Learning to Express Reward Prediction Error-like Dopaminergic Activity Requires 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

The dominant theoretical framework to account for reinforcement learning in the brain is temporal difference (TD) reinforcement learning. The normative motivation for TD theory is that the brain needs to learn about expected future rewards in order to learn how to maximize these rewards. The TD framework predicts that some neuronal elements should represent the reward prediction error (RPE), which means they signal the difference between the expected future rewards and the actual rewards. What makes the TD learning theory so prominent is that the firing properties of dopaminergic neurons in the ventral tegmental area (VTA) appear similar to those of RPE model-neurons in TD learning. Biologically plausible implementations of TD learning assume a fixed temporal basis for each stimulus that might eventually predict a reward. Here we show on the basis of first principles that such a fixed temporal basis is implausible. We also show that certain predictions of TD learning are inconsistent with experimental data. We propose instead an alternative theoretical framework, coined FLEX (**F**lexibly **L**earned **E**rrors in E**x**pected Reward). In FLEX, feature specific representations of time are learned, allowing for neural representations of stimuli to adjust their timing and relation to rewards in an online manner. As an indirect consequence, dopamine in FLEX resembles, but is not equivalent to RPE. In FLEX dopamine acts as an instructive signal which helps build temporal models of the environment. FLEX is a general theoretical framework that has many possible biophysical implementations. In order to show that FLEX is a feasible approach, we present a specific biophysically plausible model which implements the principles of FLEX. We show that this implementation can account for various reinforcement learning paradigms, and that its results and predictions are consistent with a preponderance of both existing and reanalyzed experimental data.

1. Introduction

The term reinforcement learning is used in machine learning[1], in behavioral science[2] and in neurobiology[3], to denote learning on the basis of rewards or punishment. One type of reinforcement learning is temporal difference (TD) learning, which was designed for machine learning purposes. It has the normative goal of estimating future rewards when rewards can be delayed in time with respect to the actions or cues that engendered these rewards[1].

One of the variables in TD algorithms is called reward prediction error (RPE), and it is simply the difference between the discounted predicted reward at the current state and the discounted predicted reward + actual reward at the next state. The concept of TD learning became prominent in neuroscience once it was demonstrated that firing patterns of dopaminergic neurons in ventral tegmental area (VTA) during reinforcement learning resemble RPE[4,5]

Implementations of TD using computer algorithms are straightforward but are more complex when they are mapped onto plausible neural machinery[6–8]. Current implementations of neural TD assume a set of temporal basis-functions[8,9]. These basis functions are activated by external cues. For this assumption to hold, each possible external cue must activate a separate set of basis-functions, and these basis-functions must tile all possible learnable intervals between stimulus and reward.

In this paper we demonstrate that 1) these assumptions are implausible from a fundamental conceptual level, and 2) predictions of such algorithms are inconsistent with various established experimental results. Instead, we propose that temporal basis functions used by the brain are themselves learned. We call this theoretical framework: **F**lexibly **L**earned **E**rrors in E**x**pected Reward, or FLEX for short. We also propose a biophysically plausible implementation of FLEX, as a proof-of-concept model. We show that key predictions of this model are consistent with actual experimental results but are inconsistent with some key predictions of the TD theory.

2. TD learning with a fixed feature specific temporal-basis

The original TD learning algorithms assumed that agents can be in a set of discrete labeled states (s) which are stored in memory. The goal of TD is to learn a value function such that each state becomes associated with a unique value ($V(s)$) that estimates future discounted rewards. Learning is driven by the difference between value at two subsequent states, and hence such algorithms are called temporal difference algorithms. Mathematically this is captured by the update algorithm: $V(s) \leftarrow V(s) + \alpha\big(r(s') + \gamma V(s') - V(s)\big)$, where s' is the next state and $r(s')$ is the reward in next state, γ is an optional discount factor and α is the learning rate.

The term in the brackets in the right-hand side of the equation is called the RPE. It represents the different between the estimated value at the current state and the estimated discounted value at the next state in addition to the actual reward at the next state. If RPE is zero for every state, the value function no longer changes, and learning reaches a stable state. In experiments that linked RPE to the firing patterns of dopaminergic neurons in VTA, a transient conditioned stimulus (CS) is presented to a naïve animal followed by a delayed reward (also called unconditioned stimulus or US, **Figure 1a**). It was found that VTA neurons initially respond at the time of reward, but once the association between stimulus and reward is learned, neurons stop firing at the time of the reward and start firing at the time of the stimulus (**Figure 1b**). This response pattern is what one would expect from TD learning if VTA neurons represent RPE[5].

Learning algorithms similar to TD have been very successful in machine learning[10,11]. In such implementations the state (s) could, for example, represent the state of the chess board, or the coordinates in space in a navigational task. Each of these states could be associated with a value. The state space in such examples might be very large, but the values of all these

different states could be feasibly stored in a computer's memory. In some cases, a similar formulation seems feasible for a biological system as well. For example, consider a 2-D navigation problem, where each state is a location in space. One could imagine that each state would be represented by the set of hippocampal place cells activated in this location[12], and that another set of neurons would encode the value function, while a third population of neurons (the "RPE") neurons would compare the value at the current and subsequent state. On its face, this seems to be a reasonable assumption.

However, in contrast to cases where a discrete set of states might have straightforward biological implementation, there are many cases in which this machine learning inspired algorithm cannot be implemented simply in biological machinery. For example, in experiments where reward is delivered with a temporal delay with respect to the stimulus offset (**Figure 1**), an additional assumption of a preexisting temporal basis is required[7].

2.1 Why is it implausible to assume a fixed temporal-basis in the brain?

Consider the simple canonical example of **Figure 1**. In the time interval between the stimulus and the reward, the animal does not change its location, nor does its sensory environment change in any predictable manner. The only thing that changes consistently within this interval is time itself. Hence, in order to mark the states between stimulus and reward, the brain must have some internal representation of time, an internal clock which tracks the time since the start of the stimulus. Note however that before the conditioning starts, the animal has no way of knowing that this specific sensory stimulus has unique significance and therefore each specific stimulus must a priori be assigned its own specific temporal representation.

This is the main hurdle of implementing TD in a biophysically realistic manner - figuring out how to represent the temporal basis upon which the association between cue and reward occurs (**Figure 1a**). Previous attempts were based on the assumption that there is a fixed cue-specific temporal basis, an assumption which has previously been termed "a useful fiction"[7]. The specific implementations include the commonly utilized tapped delay lines[5,13,14] (or the so-called complete serial compound), which are neurons triggered by the sensory cue, but that are active only at a specific delay, or alternatively, a set of cue specific neuronal responses which are also delayed but have a broader temporal support which increases with an increasing delay (the so called "microstimuli")[7,9] (**Figure 1c**).

For this class of temporal representations, the delay time between cue and reward is tiled by a chain of neurons, with each neuron representing a cue-specific time (sometimes referred to as a "microstate") (**Figure 1c**). In the simple case of the complete serial compound, the temporal basis is simply a set of neurons that have non-overlapping responses that start responding at the cue-relative times: t_{cue}, $t_{cue} + dt$, $t_{cue} + 2dt$, …, t_{reward}. The learned value function (and in turn the RPE) assigned to a given cue at time t is then given by a learned weighted sum of the activations of these microstates at time t (**Figure 1d**).

We argue that the conception of a fixed cue-dependent temporal basis makes biologically unrealistic assumptions. First, since one does not know *a priori* whether presentation of a cue will be followed by a reward, these models assume implicitly that every single environmental cue (or combination of environmental cues) must trigger its own sequence of neural microstates, prior to potential cue-reward pairing (**Figure 1e**). Further, since one does not know a priori *when* presentation of a cue may or may not be followed by a reward, these models also assume that microstate sequences are arbitrarily long to account for all possible (or a range of possible) cue-reward delays (**Figure 1f**). Finally, these microstates are assumed to be reliably reactivated upon subsequent presentations of the cue, e.g., a neuron that represents $t_{cue} + 3dt$ must always represent $t_{cue} + 3dt$ - across trials, sessions, days, etc. However, implementation of models that generate a chain-like structure of activity can be fragile to biologically relevant factors such as noise, neural dropout, and neural drift, all of which suggest that the assumption of reliability is

problematic as well (**Figure 1g**). The totality of these observations imply that on the basis of first principles, it is hard to justify the idea of the fixed feature-specific temporal basis, a mechanism which is required for current supposedly biophysical implementations of TD learning.

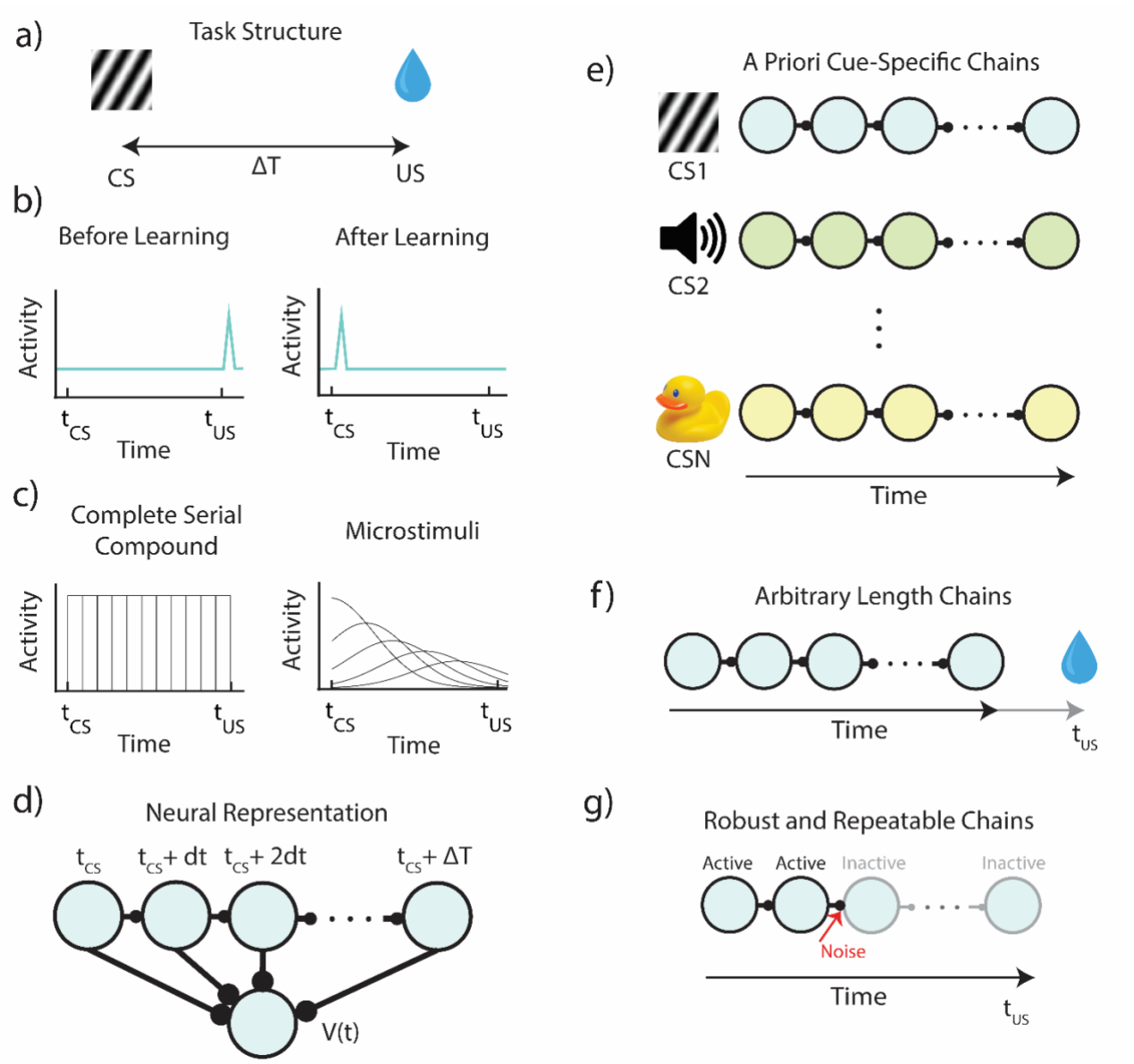


Figure 1 – Structure and Assumptions of Temporal Bases for Temporal Difference Learning a) Diagram of a simple trace conditioning task. A conditioned stimulus (CS) such as a visual grating is paired, after a delay ΔT, with an unconditioned stimulus (US) such as a water reward. **b)** According to the canonical view, neurons in VTA respond only to the US before training, and only to the CS after training. **c)** In order to represent the delay period, models generally assume neural "microstates" which span the time in between cue and reward. In the simplest case of the complete serial compound (left) the microstimuli do not overlap, and each one uniquely represents a different interval. In general, though (e.g.: microstimuli, right), these microstates can overlap with each other and decay over time. **d)** A weighted sum of these microstates determines the learned value function V(t). **e)** An agent does not know *a priori* which cue will subsequently be paired with reward. In turn, microstate TD models implicitly assume that all N unique cues or experiences in an environment each have their own independent chain of microstates *before learning*. **f)** Rewards delivered after the end of a particular cue-specific chain cannot be paired with the cue in question. The chosen length of the chain therefore determines the temporal window of possible associations. **g)** Microstate chains are assumed to be reliable and robust, but realistic levels of neural noise, drift, and variability can interrupt their propagation, thereby disrupting their ability to associate cue and reward.

Although a fixed set of basis-functions for every possible stimulus is untenable, one could assume that it is possible to replace this assumption with a single set of fixed, universal basis-functions. An example of mechanism that can generate such general basis functions is a fixed recurrent neural network (RNN). Instead of the firing of an individual neuron representing a particular time, here the entire network state can be thought of as a representation of a cue specific time. This setup is illustrated in **Figure 2a**.

To understand the consequences of this setup, we assume a simple environment in which one specific stimulus (denoted as stimulus C) is always followed 1000ms later by a reward; this stimulus is the CS. However, it is reasonable to assume for the natural world that this stimulus exists among many other stimuli that do not predict a reward. For simplicity we consider 3 stimuli, A, B and C, which can appear at any possible order, as shown in **Figure 2b**, but in which stimulus C always predicts a reward with a delay of 1000ms.

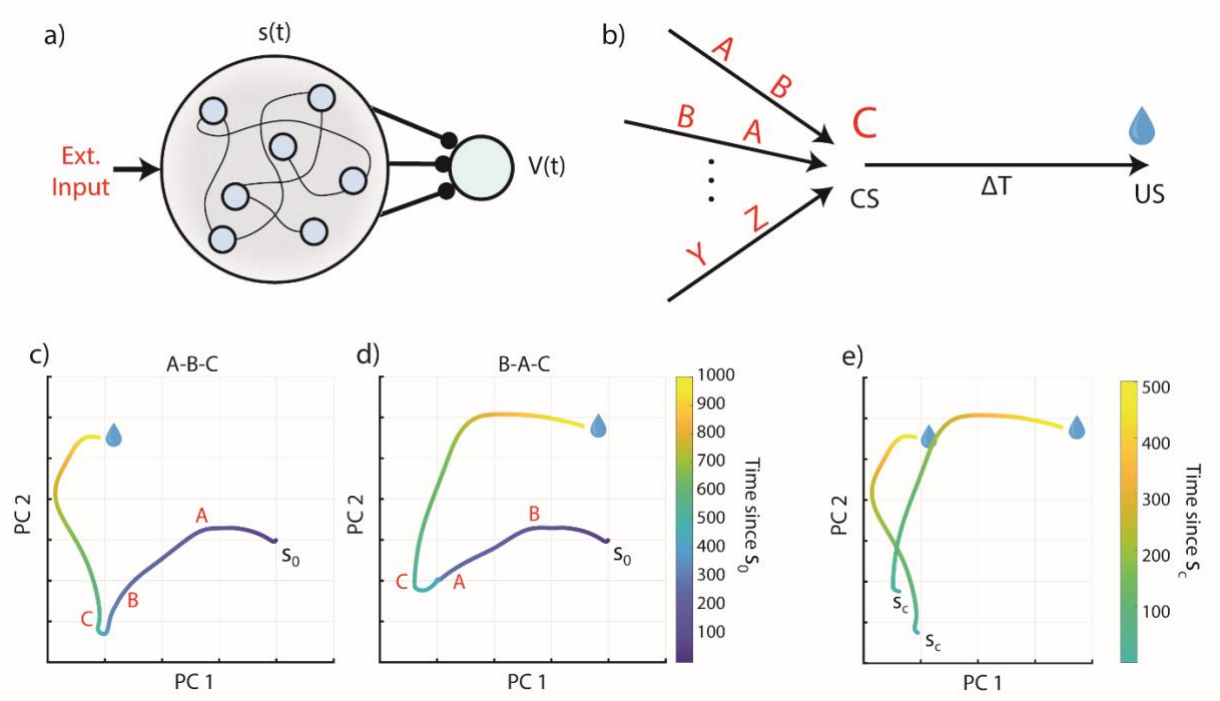


Figure 2– A fixed RNN as a basis function generator. a) Schematic of a fixed recurrent neural network as a temporal basis. The network receives external inputs and generates states s(t), which act as a basis for learning the value function V(t). Compare to figure 1c. **b)** Schematic of the task protocol. Every presentation of C is followed by a reward at a fixed delay of 1000ms. However, any combination or sequence of irrelevant stimuli may precede the conditioned stimulus C (they might also come after the CS, e.g. A,C,B). **c)** Network activity, plotted along its first two principal components, for a given initial state s_0 and a sequence of presented stimuli A-B-C (red letter is displayed at the time of a given stimulus' presentation). **d)** Same as c) but for input sequence B-A-C. **e)** Overlay of the A-B-C ad B-A-C network trajectories, starting from the state at the time of the presentation of C (state s_c). The trajectory of network activity differs in these two cases, so the RNN state does not provide a consistent temporal basis that tracks the time since the presentation of stimulus C.

We simulated the responses of such a fixed RNN to different stimulus combinations (see Methods). The complex RNN activity can be viewed as a projection to a subspace spanned by the first two principal components of the data. In **Figure 2c,d** we show a projection of the RNN response for two different sequences, A-B-C and B-A-C respectively, aligned to the time of reward. In **Figure 2e** we show the two trajectories side by side in the same subspace, starting with the presentation of stimulus C.

What these results show is that every time stimulus C appears, it generates a different temporal response in the RNN depending on the preceding stimuli. These temporal patterns can also be changed by a subsequent stimulus that may appear between the CS and US. These results mean that such a fixed RNN cannot serve as a universal basis function because its responses are not repeatable.

There are potential workarounds, such as to force the network states representing the time since stimulus C to be the same across trials. This is equivalent to learning the weights of the network such that all possible "distractor" cues pass through the network's null space. This means that the stimulus resets the network and erases its memory, but that other stimuli have no effect on the network. Generally, one would have to train the RNN to reproduce a given dynamical pattern representing C->reward, while also being invariant to noise or task-irrelevant dimensions, the latter of which can be arbitrarily high and would have to be sampled over during training.

However, this approach requires a priori knowledge that C is the conditioned stimuli (since C-> reward is the dynamical pattern we want to preserve) and that the other stimuli are nuisance stimuli. This leaves us with quite a conundrum. In the prospective view of temporal associations assumed by TD, to learn that C is associated with reward, we require a steady and repeatable labeled temporal basis (i.e. the network tracks the time since stimulus C). However, to train an RNN to robustly produce this basis, we need to have previously learned that C is associated with reward, and that the other stimuli are not. As such, these modifications to the RNN, while mathematically convenient, are based on unreasonable assumptions.

2.2 Models of TD learning with a fixed temporal-basis are inconsistent with data

Apart from being based on unrealistic assumptions, models of TD learning also make predictions that are inconsistent with experimental data. In recent years, several experiments presented evidence of neurons with temporal response profiles that resemble temporal basis functions[15–20], as depicted schematically in **Figure 3a**. While there is indeed evidence of sequential activity in the brain spanning the delay between cues and rewards (such as in the striatum and hippocampus), these sequences are generally observed *after* association learning between a stimulus and a delayed reward[15,18,19]. Some of these experiments have further shown that if the interval between stimulus and reward is extended, the response profiles either remap[18], or stretch to fill out the new extended interval[15], as depicted in **Figure 3a**. The fact that these sequences are observed after training and that the temporal response profiles are modified when the interval is changed supports the notion of plastic stimulus specific basis functions, rather than of a fixed set of basis function for each possible stimulus. Mechanistically, these results suggest that the naïve network might generate generic temporal response profile to novel stimuli before learning, resulting from the networks initial connectivity.

In the canonical version of TD learning (TD(0)), RPE neurons exhibit a bump of activity that moves back in time from the US to the CS during the learning process (**Figure 3b**-left). Subsequent versions of TD learning, called TD(λ), which speed up learning by the use of a memory trace, have a much smaller, or no noticeable moving bump (**Figure 3b**, center and right), depending on the length of the memory trace, denoted by λ. Most published experiments have not shown VTA neuron responses during the learning process. In one

prominent example by Pan et al. in which VTA neurons are observed over the learning process[21], (depicted schematically in **Figure 2c**) no moving bump is observed, prompting the authors to deduce that such memory traces exist. In a more recent paper by Amo et al. a moving bump is reported[22]. In contrast, in another recently published paper no moving bump is observed[23]. Taken together, these different results suggest that at least in some cases a moving bump is not observed. However, since a moving bump is not predicted in TD(λ) for sufficiently large λ, these results do not invalidate the TD framework in general, but rather suggest that in some cases at least the TD(0) variant is inconsistent with the data[21].

While the moving bump prediction is parameter dependent, another prediction common to all TD variants is that the integrated RPE, obtained by summing response magnitudes over the whole trial duration, does not exceed the initial US response on the first trial. This prediction is robust because the normative basis of TD is to evaluate expected reward or discounted expected reward. In versions of TD where non-discounted reward is evaluated ($\gamma = 1$) the integral of RPE activity should remain constant throughout learning. Commonly TD estimates discounted reward ($\gamma < 1$), where the discount means that rewards that come with a small delay are worth more than rewards that arrive with a large delay. With discounted rewards the integral of RPE activity will decrease with learning and become smaller than the initial US response. In contrast, we reanalyzed data from a recent experiment (Amo et al. 2022)[22,24] and found that the integrated response can transiently increase over the course of learning (**Figure 3d**).

An additional prediction of TD learning which holds across many variants is that when learning converges, and if a reward is always delivered (100% reward delivery schedule), the response of RPE neurons at the time of reward is zero. Even in the case where a small number of rewards are omitted (e.g. 10%), TD predicts that the response of RPE neurons at the time of reward is very small, much smaller than at the time of the stimulus. This seems to be indeed the case for several example neurons shown in the original experiments of dopaminergic VTA neurons[5]. However, additional data obtained more recently indicates this might not always be the case and that significant response to reward persists throughout training[21,22,25]. This discrepancy between TD and the neural data is observed both for experiments in which responses throughout learning are presented[21,22] as well as in experiments that only show results after training[25].

In experimental approaches affording large ensembles of DA neurons to be simultaneously recorded, a diversity of responses has been reported. Some DA neurons are observed to become fully unresponsive at time of reward, while others exhibit a robust response at time of reward that is no weaker than the initial US response of these cells. This is clearly exhibited by one class of dopaminergic cells (type I) that Cohen et. al[25] recorded in VTA. This diversity implies that TD is inconsistent with the results of some of the recorded neurons, but it is possible that TD does apply in some sense to the whole population. One complicating factor is that in most experiments we have no way of ascertaining that learning has reached its final state.

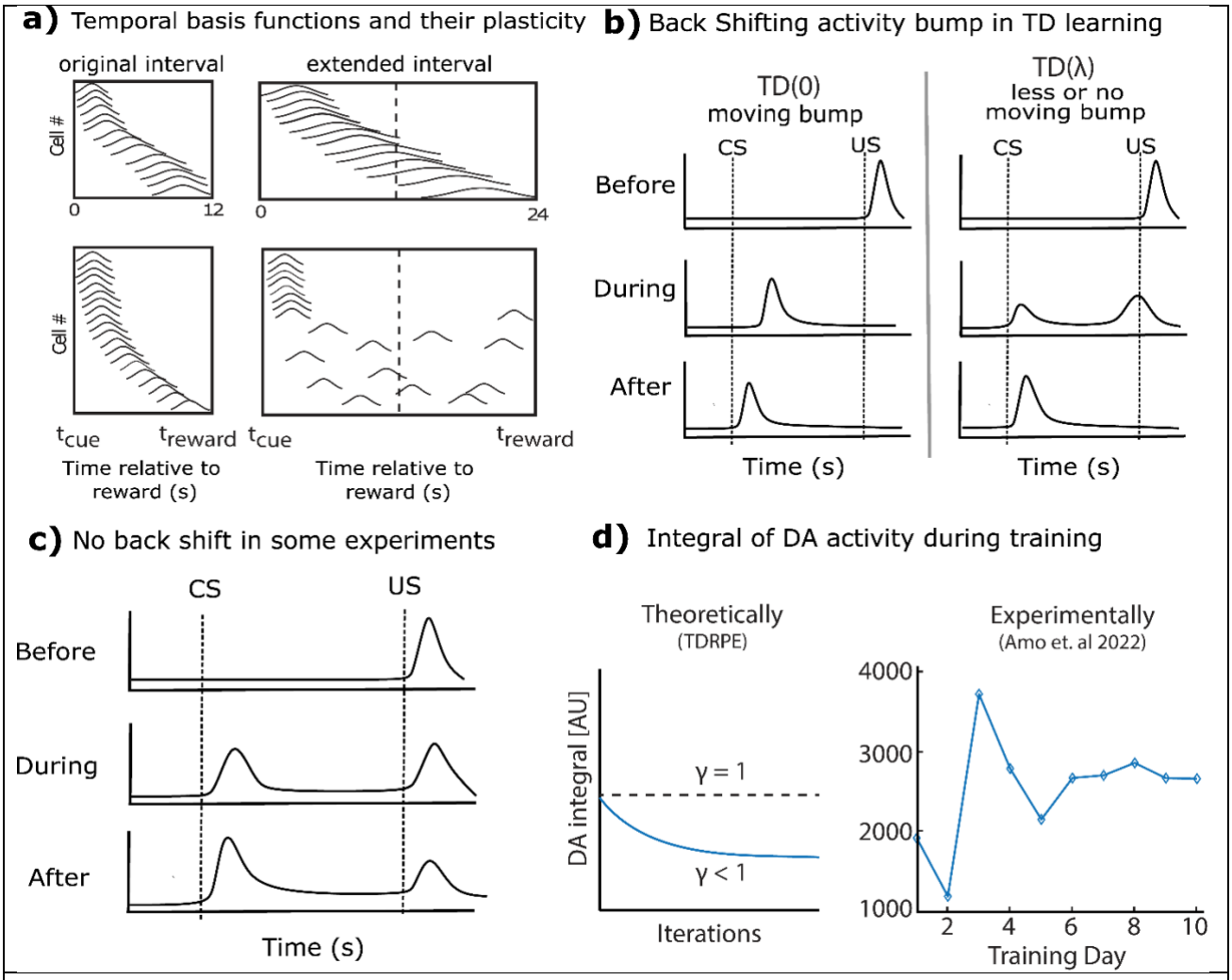


Figure 3– Certain features of experimental results run counter to predictions of TD. a) Putative temporal basis functions observed in experiments[15,18] after training are plastic, shown here schematically. If, after training, the interval between CS and US is scaled, the basis-functions also change. Recordings in striatum[15] show these basis-functions scale with the modified interval (top), while in recordings from hippocampus[18] (bottom), they are observed to redistribute to fill up the new interval. **b)** According to the $TD(0)$ theory, RPE neuron activity during learning exhibits a backward moving bump (left). For $TD(\lambda)$ the bump no longer appears (right). **c)** A schematic depiction of experiments where there is no backward shifting bump[21,23]. **d)** The integral of DA neuron activity according to TD theory (left) should be constant over training (for $\gamma = 1$) or decreasing monotonically for ($\gamma < 1$). We reanalyzed existing experimental data (right) and found that the integral can transiently significantly increase (Data from Amo et al 2022[22,24], see **Supplemental Figure 1**).

The original conception of TD is clean, elegant, and based on a simple normative foundation of estimating expected rewards. Over the years, various experimental results that do not fully conform with the predictions of TD have been interpreted as consistent with theory by making ad-hoc modifications[27,28]. Such modifications might include an assumption of different variants of the learning rule for each neuron, such that each dopaminergic neuron no longer represents RPE[27], or an assumption of additional inputs such that even when the expectation of reward is learned and fully expected dopaminergic neurons still respond at the time of reward[28].

The various modifications that are added to account for new data, and which diverge from a straightforward implementation of TD, raise the subtle question: when does TD stop being TD? We propose that changes to TD in which its normative basis of maximizing future

-discounted reward no longer holds are no longer part of the TD framework. The types of modifications added to theory to account for the experimental data are quite common in science in general. Scientific theories that no longer account for the data are often repeatedly modified before there is eventually a paradigm shift, and they are reluctantly abandoned[29].

Towards this end, a recent paper has also shown that dopamine release, as recorded with photometry, seems to be inconsistent with RPE[23]. This paper has shown many experimental results that are at odds with those expected by a RPE, and specifically, these experiments show that dopamine release at least partially represents retrospective probability of stimulus given reward. Other work has suggested that dopamine signaling is more consistent with direct learning of behavioral policy than a value-based RPE[26].

3. The FLEX theory of reinforcement learning

3.1 A theoretical framework based on a plastic temporal basis.

The FLEX theory assumes that there is a plastic (as opposed to fixed) temporal basis that evolves alongside the changing response of reward dependent neurons (such as DA neurons in VTA). The theory in general is agnostic about the functional form of the temporal basis, and several possible examples are shown in **Figure 4** (top, schematic; middle, characteristic single unit activity; bottom, population activity before and after learning). Synfire chains (**Figure 4a**), homogenous recurrent networks (**Figure 4b**), and heterogeneous recurrent networks (**Figure 4c**) could all plausibly support the temporal basis. Before learning, these basis-functions do not exist, though some neurons do respond transiently to the CS. Over learning, such basis-functions develop (**Figure 4**, bottom) in response to the rewarded stimulus, and not to unrewarded stimuli. In FLEX, we do not need a separate, predeveloped basis for every possible stimulus that spans an arbitrary amount of time. Instead, basis functions only form in the process of learning, develop only to stimuli which are tied to reward, and only span the relevant temporal interval for the behavior.

In the following, we demonstrate that such a framework can be implemented in a biophysically plausible model, and that such a model not only agrees with many existing experimental observations, but also can reconcile seemingly incongruent results pertaining to sequential conditioning. The aim of this model is to show that the FLEX theoretical framework is possible and plausible given the available data, not to claim that this implementation is a perfectly validated model of reinforcement learning in the brain. Previous models concerning hippocampus and prefrontal cortex have also considered cue memories with adaptive durations, but not explicitly in the context of challenging the fundamental idea of a fixed temporal basis[30,31].

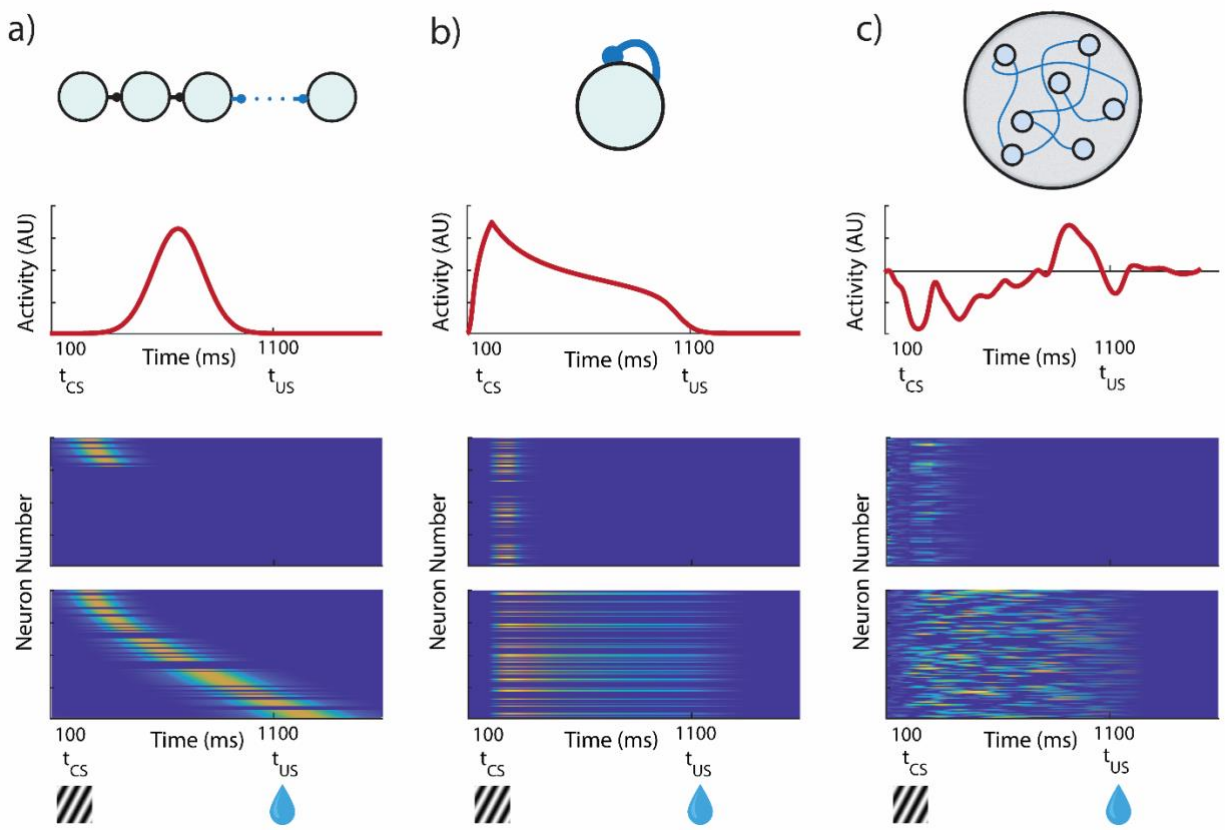


Figure 4– Potential Architectures for a Flexible Temporal Basis. Three example networks which could implement a FLEX theory. Top, network architecture. Middle, activity of one example neuron in the network. Bottom, network activity before and after training. Each network initially only has transient responses to stimuli, modifying plastic connections (in blue) during association to develop a specific temporal basis to reward-predictive stimuli. **a)** Synfire chains could support a FLEX model, if the chain could recruit more members during learning, exclusive to reward-predictive stimuli. **b)** A population of neurons with homogenous recurrent connections have a characteristic decay time that is related to the strength of the weights. The cue-relative time can then be read out by the mean level of activity in the network. **c)** A population of neurons with heterogenous and large recurrent connections (liquid state machine) can represent cue-relative time by treating the activity vector at time t as the "microstate" representing time t (as opposed to b) where only mean activity is used).

3.2 A biophysically plausible implementation of FLEX, proof-of-concept

Here we present a biophysically plausible proof-of-concept model that implements FLEX. This model is motivated by previous experimental results[16,17] and previous theoretical work in our lab[32–36]. The network's full architecture (visualized in **Figure 5a**) consists of two separate modules, a basis function module, and a reward module, here mapped onto distinct brain areas. We treat the reward module as an analogue of the VTA and the basis-function module akin to a cortical region such as the mPFC or OFC (although other cortical or subcortical regions, notably striatum[15,37] might support temporal basis functions). All cells in these regions are modeled as spiking integrate and fire neurons (see Methods).

We assume that within our basis function module are sub-populations of neurons tuned to certain external inputs, visualized in **Figure 5a** as set of discrete "columns", each responding to a specific stimulus. Within each column there are both excitatory and inhibitory cells, with a connectivity structure that includes both plastic (dashed lines, **Figure 5a**) and fixed synaptic connections (solid lines, **Figure 5a**). The VTA is composed of dopaminergic (DA) and inhibitory GABAergic cells. The VTA neurons have a background firing rate of ~5 Hz, and the DA neurons have preexisting inputs from "intrinsically rewarding" stimuli (such as a water reward). The plastic and fixed connections between the modules and from both the CS and US to these modules are also depicted in **Figure 5a**.

The model's structure is motivated by observations of distinct classes of temporally-sensitive cell responses that have evolve during trace conditioning experiments in medial prefrontal cortex (mPFC) orbitofrontal cortex (OFC) and primary visual cortex (V1)[16,17,35,38,39]. The architecture described above allows us to incorporate these observed cell classes into our basis-function module (**Figure 5b**). The first class of neurons ("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). The second class, 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)[34], a potentially canonical package of neural temporal representations. A slew of previous studies have shown these cell classes within the CNA to be a robust phenomenon experimentally[17,35,38–42], and computational work has demonstrated that the CNA can be used to learn and recall single temporal intervals[34], Markovian and non-Markovian sequences[33,43]. For simplicity, our model treats connections between populations within a single CNA as fixed (previous work has shown that such a construction is robust to perturbation of these weights[33,34]).

Learning in the model is dictated by the interaction of eligibility traces and dopaminergic reinforcement. We use a previously established two-trace learning rule[33,35,36,44] (TTL), which assumes two Hebbian activated eligibility traces, one associated with LTP and one associated with LTD (see Methods). We use this rule because it solves the temporal credit assignment problem inherent in trace conditioning, reaches stable fixed points, and since such traces have been experimentally observed in trace conditioning tasks[35]. In theory, other methods capable of solving the temporal credit assignment problem (such as a rule with a single eligibility trace[32]) 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. See the Methods section for details of the implementation of TTL used here.

Now, we will use this implementation of FLEX to simulate several experimental paradigms, showing that it can account for reported results. Importantly, some of the predictions of the model are categorically different than those produced by TD, which allows us to distinguish between the two theories based on experimental evidence.

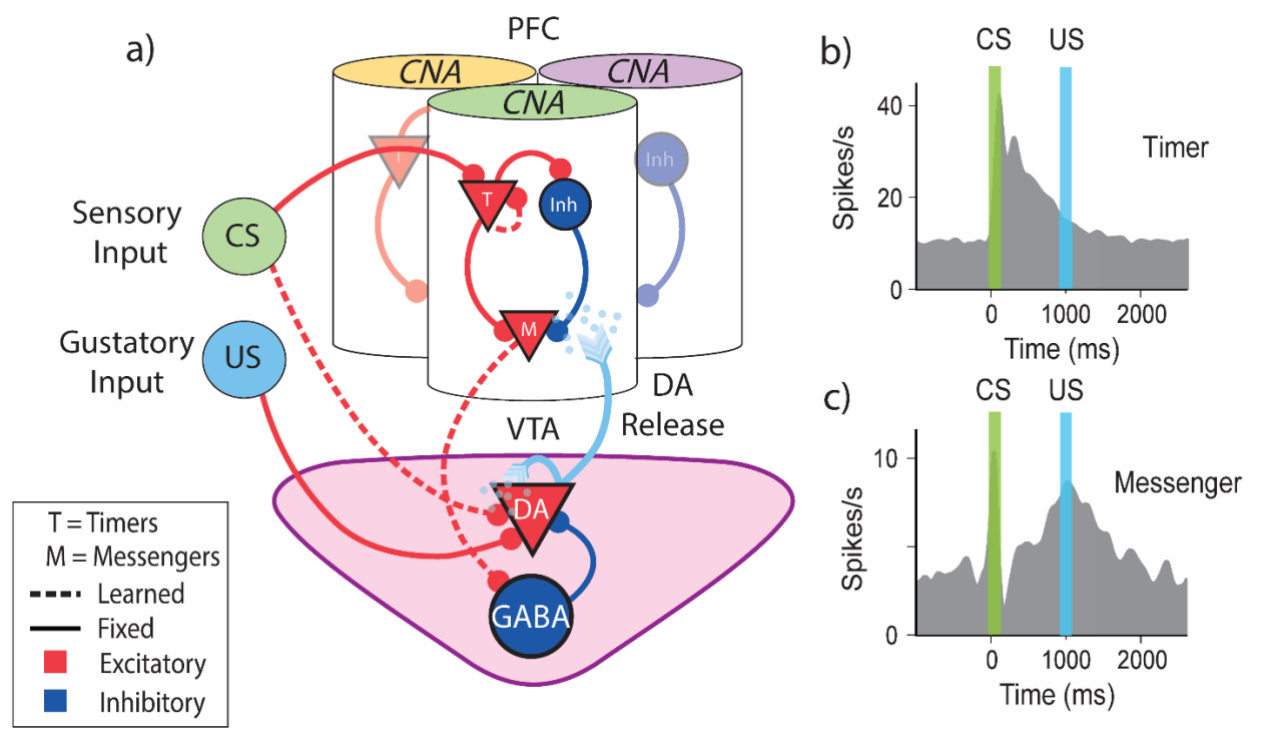


Figure 5 – Biophysically Inspired Architecture Allows for Flexible Encoding of Time **a)** Diagram of model architecture. CNAs (visualized here as columns) located in PFC are selective to certain sensory stimuli (indicated here by color atop the column) via fixed excitatory inputs (CS). VTA DA neurons 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[41]**.** **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.

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

First, we test FLEX on a basic trace conditioning task, where a single conditioned stimulus is presented, followed by an unconditioned stimulus at a fixed delay of one second (**Figure 6a**). The evolution of FLEX over training is mediated by reinforcement learning (via TTL) in three sets of weights: Timer → Timer, Messenger → VTA GABA neurons, and CS → VTA DA neurons. 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, cue neurons (CS) and feature-specific Timers are excited, producing Hebbian-activated eligibility traces at their CS->DA and T->T synapses, respectively. When the reward is subsequently presented one second later, the excess dopamine it triggers acts as a reinforcement signal for these eligibility traces (which we model as a function of the DA neuron firing rate, D(t), see Methods) causing both the cue neurons' feed-forward connections and the Timers' recurrent connections to increase (**Figure 6b**).

Over repeated trials of this cue-reward pairing, the Timers' recurrent connections continue to increase until they approach their fixed points, which corresponds to the Timers' persistent firing duration increasing until it spans the delay between CS and US

(**Supplemental Figure 2** and Methods). These mature Timers then provide a feature-specific representation of the expected cue-reward delay.

The increase of feed-forward connections from the CS to the DA neurons (6**c**) causes 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 3**). 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 Methods).

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 selectively 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 6b**). When combined with the dopamine release at t_{US}, this overlap triggers Hebbian learning at the Messenger → VTA GABA synapses (see **Figure 6c,** Methods), which indirectly suppresses the DA neurons. 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 towards baseline.

As a result of these processes, our model recaptures the traditional picture of DA neuron activity before and after learning a trace conditioning task (**Figure 6b**). 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[5], more recent calcium imaging studies have revealed that a complete suppression of the US response is not universal. Rather, many optogenetically identified dopamine neurons maintain a response to the US and show varying development of a response to the CS[21,22,25]. This diversity is also exhibited in our implementation of FLEX (**Figure 6d**) due to the connectivity structure which is based on sparse random projections from the CS to the VTA and from the US to VTA.

During trace conditioning in FLEX, the inhibition of the US-evoked dopamine response (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. At intermediate learning stages (e.g. trial 5, **Figure 6a,b**), the CS-evoked dopamine response (or equivalently, the CS → DA weights) already exhibits significant potentiation while the US-evoked dopamine response (or equivalently, inverse of the M → US weights) has only been slightly depressed. While this phenomenon has been occasionally observed in certain experimental paradigms [21,45–47], it has not been widely commented on – in FLEX, this is a fundamental property of the dynamics of learning (in particular, very early learning).

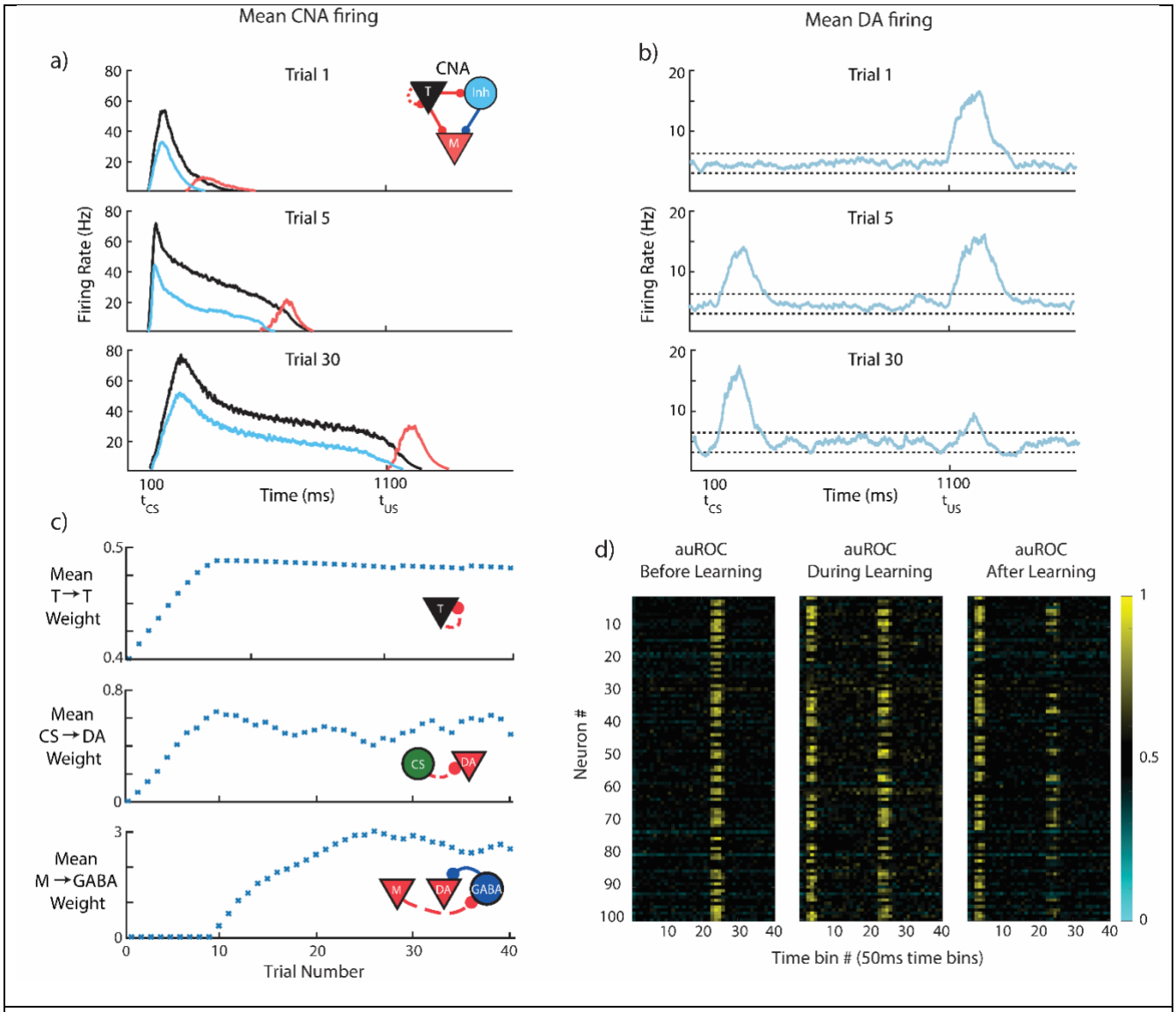


Figure 6- CS-evoked and US-evoked Model Dopamine Responses Evolve on Different Timescales. The model is trained for 40 trials while being presented with a CS at 100ms and a reward at 1100ms. **a)** Mean firing rates for the CNA (see inset for colors), for three different stages of learning. **b)** Mean firing rate over all DA neurons taken at the same three stages of learning. Firing above or below thresholds (dotted lines) evokes positive or negative D(t) in PFC. **c)** Evolution of mean synaptic weights over the course of learning. Top, middle, and bottom, mean strength of T→T, CS→DA, and M→GABA synapses, respectively. **d)** Area under receiver operating characteristic (auROC, see 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.

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 4**)[5]. 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 4**).

Dynamics of FLEX Diverge from those of TD During Conditioning

FLEX's property that the evolution of the CS responses can occur independently of (and before) depression of the US response underlies a much more fundamental and general departure of our model from TD based models. In our model, DA activity does not "travel" backwards over trials[1,5] as in TD(0), nor is DA activity transferred from one time to the other in an equal and opposite manner as in TD(λ)[6,21]. This is because our DA activity is not a strict RPE signal. Instead, while the DA neural firing in FLEX may *resemble* RPE following successful learning, the DA neural firing and RPE are *not equivalent,* as evidenced during the learning period.

To demonstrate this, we compare the putative DA responses in FLEX to the RPEs in TD(λ) (**Figure 7a**), training on the previously described trace conditioning task (see Figure 6a). We set the parameters of our TD(λ) model to match those in earlier work[21] and approximate the cue and reward as Gaussians centered at t_{CS1} and t_{CS2}, respectively. In TD models, by definition, integral of the RPE over the course of the trial is always less than or equal to the original total RPE provided by the initial unexpected presentation of reward. In other words, the error in reward expectation cannot be larger than the initial reward. In both TD(0) and TD(λ), this quantity of "integrated RPE" is conserved; for versions of TD with a temporal discounting factor γ (which acts such that a reward with value r_0 presented n timesteps in the future is only worth $r_0\gamma^n$ where $\gamma \leq 1$), this quantity decreases as learning progresses.

In FLEX, by contrast, integrated dopaminergic release D(t) during a given trial can be greater than that evoked by the original unexpected US (see **Figure 7b**), and therefore during training the DA signal in FLEX diverges from a reward prediction error. This property has not been explicitly investigated, and most published experiments do not provide continuous data during the training phase. However, to test our prediction, we re-analyzed recently published data which does cover the training phase[22], and found that there is indeed a significant transient increase in the dopamine release during training (**Figure 3d and Supplemental Figure 1**). Another recent publication found that initial DA response to the US was uncorrelated with the final DA response to the CS[26], which also supports the idea that integrated dopamine release is not conserved.

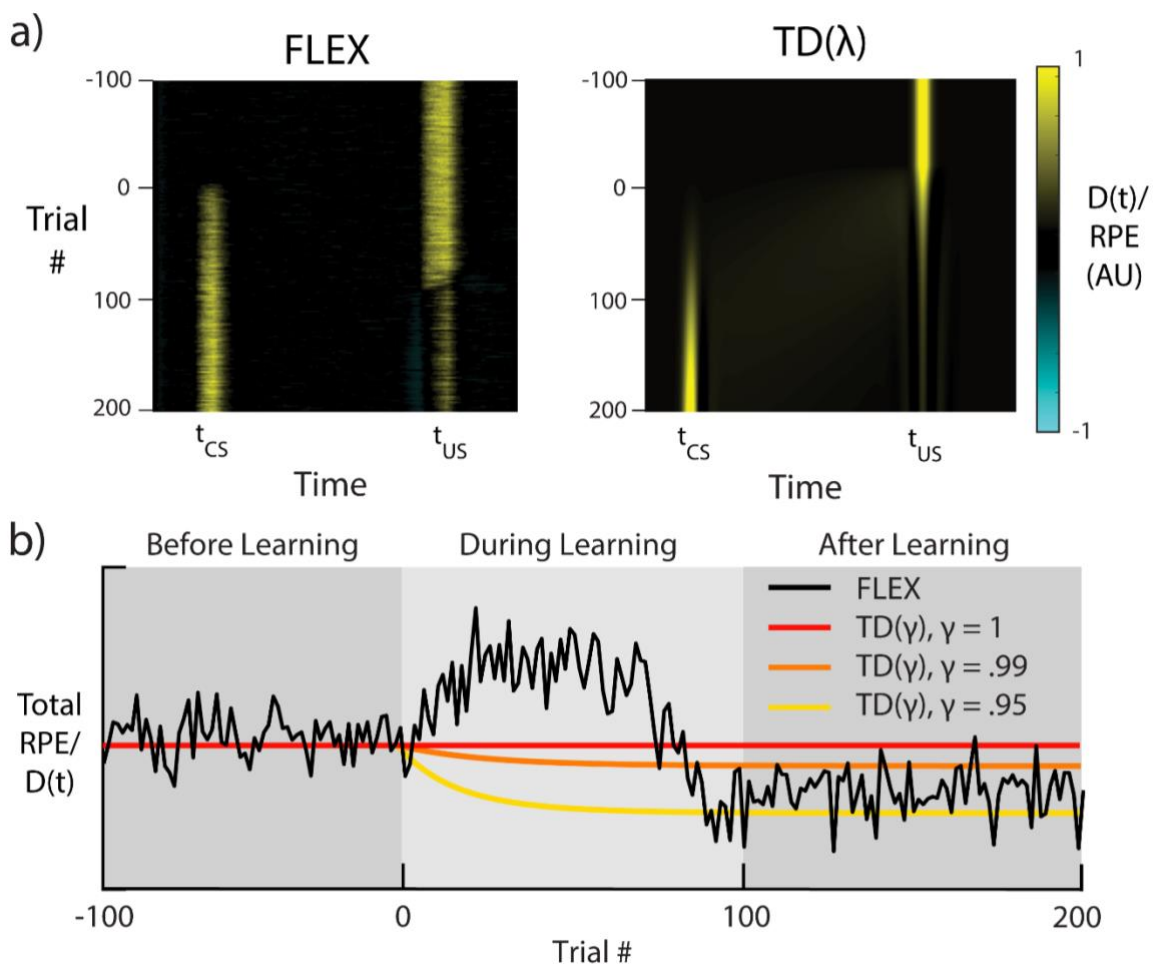


Figure 7 – FLEX Model Dynamics Diverge from those of TD Learning

Dynamics of both TD(λ) and FLEX when trained with the same conditioning protocol as shown in Figure 6. **a)** Dopaminergic release D(t) for FLEX (left), and RPE for TD(λ) (right), over the course of training. **b)** Sum total (i.e., the sum of each row in **a)** is a point on the line of **b)**) 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. The learning rate in our model is reduced in this example, as to make direct comparison with TD(λ). 100 trials of US-only presentation (before learning) are included for comparison with subsequent stages.

FLEX Unifies Sequential Conditioning Results

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[48,49]. 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 predictive of the reward[21,23]. Surprisingly, FLEX can reconcile these seemingly contradictory results. In **Figure 8** we show simulations of sequential conditioning using FLEX. In early training we observe an emerging response to both CS1 and CS2, as well as to the US (**Figure 8b,ii**). Later on the response to the US is suppressed (**Figure 8b,iii**). During late training (**Figure 8b,iv**) the response to CS2

is suppressed and all activation is transferred to the earliest predictive stimulus, CS1. These evolving dynamics can be compared to the different experimental results. Early training is similar to the results of Pan et al. (2005)[21] and Jeong et al. (2022)[23], as seen in **Figure 8c**, while the late training results are similar to the results of Schultz (1993)[48] as seen in **Figure 8d**.

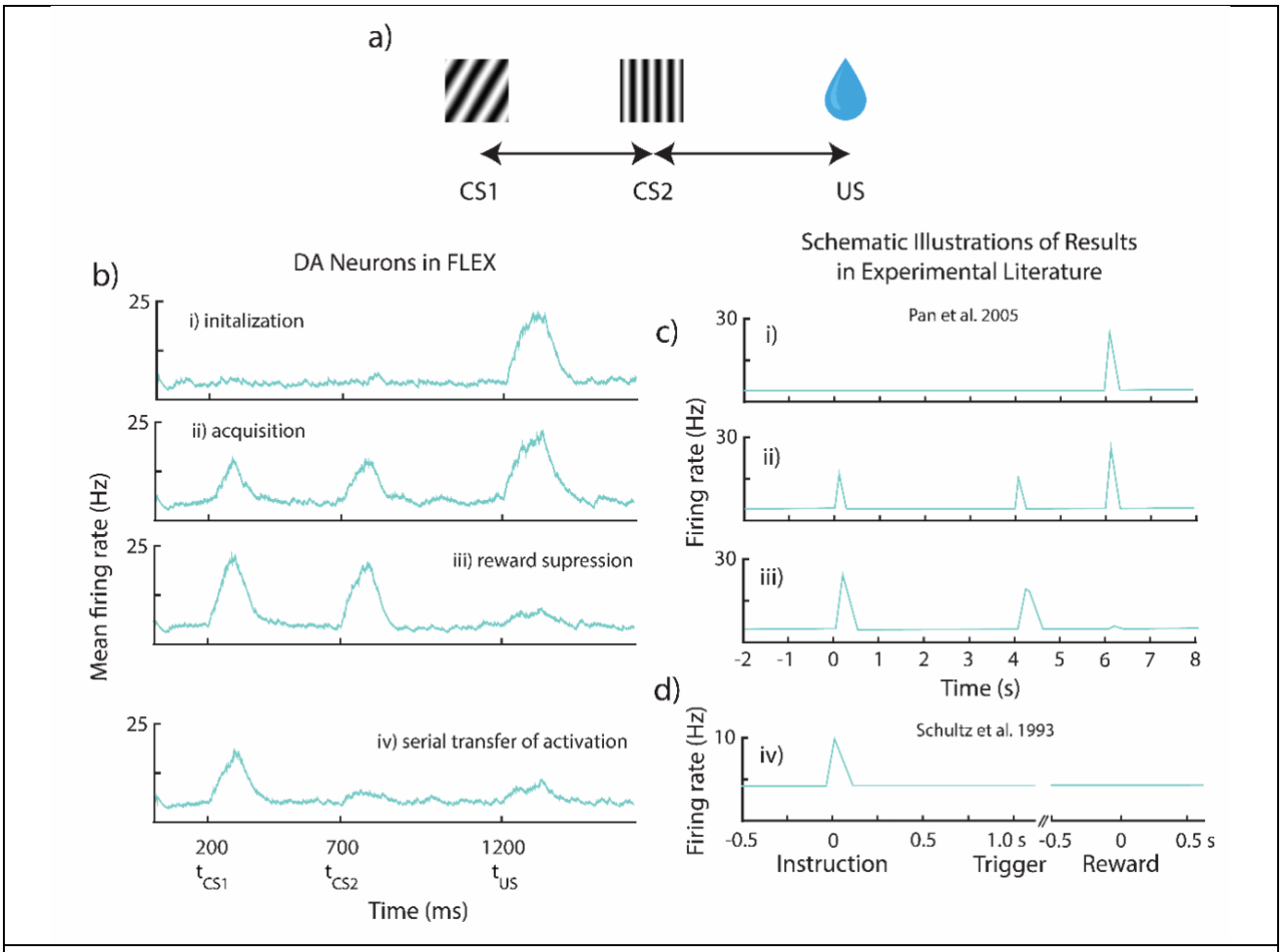


Figure 8 – FLEX model Reconciles Differing Experimental Phenomena Observed during Sequential Conditioning

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 four distinctive stages in learning – initialization(i), acquisition(ii), reward depression(iii), and serial transfer of activation(iv). **c)** 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[21]. **d)** DA neuron firing after training wherein two cues (instruction, trigger) were followed by a single reward. Adapted from Schultz et al. 1993[48].

After training, in FLEX, both sequential cues still affect the dopamine release at the time of reward, as removal of either cue results in a partial recovery of the dopamine response at t_{US} (**Supplemental Figure 5**). This is the case even in late training in FLEX, when there is a positive dopamine response only to first cue – our model predicts that removal of the second cue will result in a positive dopamine response at the time of expected reward. In contrast, the RPE hypothesis would posit that after extended training, the value function would

eventually be maximized following the first cue, and therefore removal of the subsequent cue would not change dopamine release at t_{US}.

FLEX is also capable of replicating results of a different set of sequential learning paradigms (**Supplemental Figure 6**). In these protocols, the network is initially trained on a standard trace conditioning task with a single CS. 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 on its own does not become reward predictive, a phenomenon called "blocking"[50–52]. However, if reward magnitude is increased or additional dopamine is introduced to the system, a response to the intermediate CS (CS2) emerges, a phenomenon termed "unblocking"[53,54]. Each of these phenomena can be replicated in FLEX (**Supplemental Figure 6**).

4. Discussion

TD has established itself as one of the most successful models of neural function to date, as its predictions regarding RPE have, to a large extent, matched experimental results. However, two key factors make it reasonable to consider alternatives to TD as a modeling framework for how midbrain dopamine neurons could learn RPE. First, attempts for biologically plausible implementations of TD have previously assumed that even before learning, each possible cue triggers a separate chain of neurons which tile an arbitrary period of time relative to the cue start. The a priori existence of such an immense set of fixed, arbitrarily long cue-dependent basis functions is both implausible and inconsistent with experimental evidence. Second, different conditioning paradigms have revealed dopamine dynamics that are incompatible with the predictions of models based on the TD framework[21,23,55,56].

To overcome these problems, we suggest that the temporal basis itself is not fixed, but instead plastic, and is learned only for specific cues that lead to reward. We call this theoretical framework FLEX. We also presented a biophysically plausible implementation of FLEX, and we have shown that it can generate flexible basis functions, and that it produces dopamine cell dynamics that are consistent with experimental results. Our implementation should be seen as a proof-of-concept model. It shows that FLEX can be implemented with biophysical components, and that such an implementation is consistent with much of the data. It does not show that the specific details of this implementation, (including the brain regions in which the temporal-basis is developed, the specific dynamics of the temporal basis functions, and the learning rules used) are those used by the brain.

One of the appealing aspects of TD learning is that it arises from simple normative assumption that the brain needs to estimate future expected reward. Does FLEX have a similar normative basis? Indeed, in the final state of learning, the responses of DA neurons in FLEX resemble RPE neurons in TD, however in FLEX there is no analog for the value neurons assumed in TD. Unlike in TD, activity of FLEX DA neurons in response to the cue represent an *association* with future expected rewards, independent of a *valuation*.

Instead, the goal of FLEX is to learn the association between cue and reward, develop the temporal basis functions that span the period between the two, and transfer DA signaling from the time of the reward to the time of the cue. These basis functions could then be used as a mechanistic foundation for brain activities, including the timing of actions. In essence, DA in FLEX acts to create internal models of associations and timings. Recent experimental evidence[26] which suggests that DA correlates more with the direct learning of behavioral policies rather than value-encoded prediction errors is broadly consistent with this view of dopamine's function within the FLEX framework.

Another recent publication[23] has questioned the claim that DA neurons indeed represent RPE, instead hypothesizing that DA release is also dependent on retrospective

probabilities[23,57]. The design of most historical experiments cannot distinguish between these competing hypotheses. In this recent research project, a set of new experiments were designed specifically to test these competing hypotheses, and the results obtained are inconsistent with the common interpretation that DA neurons simply represent RPE[23]. More generally, while the idea that the brain does and should estimate economic value seems intuitive, it has been recently questioned[58]. This challenge to the prevailing normative view is motivated by behavioral experimental results which instead suggest that a heuristic process, which does not faithfully represent value, often guides decisions.

Recent papers have questioned the common view of the response of DA neurons in the brain and their relation to value estimation[8,23,26,58]. Here we survey additional problems with implementations of TD algorithms in neuronal machinery, and propose an alternative theoretical formulation, FLEX, along with a computational implementation of this theory. The fundamental difference from previous work is that FLEX postulates that the temporal basis-functions necessary for learning are themselves learned, and that neuromodulator activity in the brain is an instructive signal for learning these basis functions. Our computational implementation has predictions that are different than those of TD and are consistent with many experimental results. Further, we tested a unique prediction of FLEX (that integrated DA release across a trial can change over learning) by re-analyzing experimental data, showing that the data was consistent with FLEX but not the TD framework.

Methods

Network Dynamics

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

Equation 1 - 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 2 - Synaptic Activation Dynamics

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

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 3 - 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 4 - Firing Rate

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

Fixed RNN

For the network in **Figure 2**, the dynamics of the units u_i in the RNN are described by the equation below:

Equation 5 – RNN Dynamics

$$\tau_{net} \frac{du_i}{dt} = -u_i + \sum_k W_{ik} \phi(u_k)$$

Where u_i are the firing rates of the units in the RNN. W_{ik} are the recurrent weights of the RNN, each of which is drawn from a normal distribution $N(0, \frac{g}{\sqrt{K}})$, where g is the "gain" of the network[59] and ϕ is a sigmoidal activation function. Each of the inputs given to the network (A, B, or C) is a unique, normally distributed projection of a 100ms step function.

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[33,35,36]. 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

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

Equation 8 - LTD Trace Dynamics

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

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, $\mathrm{H_{ij}}$, is a simple multiplication $r_i \cdot r_j$ for application of this rule in VTA, where r_j and $\mathrm{r_i}$ are the time-averaged firing rates at the pre- and post-synaptic cells. Experimentally, the "Hebbian" terms ($\mathrm{H_{ij}}$) which impact LTP and LTD trace generation are complex[35], but in VTA we approximate with the simple multiplication $r_i \cdot r_j$. For synapses in PFC, we make the alteration that $\mathrm{H_{ij}} = \frac{r_i \cdot r_j}{1+\alpha D(t)}$, acting to restrict PFC trace generation for large positive RPEs. The alteration of $\mathrm{H_{ij}}$ 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[60]. Functionally, this altered $\mathrm{H_{ij}}$ 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 $\mathrm{T^a_{max}}$) 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

$$\frac{\mathrm{dW_{ij}}}{\mathrm{dt}} = \eta \mathrm{D(t)} \left(\mathrm{T^p_{ij}(t)} - \mathrm{T^d_{ij}(t)} \right)$$

where η is the learning rate.

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[35,61–64]. Notably, in such experiments, the eligibility traces in prefrontal cortex were found to convert into positive synaptic changes via delayed application of dopamine[35], 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, $\frac{\mathrm{dW_{ij}}}{\mathrm{dt}} = \eta \mathrm{D(t) r_i r_j}$. 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 $\mathrm{T^p_{ij}(t)} = \mathrm{T^d_{ij}(t)}$. This is because, by construction, the function of M→GABA learning is to suppress D(t) down to zero. Therefore, the fixed point $\mathrm{T^p_{ij}(t)} = \mathrm{T^d_{ij}(t)}$ 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[34], and further that a single mean field neuron can well approximate the activity of each of these subpopulations of spiking neurons[65].

The two-trace learning rule we utilize for our model is described in further detail in previous work[33,35,36,44]. 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 2 and **Supplemental Figure 3** 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 2**). 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 2**). After the recurrent weights have reached this fixed point and the Timer neurons encode the cue-reward delay (Trial 30 in **Supplemental Figure 2**), M→GABA learning acts to suppress D(t) down to zero as well (Trial 40 in **Supplemental Figure 2**). 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[60].

For feed-forward learning, the weights initially evolve identically to the recurrent weights (Trial 1 in **Supplemental Figure 3**). 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 3**). 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 3**). 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 3**).

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. (2012)[25], 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 6d**,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 maximally discriminable increase in spikes compared to baseline, 0 demonstrating a maximally discriminable decrease of spikes compared to baseline, and .5 demonstrating an inability to discriminate between the two distributions.

Data from Amo et al. 2022 was used for **Supplemental Figure 1**[22,24]. The data from 7 animals (437-440, 444-446) is shown here. This data is already z-scored as described by Amo et al. On every training day only the first 40 trials are used, unless there was a smaller number of rewarded trials, in which case that number was used. For every animal per every day, the integral from the time of the CS to the end of the trial is calculated and averaged over all trials in that day. These are the data points in **Supplemental Figure 1a**, with each animal represented by a different color. The blue line is the average over animals. The bar graphs in **Supplemental Figure 1b** represent the average over all animals in early (day 1-2), intermediate (day 3-4) and late (day 8-10) periods. Wilcoxon rank sum tests find that intermediate is significantly higher than early (p=0.004) and that late is significantly higher than early (p=0.006). Late is not significantly lower than intermediate. We also find that late is significantly higher than early if we take the average per animal over the early and intermediate days (p=0.02). A MATLAB code that carries out this analysis is posted on modelDB together with the code for FLEX.

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. Sutton, R. S. & Barto, A. G. *Reinforcement Learning, second edition: An Introduction*. (MIT Press, 2018).
2. Glickman, S. E. & Schiff, B. B. A biological theory of reinforcement. *Psychol. Rev.* **74**, 81–109 (1967).
3. Lee, D., Seo, H. & Jung, M. W. Neural Basis of Reinforcement Learning and Decision Making. *Annu. Rev. Neurosci.* **35**, 287–308 (2012).
4. A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement. in *Models of Information Processing in the Basal Ganglia* (eds. Houk, J. C., Davis, J. L. & Beiser, D. G.) (The MIT Press, 1994). doi:10.7551/mitpress/4708.003.0020.
5. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593–1599 (1997).
6. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
7. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Evaluating the TD model of classical conditioning. *Learn. Behav.* **40**, 305–319 (2012).
8. Namboodiri, V. M. K. How do real animals account for the passage of time during associative learning? *Behav. Neurosci.* **136**, 383–391 (2022).
9. 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).
10. Tesauro, G. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Comput.* **6**, 215–219 (1994).
11. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Process. Mag.* **34**, 26–38 (2017).
12. Foster, D. J., Morris, R. G. & Dayan, P. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* **10**, 1–16 (2000).

13. Montague, P., Dayan, P. & Sejnowski, T. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
14. Schultz, W. Behavioral Theories and the Neurophysiology of Reward. *Annu. Rev. Psychol.* **57**, 87–115 (2006).
15. Mello, G. B. M., Soares, S. & Paton, J. J. A Scalable Population Code for Time in the Striatum. *Curr. Biol.* **25**, 1113–1122 (2015).
16. 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).
17. Kim, E., Bari, B. A. & Cohen, J. Y. Subthreshold basis for reward-predictive persistent activity in mouse prefrontal cortex. *Cell Rep.* **35**, 109082 (2021).
18. Taxidis, J. *et al.* Differential Emergence and Stability of Sensory and Temporal Representations in Context-Specific Hippocampal Sequences. *Neuron* **108**, 984-998.e9 (2020).
19. MacDonald, C. J., Lepage, K. Q., Eden, U. T. & Eichenbaum, H. Hippocampal "Time Cells" Bridge the Gap in Memory for Discontiguous Events. *Neuron* **71**, 737–749 (2011).
20. Parker, N. F. *et al.* Choice-selective sequences dominate in cortical relative to thalamic inputs to NAc to support reinforcement learning. *Cell Rep.* **39**, 110756 (2022).
21. 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).
22. Amo, R. *et al.* A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nat. Neurosci.* **25**, 1082–1092 (2022).
23. Jeong, H. *et al.* Mesolimbic dopamine release conveys causal associations. *Science* **0**, eabq6740 (2022).

24. Amo, R., Matias, S., Uchida, N. & Watabe-Uchida, M. Population and single dopamine neuron activity during classical conditioning. Dryad, Dataset (2022) doi:10.5061/DRYAD.HHMGQNKJW.

25. 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).

26. Coddington, L. T., Lindo, S. E. & Dudman, J. T. Mesolimbic dopamine adapts the rate of learning from action. *Nature* **614**, 294–302 (2023).

27. Dabney, W. *et al.* A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).

28. Mikhael, J. G., Kim, H. R., Uchida, N. & Gershman, S. J. The role of state uncertainty in the dynamics of dopamine. *Curr. Biol.* **32**, 1077–1087 (2022).

29. Kuhn Thomas. *The structure of scientific revolutions*. vol. 111 (University of Chicago Press:, 1970).

30. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).

31. O'Reilly, R. C., Frank, M. J., Hazy, T. E. & Watz, B. PVLV: The Primary Value and Learned Value Pavlovian Learning Algorithm. *Behav. Neurosci.* **121**, 31–49 (2007).

32. 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).

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

34. 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).

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

36. 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).

37. Aosaki, T. *et al.* Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. *J. Neurosci.* **14**, 3969–3984 (1994).

38. 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).

39. Shuler, M. G. Reward Timing in the Primary Visual Cortex. *Science* **311**, 1606–1609 (2006).

40. Namboodiri, V. M. K., Huertas, M. A., Monk, K. J., Shouval, H. Z. & Hussain Shuler, M. G. Visually Cued Action Timing in the Primary Visual Cortex. *Neuron* **86**, 319–330 (2015).

41. Liu, C.-H., Coleman, J. E., Davoudi, H., Zhang, K. & Hussain Shuler, M. G. Selective Activation of a Putative Reinforcement Signal Conditions Cued Interval Timing in Primary Visual Cortex. *Curr. Biol.* **25**, 1551–1561 (2015).

42. Monk, K. J., Allard, S. & Hussain Shuler, M. G. Reward Timing and Its Expression by Inhibitory Interneurons in the Mouse Primary Visual Cortex. *Cereb. Cortex N. Y. NY* **30**, 4662–4676 (2020).

43. Zajzon, B., Duarte, R. & Morrison, A. Towards reproducible models of sequence learning: replication and analysis of a modular spiking network with reward-based learning. *bioRxiv* 2023.01.18.524604 (2023) doi:10.1101/2023.01.18.524604.

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

45. Ljungberg, T., Apicella, P. & Schultz, W. Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.* **67**, 145–163 (1992).

46. 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).

47. 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).

48. 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).

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

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

51. 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).

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

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

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

55. 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).

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

57. Namboodiri, V. M. K. & Stuber, G. D. The learning of prospective and retrospective cognitive maps within neural circuits. *Neuron* **109**, 3552–3575 (2021).

58. Hayden, B. Y. & Niv, Y. The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behav. Neurosci.* **135**, 192 (2021).

59. Rajan, K., Abbott, L. F. & Sompolinsky, H. Stimulus-Dependent Suppression of Chaos in Recurrent Neural Networks. *Phys. Rev. E* **82**, 011903 (2010).

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

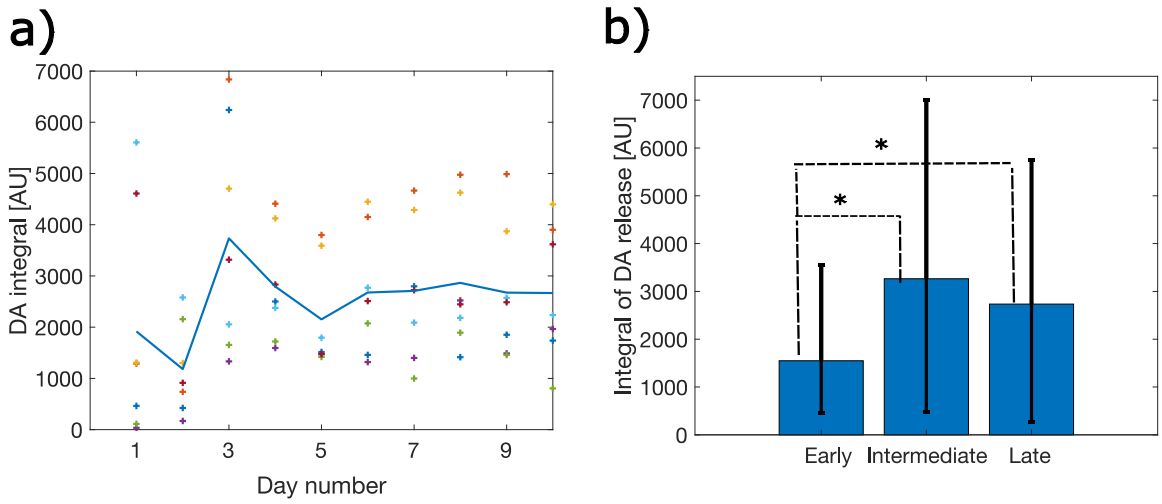
62. Hong, S. Z. *et al.* Norepinephrine potentiates and serotonin depresses visual cortical responses by transforming eligibility traces. *Nat. Commun.* **13**, 3202 (2022).

63. Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D. & Brea, J. Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. *Front. Neural Circuits* **12**, (2018).

64. Brzosko, Z., Schultz, W. & Paulsen, O. Retroactive modulation of spike timing-dependent plasticity by dopamine. *eLife* **4**, e09685 (2015).

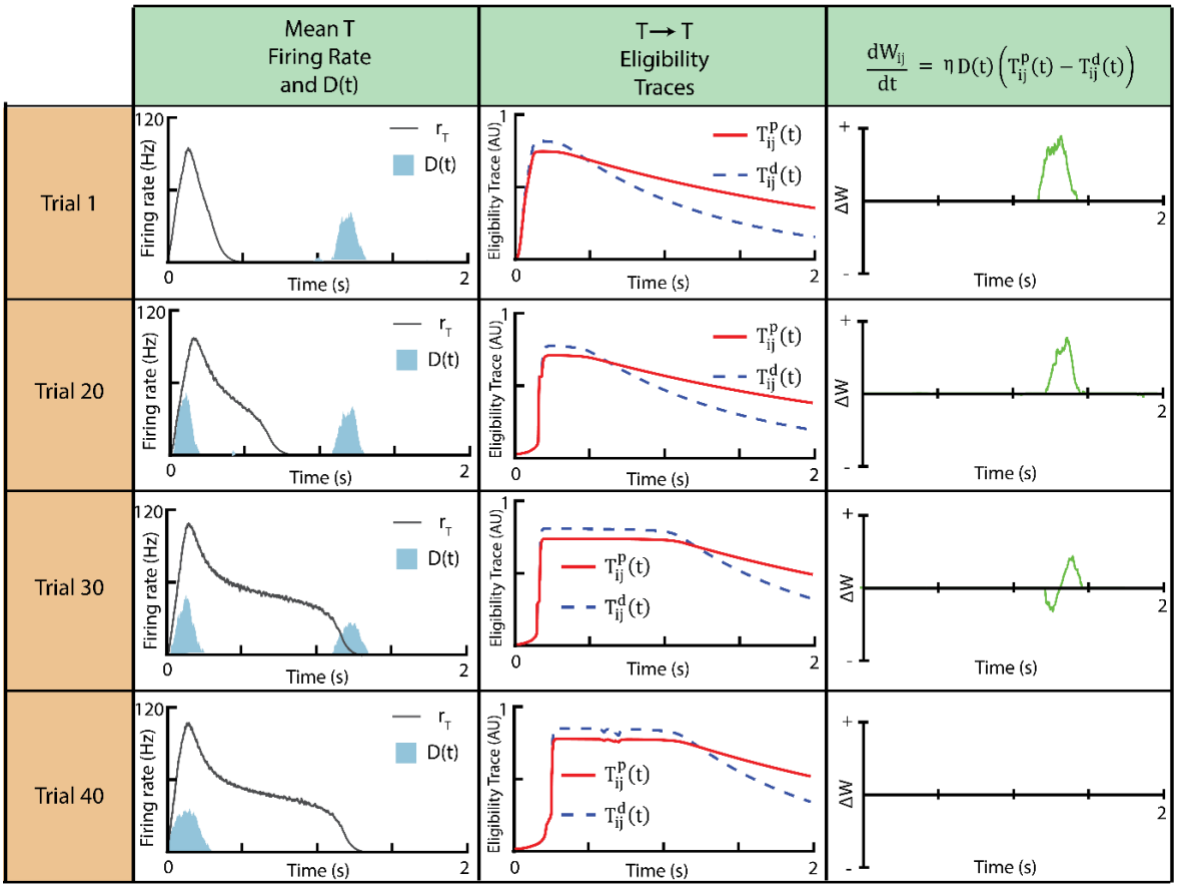
65. Gavornik, J. P. & Shouval, H. Z. A network of spiking neurons that can represent interval timing: mean field analysis. *J. Comput. Neurosci.* **30**, 501–513 (2011).

Supplemental Figures

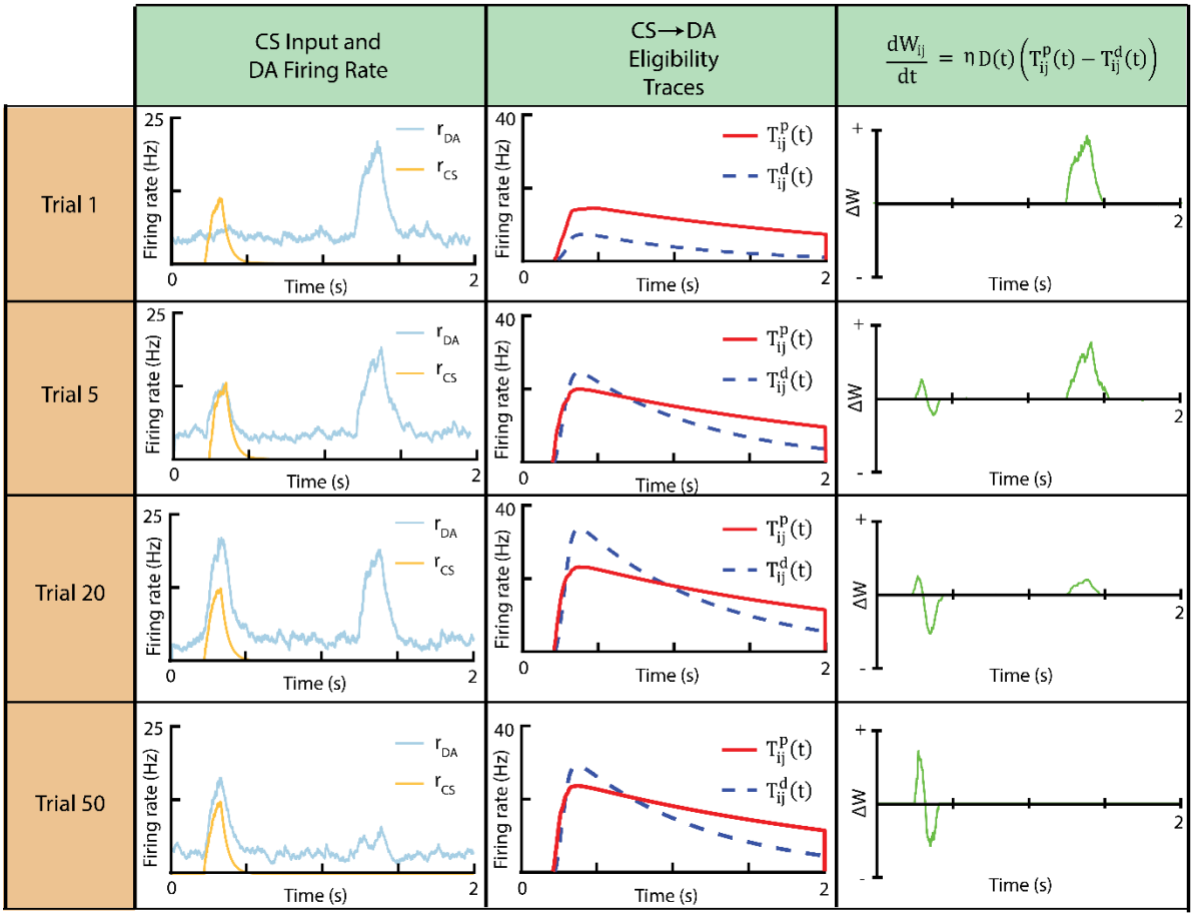


Supplemental Figure 1 – Dynamics of the DA integral.

a) Re-analyzed data from Amo et al (2022)[22]. Here we plot the time evolution of the integral of DA response over training days. The different color symbols are the average DA integral (averaged over trials) per each training day, per each animal. Different animals are color coded. The blue line is the mean over animals. **b)** Average integrals for early (days 1-2), intermediate (days 3-4) and late (days 8-10) of training. Early is significantly lower than intermediate (Wilcoxon rank sum test p=0.004) and late (p=0.006), but intermediate is not significantly higher than late.

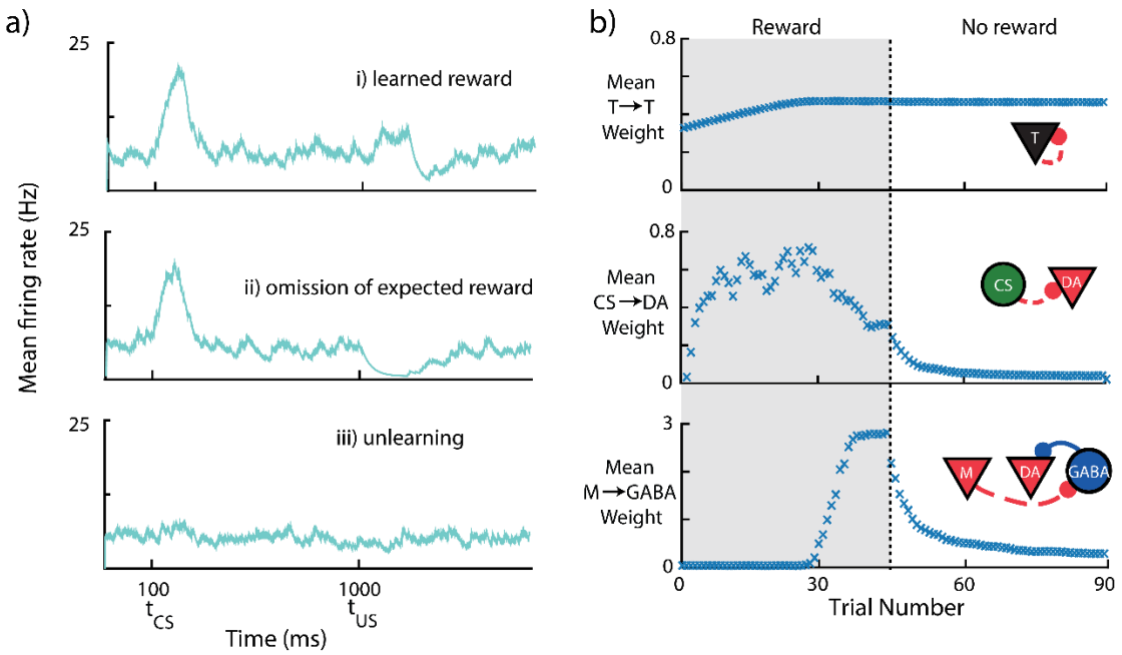


Supplemental Figure 2 – Two-Trace Learning Flexibly Encodes Cue-Reward Delay in Recurrent Connections of Timer Neurons Demonstration of the dynamics of two-trace learning for recurrent connections. Rows consist of different trials. Left column, mean firing rate of Timer neurons (black) and dopamine reinforcement 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 ($T_{ij}^{p} - T_{ij}^{d}$) 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).



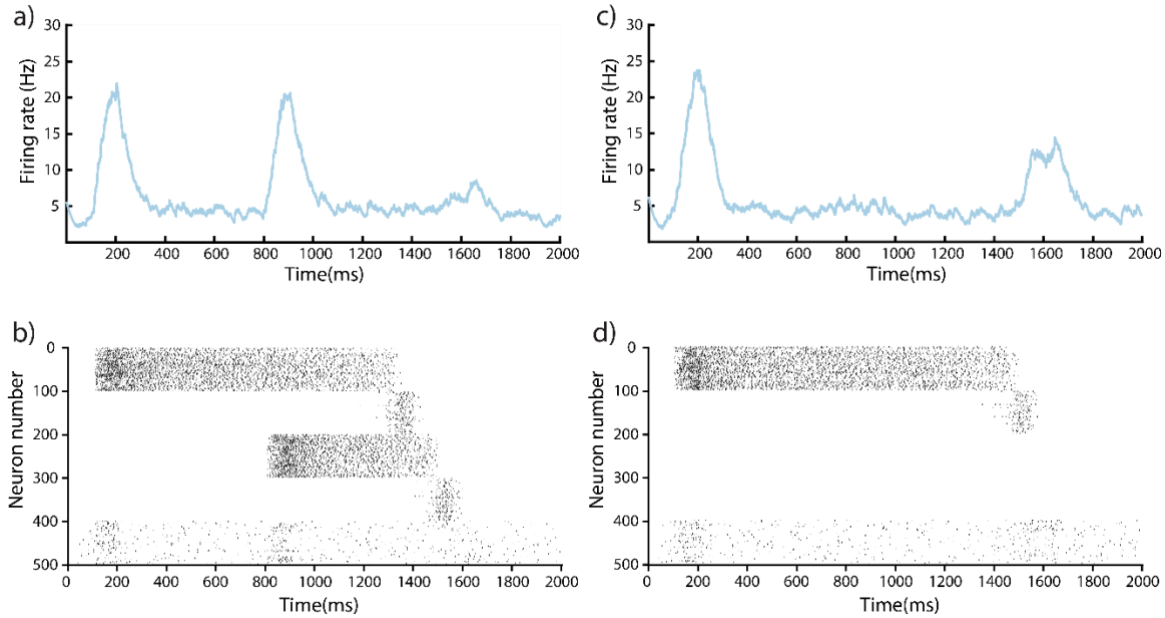
Supplemental Figure 3 – Two-Trace Learning Encodes Reward Predictive Cues via Feed-Forward Connections to DA Neurons

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.



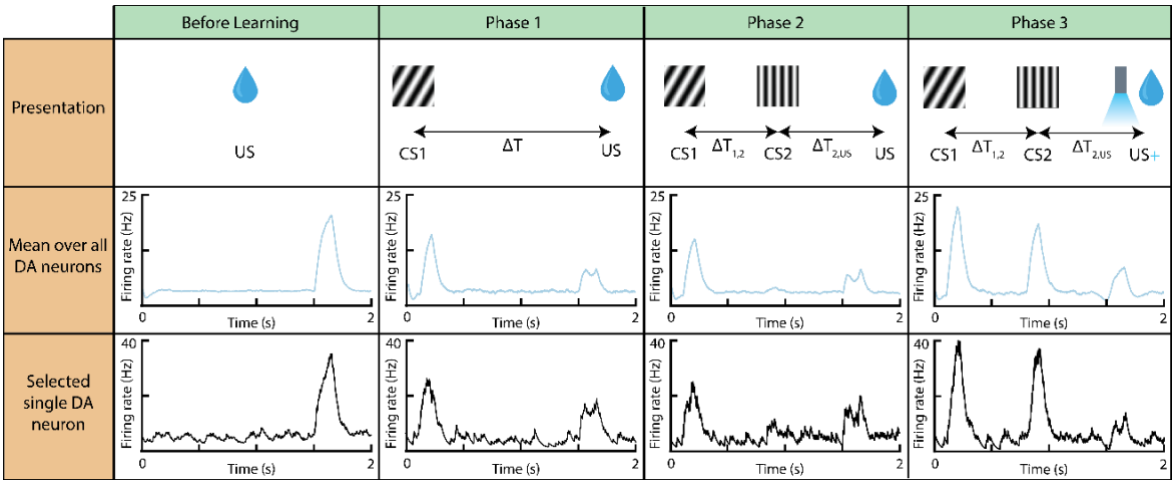
Supplemental Figure 4 – Consistent Unpairing Leads to Negative RPEs and Unlearning

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 2**) is agnostic to changes in to D(t).



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

a) Mean firing rate of all VTA DA 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 DA 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 – 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.