row, mean over all DA neurons and trials for given presentation protocol. Bottom row, a selected single unit response averaged over trials of the given presentation protocol. Before learning, unpredicted rewards trigger dopamine release at t_{US}. In phase 1, CS2→US is learned, and DA neurons develop a response to CS2 and have suppressed their response to the US. In phase 2, CS1→CS2→US is presented, but since the dopamine response at t_{US} has already been "transferred" to t_{CS2}, phase 2 evolves equivalently to phase 1, but this time CS1 is acting as the "CS" and CS2 as the "US". The result is second-order conditioning, where a cue becomes predictive of another reward-predictive cue. In this case in our model, only CS2 is directly predictive of reward, and CS1 is only predictive of CS2.



Supplemental Figure 7 – Expected Rewards Block Learning of New Cue-Reward Associations Unless Reward Magnitude is Increased

Each column marks a different phase of conditioning in a blocking/unblocking paradigm. Results shown are averaged over 25 trials. Top row, visual representation of protocol in given column. Middle row, mean over all DA neurons and trials for given presentation protocol. Bottom row, a selected single unit response averaged over trials of the given presentation protocol. Before learning, unpredicted rewards trigger dopamine release at t_{US}. In phase 1, CS1→US is learned, and DA neurons develop a response to CS1 and have suppressed their response to the US. In phase 2, CS1→CS2→US is presented, but a dopamine response fails to develop to CS2, as the US is already fully predicted by CS1 (and therefore CS2 is "blocked"). Phase 3 results in "unblocking" via increasing dopamine release at t_{US} – doing so recovers the CS2-evoked response and results in both stimuli becoming reward-predictive.



Supplemental Figure 8 – Omission of Expected Reward after Overtraining Results in Negative RPE

a) Mean firing rate over all VTA DA neurons for a given trial after successful training (i), following overtraining (ii), and finally, upon reward omission after overtraining (iii). Despite the network reaching an overtrained state, the omission of the expected reward still produces a negative RPE "dip" in the dopamine firing. This occurs because the feature-specific cue-reward delay in PFC, which is responsible for temporally specific suppression of the reward, is maintained despite overtraining. **b)** Mean Timer→Timer (top), CS→DA (middle), and M→GABA (bottom) synaptic weights over the course of training. At around trial 340, spurious noise happens to accumulate in such a manner to kick the network out of its attractive basin. In response to this noise, M→GABA weights transiently increase beyond their fixed-point vales, which in turn increases inhibition to the DA neurons at the time of the US. The resulting E/I imbalance at the time of the US results in negative RPEs on average, dragging down the CS→DA weights down towards zero and taking the network into an overtrained state.

Supplemental Methods

The two-trace learning rule we utilize for our model is described in further detail in previous work[24,30–32]. However, we will attempt to clarify how it functions below.

As a first approximation, the traces from the two-trace learning rule we utilize effectively act such that when they interact with dopamine above baseline, the learning rule will favor potentiation, and when they encounter dopamine below baseline, the learning rule will favor depression. Formally, the rule has fixed points which depend on both the dynamics of the traces and the dopamine release D(t):

Equation 10 – Fixed Point

$$\int_{0}^{T_{trial}} dt\, \mathrm{D(t)} \left(\mathrm{T_{ij}^{p}(t)} - \mathrm{T_{ij}^{d}(t)} \right) = 0$$

A trivial fixed point exists when D(t) = 0 for all t. Another simple fixed point exists in the limit that $\mathrm{D(t)} = \delta(t_R - t)$, where t_R is the time of reward, as **Equation 10** then reduces to $\mathrm{T_{ij}^{p}}(t_R) = \mathrm{T_{ij}^{d}}(t_R)$. In this case, the weights have reached their fixed point when the traces cross at the time of reward. In practice, the true fixed points of the model are a combination of these two factors (suppression of dopamine and crossing dynamics of the traces). In reality, D(t) is not a delta function (and may have multiple peaks during the trial), so to truly calculate the fixed points, one must use **Equation 10** as a whole. However, the delta approximation used above gives a functional intuition for the dynamics of learning in the model.

Supplemental Figure 1 and **Supplemental Figure 2** demonstrate examples of fixed

points for both recurrent and feed-forward learning, respectively. Note that in these examples the two "bumps" of excess dopamine (CS-evoked and US-evoked) are the only instances of non-zero D(t). As such, we can take the integral in **Equation 10** and split it into two parts:

Equation 11 – Two-part Fixed Point

$$\int_{t_{CS,start}}^{t_{CS,end}} dt\, \mathrm{D(t)} \left(\mathrm{T_{ij}^{p}(t)} - \mathrm{T_{ij}^{d}(t)}\right) + \int_{t_{US,start}}^{t_{US,end}} dt\, \mathrm{D(t)} \left(\mathrm{T_{ij}^{p}(t)} - \mathrm{T_{ij}^{d}(t)}\right) = 0$$

For recurrent learning, the dynamics evolve as follows. At the beginning, only the integral over Δt_{US} exists, as D(t) is initially zero over Δt_{CS} (Trial 1 in **Supplemental Figure 1**). As a result, the learning rule evolves to approach the fixed point mediated by $\left(\mathrm{T_{ij}^{p}(t)} - \mathrm{T_{ij}^{d}(t)}\right)$ (Trial 20 in **Supplemental Figure 1**). After the recurrent weights have reached this fixed point and the Timer neurons encode the cue-reward delay (Trial 30 in **Supplemental Figure 1**), M→GABA learning acts to suppress D(t) down to zero as well (Trial 40 in **Supplemental Figure 1**). Note again that we make the assumption that trace generation in PFC is inhibited during large positive RPEs. This acts to encourage the Timers to encode a single "duration" (whether cue-cue or cue-reward). In line with our assumption, experimental evidence has shown these large positive RPEs act as event boundaries and disrupt across-boundary (but not within-boundary) reports of timing[34].

For feed-forward learning, the weights initially evolve identically to the recurrent weights (Trial 1 in **Supplemental Figure 1**). Again, only the integral over Δt_{US} exists, so the feed-forward weights evolve according to $\left(\mathrm{T_{ij}^{p}(t)} - \mathrm{T_{ij}^{d}(t)}\right)$. However, soon the potentiation of these feed-forward CS→DA weights themselves cause release of CS-evoked dopamine, and therefore we must consider both integrals to explain the learning

dynamics (Trial 5 in **Supplemental Figure 1**). This stage of learning is harder to intuit, but an intermediate fixed point is reached when the positive ΔW produced by the traces' overlap with US-evoked dopamine is equal and opposite to the negative ΔW produced by the traces' overlap with CS-evoked dopamine (Trial 20 in **Supplemental Figure 1**). Finally, after US-evoked dopamine has been suppressed to baseline, the feed-forward weights reach a final fixed point where both positive and negative contributions to ΔW over the course of the CS offset each other (Trial 50 in **Supplemental Figure 1**).

The measure of "area under receiver operating characteristic" (auROC) is used throughout this paper, for the purpose of making direct comparison to a slew of calcium imaging results that use auROC as a measure of statistical significance. Following the methods of Cohen et al., time is tiled into 50ms bins. For a single neuron, within each 50ms bin, the distribution of spike counts for 25 trials of baseline spontaneous firing (no external stimuli) is compared to the distribution of spike counts during the same time bin for 25 trials of the learning phase in question. For example, in **Figure 3a** left, the baseline distributions of spike counts are compared to the distributions of spike counts when US only is presented. ROC is calculated for each bin by sliding the criteria from zero to the max spike count within the bin, and then plotting P(active>criteria) versus P(baseline>criteria). The area under this curve is then a measure of discriminability between the two distributions, with an auROC of 1 demonstrating a discriminable increase in spikes compared to baseline, 0 demonstrating a discriminable decrease of spikes compared to baseline, and .5 demonstrating an inability to discriminate between the two distributions.

Acknowledgements

This work was supported by Biotechnology and Biological Sciences Research Council (BBSRC) Grants BB/N013956/1 and BB/N019008/1, Wellcome Trust 200790/Z/16/Z, Simons Foundation 564408, Engineering and Physical Sciences Research Council (EPSRC) EP/R035806/1, National Institute of Biomedical Imaging and Bioengineering 1R01EB022891-01 and Office of Naval Research N00014-16-R-BA01.

Competing Interests

The authors declare no competing interests.

References

1. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593–1599 (1997).
2. Schultz, W., Apicella, P. & Ljungberg, T. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neurosci.* **13**, 900–913 (1993).
3. Ljungberg, T., Apicella, P. & Schultz, W. Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.* **67**, 145–163 (1992).
4. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
5. Pan, W.-X. Dopamine Cells Respond to Predicted Events during Classical Conditioning: Evidence for Eligibility Traces in the Reward-Learning Network. *J. Neurosci.* **25**, 6235–6242 (2005).
6. Dabney, W. *et al.* A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
7. Eshel, N. *et al.* Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* **525**, 243–246 (2015).
8. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* **19**, 479–486 (2016).
9. Sutton, R. S. & Barto, A. G. *Reinforcement Learning, second edition: An Introduction*. (MIT Press, 2018).
10. Montague, P., Dayan, P. & Sejnowski, T. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
11. Schultz, W. Behavioral Theories and the Neurophysiology of Reward. *Annu. Rev. Psychol.* **57**, 87–115 (2006).

12. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Stimulus Representation and the Timing of Reward-Prediction Errors in Models of the Dopamine System. *Neural Comput.* **20**, 3034–3054 (2008).
13. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Evaluating the TD model of classical conditioning. *Learn. Behav.* **40**, 305–319 (2012).
14. Mello, G. B. M., Soares, S. & Paton, J. J. A Scalable Population Code for Time in the Striatum. *Curr. Biol.* **25**, 1113–1122 (2015).
15. Menegas, W., Babayan, B. M., Uchida, N. & Watabe-Uchida, M. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife* **6**, e21886 (2017).
16. Flagel, S. B. *et al.* A selective role for dopamine in stimulus–reward learning. *Nature* **469**, 53–57 (2011).
17. Roesch, M. R., Calu, D. J. & Schoenbaum, G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci.* **10**, 1615–1624 (2007).
18. Hollerman, J. R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1**, 304–309 (1998).
19. Amo, R., Yamanaka, A., Tanaka, K. F., Uchida, N. & Watabe-Uchida, M. A gradual backward shift of dopamine responses during associative learning. *bioRxiv* 2020.10.04.325324 (2020) doi:10.1101/2020.10.04.325324.
20. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
21. Clark, J. J., Collins, A. L., Sanford, C. A. & Phillips, P. E. M. Dopamine Encoding of Pavlovian Incentive Stimuli Diminishes with Extended Training. *J. Neurosci.* **33**, 3526–3532 (2013).
22. Chubykin, A. A., Roach, E. B., Bear, M. F. & Shuler, M. G. H. A Cholinergic Mechanism for Reward Timing within Primary Visual Cortex. *Neuron* **77**, 723–735 (2013).
23. Shuler, M. G. Reward Timing in the Primary Visual Cortex. *Science* **311**, 1606–1609 (2006).

24. He, K. *et al.* Distinct Eligibility Traces for LTP and LTD in Cortical Synapses. *Neuron* **88**, 528–538 (2015).

25. Kim, E., Bari, B. A. & Cohen, J. Y. Subthreshold basis for reward-predictive persistent activity in mouse prefrontal cortex. *Cell Rep.* **35**, 109082 (2021).

26. Namboodiri, V. M. K. *et al.* Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. *Nat. Neurosci.* **22**, 1110–1121 (2019).

27. Bellemare, M. G., Dabney, W. & Munos, R. A Distributional Perspective on Reinforcement Learning. in *Proceedings of the 34th International Conference on Machine Learning* 449–458 (PMLR, 2017).

28. Rivest, F., Kalaska, J. F. & Bengio, Y. Alternative time representation in dopamine models. *J. Comput. Neurosci.* **28**, 107–130 (2010).

29. Huertas, M. A., Hussain Shuler, M. G. & Shouval, H. Z. A Simple Network Architecture Accounts for Diverse Reward Time Responses in Primary Visual Cortex. *J. Neurosci.* **35**, 12659–12672 (2015).

30. Cone, I. & Shouval, H. Z. Learning precise spatiotemporal sequences via biophysically realistic learning rules in a modular, spiking network. *eLife* **10**, e63751 (2021).

31. Huertas, M. A., Schwettmann, S. E. & Shouval, H. Z. The Role of Multiple Neuromodulators in Reinforcement Learning That Is Based on Competition between Eligibility Traces. *Front. Synaptic Neurosci.* **8**, (2016).

32. Cone, I. & Shouval, H. Z. Behavioral Time Scale Plasticity of Place Fields: Mathematical Analysis. *Front. Comput. Neurosci.* **15**, (2021).

33. Gavornik, J. P., Shuler, M. G. H., Loewenstein, Y., Bear, M. F. & Shouval, H. Z. Learning reward timing in cortex through reward dependent expression of synaptic plasticity. *Proc. Natl. Acad. Sci.* **106**, 6826–6831 (2009).

34. Rouhani, N., Norman, K. A., Niv, Y. & Bornstein, A. M. Reward prediction errors create event boundaries in memory. *Cognition* **203**, 104269 (2020).

35. Coddington, L. T. & Dudman, J. T. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nat. Neurosci.* **21**, 1563–1573 (2018).

36. Schultz, W. Predictive Reward Signal of Dopamine Neurons. *J. Neurophysiol.* **80**, 1–27 (1998).

37. Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y. & Hikosaka, O. Dopamine Neurons Can Represent Context-Dependent Prediction Error. *Neuron* **41**, 269–280 (2004).

38. Tian, J. & Uchida, N. Habenula Lesions Reveal that Multiple Mechanisms Underlie Dopamine Prediction Errors. *Neuron* **87**, 1304–1316 (2015).

39. Kamin, L. J. Predictability, surprise, attention, and conditioning. *B Campbell R M Church Eds Punishm. Aversive Behav.* 279–296 (1969).

40. Mackintosh, N. J. & Turner, C. Blocking as a function of novelty of CS and predictability of UCS. *Q. J. Exp. Psychol.* **23**, 359–366 (1971).

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

42. Holland, P. C. Unblocking in Pavlovian appetitive conditioning. *J. Exp. Psychol. Anim. Behav. Process.* **10**, 476–497 (1984).

43. Steinberg, E. E. *et al.* A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).

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

45. Gardner, M. P. H., Schoenbaum, G. & Gershman, S. J. Rethinking dopamine as generalized prediction error. *Proc. R. Soc. B Biol. Sci.* **285**, 20181645 (2018).

46. Sharpe, M. J. *et al.* Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).

47. Langdon, A. J., Sharpe, M. J., Schoenbaum, G. & Niv, Y. Model-based predictions for dopamine. *Curr. Opin. Neurobiol.* **49**, 1–7 (2018).

48. McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y. & Schoenbaum, G. Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning. *J. Neurosci.* **31**, 2700–2705 (2011).

49. Hamid, A. A., Frank, M. J. & Moore, C. I. Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell* **184**, 2733-2749.e16 (2021).

50. Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E. M. & Graybiel, A. M. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* **500**, 575–579 (2013).

51. Gershman, S. J. Dopamine Ramps Are a Consequence of Reward Prediction Errors. *Neural Comput.* **26**, 467–471 (2014).

52. Mohebi, A. *et al.* Dissociable dopamine dynamics for learning and motivation. *Nature* **570**, 65–70 (2019).

53. Kim, H. R. *et al.* A Unified Framework for Dopamine Signals across Timescales. *Cell* **183**, 1600-1616.e25 (2020).

54. Jo, Y. S., Lee, J. & Mizumori, S. J. Y. Effects of Prefrontal Cortical Inactivation on Neural Activity in the Ventral Tegmental Area. *J. Neurosci.* **33**, 8159–8171 (2013).

55. Stalnaker, T. A., Liu, T.-L., Takahashi, Y. K. & Schoenbaum, G. Orbitofrontal neurons signal reward predictions, not reward prediction errors. *Neurobiol. Learn. Mem.* **153**, 137–143 (2018).

56. Jo, Y. S. & Mizumori, S. J. Y. Prefrontal Regulation of Neuronal Activity in the Ventral Tegmental Area. *Cereb. Cortex* **26**, 4057–4068 (2016).

57. Starkweather, C. K., Gershman, S. J. & Uchida, N. The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. *Neuron* **98**, 616-629.e6 (2018).

58. Lodge, D. J. The Medial Prefrontal and Orbitofrontal Cortices Differentially Regulate Dopamine System Function. *Neuropsychopharmacology* **36**, 1227–1236 (2011).

59. Takahashi, Y. K. *et al.* Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat. Neurosci.* **14**, 1590–1597 (2011).

60. Wang, P. Y. *et al.* Transient and Persistent Representations of Odor Value in Prefrontal Cortex. *Neuron* **108**, 209-224.e6 (2020).

61. Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616–1620 (2014).

62. Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S. & Magee, J. C. Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033–1036 (2017).

63. Brzosko, Z., Zannone, S., Schultz, W., Clopath, C. & Paulsen, O. Sequential neuromodulation of Hebbian plasticity offers mechanism for effective reward-based navigation. *eLife* **6**, e27756 (2017).

64. Gavornik, J. P. & Shouval, H. Z. A network of spiking neurons that can represent interval timing: mean field analysis. *J. Comput. Neurosci.* **30**, 501–513 (2011).