Reinforcement learning with dopamine: a convergence of natural and artificial intelligence

Paul Masset[1] and Samuel J. Gershman[2,3]
[1]Department of Psychology, McGill University
[2]Department of Psychology and Center for Brain Science, Harvard University
[3]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University

July 18, 2024

Abstract

Reinforcement learning is the problem of predicting and maximizing long-term reward. Computer scientists recognized that this problem could be solved by updating predictions and action policies based on prediction errors (discrepancies between observed and expected reward). Remarkably, a similar strategy appears to be used by the basal ganglia, where dopamine provides the prediction errors for updating predictions and action policies in the striatum. We review how this convergence of natural and artificial intelligence has been elaborated and challenged, focusing on recent developments that connect cutting-edge machine learning algorithms to experimental observations. A recurring theme, both theoretically and empirically, is the surprising power of simple error-driven learning algorithms when equipped with suitably rich (and potentially distributed) state representations. These representations are in turn modified by dopaminergic prediction errors, creating a virtuous cycle whereby learning algorithms can amplify their ability to solve more complex tasks.

1 Introduction

Reinforcement learning (RL)—the study of how agents (both natural and artificial) can learn to predict and control future reward—emerged from a multidisciplinary synthesis of ideas from engineering, psychology, and neuroscience. The original inspiration for RL algorithms came from animal studies of learning by trial and error (Thorndike, 1898; Pavlov, 1927; Skinner, 1938). Early computational models formalized how this kind of learning might work, using simple scalar errors to update reward expectations (Widrow et al., 1973; Klopf, 1972; Sutton and Barto, 1981). The apotheosis of these ideas was the temporal difference (TD) learning algorithm, which could provably solve a broad class of sequential decision problems (described further below). Variants of this algorithm have since been used to achieve human-level performance on challenging tasks such as arcade games (Mnih et al., 2015), computer chip design (Mirhoseini et al., 2021), language modeling (Ouyang et al., 2022), and many more.

The TD algorithm has also been used to explain many aspects of animal learning and its neurophysiological basis. Most notably, the hypothesis that dopamine reports the error signal used by TD for updating reward expectations (Montague et al., 1996; Schultz et al., 1997) has had a huge impact in neuroscience. The fact that the same algorithms appear to be useful for engineering and

implemented by the brain is unlikely to be a coincidence—it is an example of convergence, where intelligent systems are destined to hit upon certain algorithms because they solve a broad class of problems (Gershman, 2024). Biology is filled with examples of convergent evolution (e.g., echolocation, eyes, flight, opposable thumbs), where species independently attain similar phenotypes. This phenomenon may be at work in the emergence of intelligence across natural and artificial systems.

In this chapter, we will introduce the mathematical concepts necessary to understand the key ideas behind RL, with a focus on the TD learning algorithm.[1] We will then look at the neurobiology of RL through the lens of TD learning, summarizing and evaluating the available evidence. Finally, we will describe how recent extensions developed in the machine learning literature can help us understand some of the currently perplexing aspects of dopamine and its interactions with other parts of the brain.

2 The reinforcement learning problem

The goal of RL is to predict and maximize long term reward. Formally, an agent collects immediate reward r_t at time t, accumulating a return R over a horizon of H time steps:

$$R = r_1 + r_2 + \quad r_H: \tag{1}$$

If the horizon is infinite ($H = 1$), as commonly assumed, this return may diverge. One way to deal with this problem is to assume that later rewards are discounted exponentially, which ensures that the sum converges:

$$R^\gamma = r_1 + \quad r_2 + \quad {}^{H\ 1} r_H; \tag{2}$$

where $2\ [0;1)$ is known as the *discount factor*. Psychologically, we can interpret in terms of an intertemporal preference for sooner rewards.[2] The discount factor can also be interpreted as a fixed survival probability: if the agent terminates the task with a fixed probability of 1 at each time step, then on average its undiscounted return (truncated by the random termination) will equal the discounted return (Sozou, 1998).

In addition to randomness induced by termination, the rewards themselves may be random variables (e.g., how much food a foraging animal harvests may depend on randomness in patch yields, which patch it visits, etc.). The agent therefore needs to consider its *expected* return, $\mathsf{E}[R^\gamma]$, averaging over all these sources of randomness. This is the quantity that most RL models assume animals are trying to estimate and maximize.[3]

The RL problem is easy to state but not easy to solve. How are agents supposed to estimate a quantity that involves averaging a possibly infinite sequence of random events? Rendering this

[1] In the interest of broad accessibility, we will avoid extensive formalism. For a more complete introduction and formal mathematical treatment of reinforcement learning, we refer the reader to excellent books on the topic (Sutton and Barto, 2018; Bertsekas, 2019; Szepesvári, 2022).

[2] Evidence suggests that intertemporal choice behavior is better described by hyperbolic, rather than exponential, discounting (Berns et al., 2007; Vanderveldt et al., 2016). We will return to this discrepancy below.

[3] Some models formalize the RL problem in terms of other objective functions, such as average reward. These models have been important for understanding some aspects of dopamine (Daw and Touretzky, 2002; Niv et al., 2007; Mikhael and Gershman, 2022), though for brevity we do not address them here.

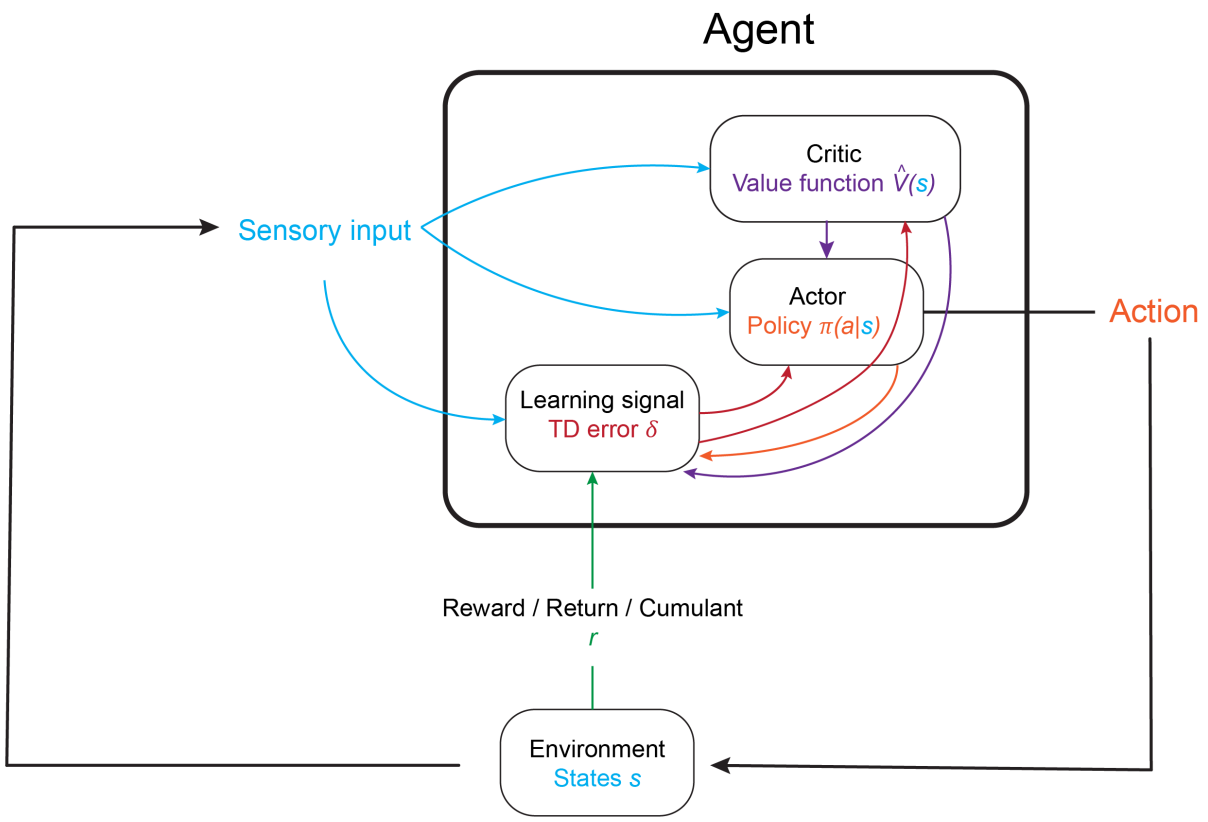

Figure 1: **The reinforcement learning problem**. An agent interacts with its environment by generating actions from a policy. The environment then produces state transitions and rewards, from which the agent updates its value estimates (e.g., using TD learning) and possibly also a world model.

problem tractable requires some additional assumptions (Figure 1). In particular, most models assume that the rewards are *conditionally independent* given the time-varying *state* of the environment, denoted s. Conditional independence means that the probability of reward at time t depends only on the state at time t—all other past and future rewards can be ignored. If we additionally assume that the states themselves are conditionally independent (i.e., the probability of the next state depends only on the current state), we have what is known as a *Markov decision process*. The key advantage of making these assumptions is that the expected return can be decomposed into a recursive form known as the *Bellman equation*:

$$V(s) = \mathrm{E}[R^{\gamma} | s] = \mathrm{E}[r + \quad V(s^{0})]; \tag{3}$$

where $V(s)$ is the expected return starting at state s, also known as the *value function*, and s^{0} is the state at the next time step.[4] The Bellman equation is the basis of most RL algorithms studied in both AI and neuroscience, as we discuss in the next section.

[4] In the interest of notational compactness, we will omit time subscripts where possible.

3 The temporal difference learning algorithm

The Bellman equation enables an elegant and simple algorithmic solution to the RL problem, the TD learning algorithm (Sutton and Barto, 2018), which we mentioned in the Introduction. The key idea is to start with some initial estimate of the value function, $\hat{V}$, and progressively improve this estimate by interacting with the environment. The Bellman equation defines a consistency condition between consecutive values; violations of this consistency instruct the agent how to improve its estimate. To see this, let's define a new random variable, the *TD error*:

$$= r + \quad \hat{V}(s^0) \quad \hat{V}(s): \tag{4}$$

Using the Bellman equation, we can see that $\mathsf{E}[\ ] = 0$ when $\hat{V}(s) = V(s)$. In other words, the TD error measures (on average) the discrepancy between the true and estimated value function. Furthermore, the sign of the TD error is informative about the direction of the discrepancy: if < 0, this indicates that the agent is overestimating the value of the current state and should therefore decrease $\hat{V}(s)$, whereas if > 0, then the agent is underestimating the value and should therefore increase $\hat{V}(s)$. These prescriptions are the essence of the TD algorithm. In its simplest form, it says that the estimate for the current state should be updated in proportion to the TD error:

$$\hat{V}(s) \ / \ : \tag{5}$$

In practice, this only works when states are discrete and their value estimates are stored in a look-up table (the "tabular" setting). Most natural environments include continuous and/or high-dimensional states, which make look-up tables impractical. It is therefore more common to define the value estimates in terms of a parametrized function approximator, $\hat{V}_{\mathbf{w}}$, and then update the parameters $\mathbf{w}$ based on the TD error:

$$\mathbf{w} \ / \quad r_{\mathbf{w}} \hat{V}_{\mathbf{w}}(s); \tag{6}$$

where $r_{\mathbf{w}} \hat{V}_{\mathbf{w}}(s)$ is the gradient of the function approximator, which determines how credit (for positive errors) and blame (for negative errors) should be assigned to the parameters.

To make this more concrete, consider a linear function approximator (widely used in the computational neuroscience literature):

$$\hat{V}_{\mathbf{w}}(s) = \sum_i w_i \ _i(s); \tag{7}$$

where $f \ _i g$ is a set of *basis functions* defining a feature space for value function approximation. Henceforth we will use the abbreviation $_i(s) = \ _i$. The parameters $\mathbf{w}$ correspond to weights on the basis functions. Eq. 6 then becomes:

$$w_i \ / \quad _i: \tag{8}$$

Intuitively, credit or blame is assigned to each weight in proportion to the activation of its corresponding feature. We will defer a discussion of what these features are and where they come from to a later section.

So far, we have focused on the problem of value estimation, but this is only half of the RL problem; the other half is the problem of value maximization. The brain can interact with the

environment by selecting actions, which affect the state transitions and rewards. We can formalize this as a state-dependent action policy $\pi_\theta(a|s)$, parametrized by θ. In the tabular setting, the policy parameters are often modeled as action "preferences" which are mapped into action probabilities via a softmax function:

$$\pi_\theta(a = j|s) = \exp[\theta_{sj}] / \sum_i \exp[\theta_{si}]. \tag{9}$$

In the linear function approximation setting, the policy parameters typically correspond to action weights:

$$\pi_\theta(a = j|s) = \exp\left[\sum_i \theta_{ij} f_i\right] / \sum_i \exp\left[\sum_i \theta_{ij} f_i\right]. \tag{10}$$

The action weights can be updated to improve the expected return. We can derive an update rule by following the gradient of the value with respect to the policy parameters:

$$\theta_{ij} \propto \alpha \delta f_j, \tag{11}$$

where

$$f_j = \mathbb{I}[a = j] \quad \pi_\theta(j|s), \tag{12}$$

what we will refer to as the *action trace*; $\mathbb{I}[\cdot]$ is the indicator function (equaling 1 when its argument is true, 0 otherwise), and a is the chosen action. This update uses the TD error for policy improvement. Intuitively, the TD error is positive when the agent tries an action which improves the expected return; the update then shifts the weights to make this action more likely. The opposite happens when the TD error is negative. The TD error thus acts as a "critic" of the "actor" (the policy), which is why this is known as an *actor-critic* architecture (Barto et al., 1983). One reason to focus on the actor-critic architecture here is that it has become one of the workhorses of modern deep RL, responsible for some of its most spectacular successes (e.g., Lillicrap et al., 2015; Haarnoja et al., 2018; Andrychowicz et al., 2020).

Our brief synopsis of RL concepts does not exhaust the space of value estimation and policy improvement algorithms (see Sutton and Barto, 2018), focusing on those concepts which have been most influential in computational neuroscience. We next turn to a discussion of how these have been mapped onto brain circuitry.

4 The neurobiology of temporal difference learning

The TD learning algorithm has played an important role in the contemporary understanding of dopamine and the basal ganglia, starting with the observation that phasic dopamine activity appears to track the TD error (Eq. 4; see Chapter 22). As noted by Schultz et al. (1997), dopamine neurons produce a burst of activity following delivery of an unexpected reward, and pause firing when an expected reward is omitted. When reward is reliably predicted by a cue, the burst of activity moves backward toward the cue (see also Amo et al., 2022), consistent with the hypothesis that stimulus features are temporally distributed and thus structure the progression of credit assignment. Importantly, studies have shown that dopamine activity conforms to detailed quantitative properties of TD errors (Bayer and Glimcher, 2005; Eshel et al., 2015, 2016; Kim et al.,

2020), and that causal manipulations of dopamine produce behaviorally measurable effects consistent with the TD learning algorithm (Tsai et al., 2009; Steinberg et al., 2013; Chang et al., 2016; Salinas-Hernández et al., 2018; Xie et al., 2023).

What is the site of plasticity induced by dopamine? A common assumption is that cortex conveys sensory information (feature vectors), which are then mapped into values via a function approximation architecture in the medium spiny neurons (MSNs) of the striatum (Doya, 2008). Using the linear function approximator as the simplest form of such an architecture, the parameters $\mathbf{w}$ correspond to corticostriatal synaptic strengths. This implies that the value update (Eq. 8) is a synaptic plasticity rule in which dopamine (signaling) interacts multiplicatively with the presynaptic firing rate (signaling $_i$). Similarly, the policy update (Eq. 11) is a three-factor rule which depends also on the postsynaptic firing rate (signaling f_j). This plasticity rule is broadly consistent with studies of corticostriatal plasticity (Reynolds and Wickens, 2002), although the literature discloses many complexities (Perrin and Venance, 2019; Sippy and Tritsch, 2023, see also Chapter 18).

One important complexity is the physiology of different cell types. Most MSNs in the striatum express either D1 or D2 receptors, which respond to dopamine in different ways and have different downstream consequences. Briefly, dopamine excites D1-expressing MSNs and inhibits D2-expressing MSNs (Surmeier et al., 2007). D1-expressing MSNs in the dorsal striatum project primarily to the "direct" (striatonigral) pathway, which ultimately facilitates motor commands, thereby facilitating action. These cells are therefore commmonly referred to as *direct spiny projection neurons* (dSPNs). In contrast, D2-expressing MSNs project primarily to the "indirect" (striatopallidal) pathway, which ultimately suppresses motor commands. These cells are therefore commmonly referred to as *indirect spiny projection neurons* (iSPNs). The differential activation of these two cell types influences action selection (Kravitz et al., 2012; Freeze et al., 2013; Lee and Sabatini, 2021). They also exhibit different patterns of corticostriatal plasticity, with dopamine promoting synaptic potentiation in dSPNs and depression in iSPNs (Shen et al., 2008; Sippy and Tritsch, 2023). Computational models (e.g., Frank, 2005; Collins and Frank, 2014; Möller and Bogacz, 2019; Pinto and Uchida, 2023; Lindsey et al., 2024) have incorporated these ideas into biologically plausible policy parametrizations.

To illustrate, let us assume that each cell type's activation is a linear function of the cortical state features:

$$y_j^{\mathrm{D1}} = \sum_i \; {}^{\mathrm{D1}}_{ij} \; {}_i \tag{13}$$

$$y_j^{\mathrm{D2}} = \sum_i \; {}^{\mathrm{D2}}_{ij} \; {}_i; \tag{14}$$

where we have assumed separate cells selective to each action j. We can then parametrize a version of the softmax policy in which the probability of selecting action j depends on the difference between these two activations (the action preferences):

$$(a = j|s) \; / \; \exp\left[y_j^{\mathrm{D1}} \quad y_j^{\mathrm{D2}}\right]: \tag{15}$$

The policy improvement update then becomes:

$${}^{\mathrm{D1}}_{ij} \; / \quad {}_i f_j \tag{16}$$

$${}^{\mathrm{D2}}_{ij} \; / \quad {}_i f_j: \tag{17}$$

The updates are identical except for the sign, consistent with the opposite effects of dopamine on plasticity in dSPNs and iSPNs. In order for the updates to be interpreted as Hebbian (i.e., based on the co-occurrence of presynaptic and postsynaptic activity), we would need to interpret f_j (the action trace) as the activity of both cell types. This implies that activity of the two cell types become correlated after action selection, with co-activation of cells tuned to the chosen action. Evidence for this hypothesis has been reported by Lindsey et al. (2024).

Although we have been referring to f_j as the action trace, it might be more appropriately termed an action *prediction error*, because it reflects the difference between the chosen and predicted action (see also Lindsey and Litwin-Kumar, 2022). If we follow the hypothesis that this is signaled by post-choice activity of SPNs, then we expect these cells to be maximally excited when their preferred action is chosen and low probability, whereas they will be maximally suppressed when their preferred action is unchosen and high probability. In partial support of this hypothesis, Markowitz et al. (2018) showed that SPN activity is lower following high probability actions compared to low probability actions.

The actor-critic architecture requires a segregation of circuits computing the value estimate ($\hat{V}$) and the action preferences (parametrized by the dSPNs and iSPNs). A venerable hypothesis (Houk et al., 1995; Joel et al., 2002) asserts that the value estimate is computed in the ventral striatum and the action preferences are computed in the dorsal striatum. This division of labor is consistent with the observation that action preference signals are typically prevalent in the dorsal striatum but relatively weak or absent in ventral striatum (Samejima et al., 2005; Pasquereau et al., 2007; Lau and Glimcher, 2008; Kim et al., 2009; Ito and Doya, 2015). The hypothesis also asserts that the TD error is computed with respect to the ventral striatal value estimate, and that the same error signal is projected to both the ventral and dorsal striatum, driving plasticity in both regions (though the plasticity rules are different, as summarized above). Recordings of activity in different dopamine projections confirm that ventral and dorsal striatum receive essentially the same TD error signal (Tsutsui-Kimura et al., 2020), though other data suggest more regional heterogeneity (van Elzelingen et al., 2022).

5 A dialogue between artificial intelligence and neuroscience

So far, we have been discussing a fairly classical computational picture of the brain's RL system. In the rest of the chapter, we broaden this picture to consider how recent developments in AI (particularly deep learning) have changed this picture (see also Botvinick et al., 2020; Gershman and Ölveczky, 2020). These developments have influenced modern thinking about every aspect of the RL framework, including how we define values, rewards, states, features, and policies. Our goal is to show how some of these developments also shed light on the neuroscience of RL, and in particular on some aspects of dopamine function that at first glance does not seem to fit with the canonical TD error hypothesis. We explore how some of these non-canonical responses could be signatures of more complex RL algorithms.

5.1 What is the state?

The concept of state has a technical definition in RL: it is the variable which renders the conditional independence assumptions of a Markov decision process true. In other words, the state summarizes all the information about the past that is needed to predict the future. As discussed

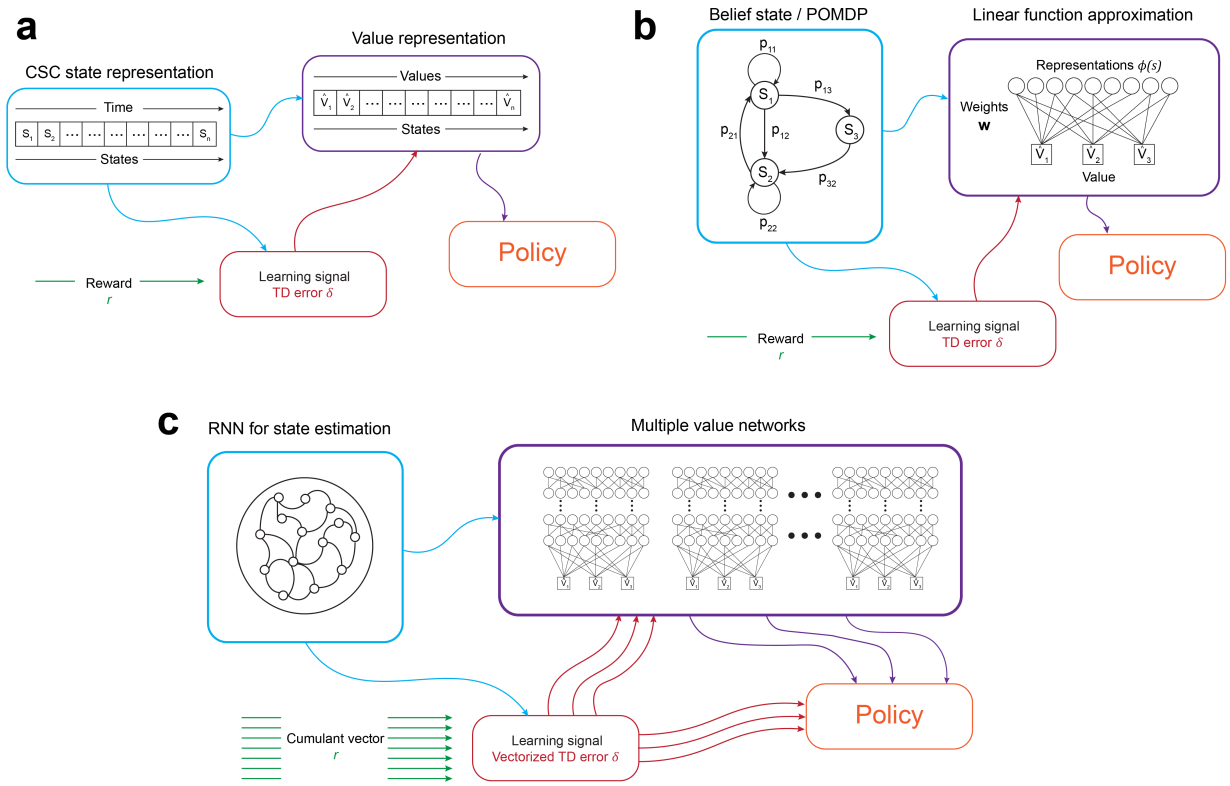

Figure 2: Increasingly complex reinforcement learning architectures. **a.** In the complete serial compound (CSC) model, states correspond to a discretized axis of the time spent in the episode. A value is learned for each time bin. Most models using the CSC as the state representation have been used to simulate Pavlovian tasks, so the policy is to simply choose the next state. **b.** Belief states and linear function approximation. A more advanced implementation of RL would build a model of the world based on a Markov decision process (MDP) or a partially observable Markov process (POMDP) that captures the transition probabilities across states. In the POMDP setting, there is the additional uncertainty that states are not directly observed and there is therefore uncertainty about the current state of the environment. Value is also not directly represented through a 1-to-1 mapping with state but instead is approximated using linear function approximation (eq 8. In a more modern RL architecture, the state representations could be learned via a recurrent neural network (RNN), allowing it to learn the dynamics of state transitions. The system could potentially learn from several reward streams with separate value networks, allowing the implementation of distributional RL or general value functions (see Section 5). The policy network would allow the system to learn how to combine the value networks to achieve optimal performance.

in Section 2, this is critical to the validity of efficient algorithms like TD learning. Thus, if the brain uses these algorithms, we should expect it to extract states from its sensory inputs.

Understanding how states in the environment are represented is crucial to estimate the form the TD error will take when an agent experiences or explores an environment. In the canonical reward prediction error model of Schultz et al. (1997), states are represented using the complete serial compound (CSC) representation in which the time within a trial is discretized and each timestep corresponds to a state (Figure 2a). This representation is useful in tasks where a few potential cues predict the timing and size of rewards and punishments, as is often the case in neuroscience experiments, but it might be very inefficient in more complex environments. In this section, we will discuss several examples in which different state representations can affect the form of the TD error as it would be measured experimentally though dopamine signaling. A key takeaway is that in order to make strong claims about the link between dopamine signaling and TD error, it is necessary to have a good handle on the structure of state representations and of the internal state of an agent at any given time.

Implicit in the models described earlier is the assumption that the state is directly observable—it can be extracted unambiguously from sensory inputs. This assumption is unrealistic, since sensory inputs often do not perfectly disambiguate the state. This "partial observability" arises in two common scenarios. One is where sensory data are noisy or incomplete (e.g., observing the silhouette of a person in the distance). Another is where predicting the future depends on something that happened in the past which is not recorded in current sensory inputs.

It is possible to address partial observability by mapping sensory inputs to a probability distribution over the hidden state, known as the *belief state* (Kaelbling et al., 1998). This is truly a state in the sense that it satisfies the conditional independence assumptions, and hence is compatible with algorithms like TD learning. More formally, let $\mathbf{x}$ denote the vector of sensory inputs, and let $\mathbf{b}$ denote the belief state, where each dimension corresponds to a single state such that $b_s = P(s|\mathbf{x})$, the posterior distribution over states conditional on $\mathbf{x}$. The posterior can be computed using Bayes' rule:

$$b_s \propto P(\mathbf{x}|s)P(s); \tag{18}$$

where $P(\mathbf{x}|s)$ is the *likelihood* of the sensory inputs given a hypothetical state s, and $P(s)$ is the *prior probability* of s. The likelihood expresses the consistency between the sensory inputs and the hypothetical state. The prior is the agent's belief state before observing $\mathbf{x}$. Because the belief state is a sufficient statistic for prediction, it can substitute for the hidden state (which the agent doesn't have access to) in RL algorithms.

From a neuroscience perspective, the general picture is that the cortical inputs to striatum signal belief states, and corticostriatal plasticity learns the mapping from belief states to values and policy parameters (Daw et al., 2006; Rao, 2010). This picture is consistent with the idea that cortex (or at least parts of it) computes posterior probabilities via some form of approximate Bayesian inference (Lee and Mumford, 2003; Knill and Pouget, 2004; Friston, 2012; Sohn and Narain, 2021). Here we will focus on the implications for dopamine. If TD learning is operating on belief states, and dopamine is reporting a TD error, then we should be able to identify signatures of belief-dependence in dopamine.

In a standard Pavlovian conditioning experiment, an initially neutral cue (such as an odor) is followed by a reward (such as water, juice, or food). The structure of the task is defined by the joint probability of the interstimulus interval (ISI), the intertrial interval (ITI), and the US delivery. We

can think of the ISI and ITI as "macro-states" that can be broken down into more temporally fine-grained "micro-states" (see subsection 5.2.4 for further discussion of time representation). If the US is always delivered following the cue, and the agent can perfectly keep track of time, then the agent has no state uncertainty—the environment is fully observable. Real animals, on the other hand, are imperfect time-keepers, as evidenced by the fact that behavioral responses to the anticipated time of US delivery become progressively more spread out in time for longer ISIs (Holland, 2000; Kirkpatrick and Church, 2000; Tsao et al., 2022). This implies that the micro-states are partially observable. If reward delivery is stochastic (i.e., it is omitted on some proportion of trials), then the macro-states also become partially observable. Both forms of partial observability affect dopamine responses. Using fixed ISIs that were fully predictable based on distinct visual cues, Fiorillo et al. (2008) found that dopamine neuron responses to reward were larger after longer ISIs (see also Starkweather et al., 2017). This is consistent with the view that rewards are more surprising after longer ISIs because temporal precision is lower—the animal has greater uncertainty about what micro-state it is in.

A quite different pattern of results is observed when the same cue is followed by randomly distributed ISIs. Fiorillo et al. (2008) reported that responses of dopamine neurons to reward delivery were *lower* following longer ISIs, consistent with an increase in reward expectation which suppressed the TD error at the time of reward. Starkweather et al. (2017) reported the same finding, showing that it could be explained by a belief state TD model without any timing noise. Importantly, Starkweather et al. also showed that another condition, in which rewards were omitted on 10% of trials, reversed this pattern: reward responses were *higher* following longer ISIs. This pattern could be explained by a belief state TD model with macro-state uncertainty. Intuitively, if the animal has not yet received reward, it's uncertain whether the reward is forthcoming or the trial has silently transitioned into the ITI. The more time that has elapsed without reward, the higher the probability that the animal is in the ITI state, and thus the greater its surprise when reward is actually delivered.

The uncertainty about the state of the environment can also arise from uncertainty about the physical state of the environment. In perceptual decision making tasks, subjects are asked to make decisions based on ambiguous sensory stimuli. In the well-known random-dot kinematogram task, animals have to identify the direction of coherent motion of a subset of dots in a background of randomly moving dots. Dopamine neurons recorded in this task were shown to exhibit the patterns expected if their activity was driven by the confidence that the decision was correct given the sensory evidence and the choice on a given trial (Lak et al., 2017).

This presence of model-based prediction errors raises an intriguing possibility. Can we flip the problem and infer the internal state of the agent and its model of the task from the dopamine activity? When inferring subjective state signals from other brain areas, the state is usually decoded from a high-dimensional neural population. Here, the state can be read out from a scalar (or low-dimensional) signal. The graded responses at sharp transitions in the value can be used to infer the state structure governing prediction error computation. For example, in the work of Starkweather et al. (2017) mentioned above, from analyzing the RPE as a function of reward delay (decreasing or increasing with reward delay), we could potentially infer the internal subjective model of the task used by the animal (100% or 90% reward probability). In recent work, Blanco-Pozo et al. (2024) use this approach to disambiguate two competing models. In most behavioral tasks, the behavioral readout is relatively low dimensional and multiple competing models could explain the observed behavioral patterns. By analyzing the dopamine responses during the task,

Blanco-Pozo and colleagues were able to distinguish the two models, supporting a model in which dopamine contributes to model-based learning without directly affecting choices.

Our brief survey covers only a subset of the empirical evidence supporting the belief state hypothesis (see also Daw et al., 2006; Babayan et al., 2018; Nour et al., 2018; Gershman and Uchida, 2019; Mikhael et al., 2022). In the remainder of this section, we want to interrogate some of our assumptions about computational architecture, in particular the division into belief computation and value approximation stages. While this division is conceptually convenient, and consistent with a wealth of data, it suffers from the curse of dimensionality that arises when the number of states grows large: for discrete hidden states, representing a belief state require a dimensionality that is exponential in the number of hidden states. For high-dimensional continuous states, belief state representation may be similarly intractable. Fortunately, it has long been recognized within AI that it is not necessary to represent all possible belief states, since only a small fraction of them are likely to be visited (Roy et al., 2005). While classical solution methods tried to find low-dimensional approximations of belief states, modern methods (e.g., Ni et al., 2022, 2024) try to avoid them entirely by learning a mapping (usually a neural network) directly from sensory inputs to value estimates. This does not, however, mean that belief states have entirely left the scene; we can often think of the learned mapping as implicitly computing a form of compressed belief state.

Hennig et al. (2023) examined this idea in detail, showing that belief states could be decoded from the internal representations of a recurrent neural network (RNN) trained end-to-end using TD learning to estimate value. The same RNN was able to recapitulate the aforementioned empirical results from the study of Starkweather et al. (2017). This suggests a different kind of computational architecture, where striatum represents the final stage of a complex value approximation architecture, which may include recurrent cortical dynamics. A similar line of argument has been made about other forms of model-based RL (Wang et al., 2018; Botvinick et al., 2019; Hattori et al., 2023). The key idea is that cortical networks can be trained to meta-learn cheap approximations of algorithms like Bayesian updating, planning, etc. which can be employed for sample-efficient RL. Striatum may act as a relatively "shallow" readout of these computations, translating them into values and action preferences.

To conclude, RL algorithms have moved from simple representations of states as time elapsed in a trial (CSC representations, Figure 2a) to more complex representations for example through POMDPs (Figure 2b) or a trained RNN (Figure 2c). In a neuroscience setting, understanding the structure of these state representations is essential to interpret dopamine signaling within the framework of RL.

5.2 What is value?

So far, all the models we've discussed share the assumption that value corresponds to expected discounted return. Modern work in AI has explored the consequences of broadening the definition of value. This line of work was motivated the observation that agents operating in complex worlds benefit from being trained on *auxiliary tasks*—intrinsic reward functions which encourage the agents to acquire more general skills and richer representations, which in turn facilitate attainment of extrinsic reward (Sutton et al., 2011; Jaderberg et al., 2016; Mirowski et al., 2016; Veeriah et al., 2019; Lyle et al., 2021). Recent evidence suggests that the brain may also learn and utilize such generalized values.

5.2.1 Distributional reinforcement learning

One way to generalize the expected return definition of value is to move from a single summary statistic (the mean) to a set of summary statistics that captures the distribution of return. A return distribution definition can distinguish between states that lead to the same expected return but with different distributions. For example, if from state A the return is always 1 while from state B the return is 102 with probability $p = 0.5$ or -100 with probability $p = 0.5$, both states will have the same expected return, but intuitively state B is riskier. A risk-averse agent would therefore avoid this state, but of course they would have to represent the return distribution in order to express such a preference. The fact that animals express both risk-aversion and risk-seeking in different circumstances suggests that they understand something about the return distribution.

This idea can be applied to *distributional RL* (Bellemare et al., 2017, 2023), where the goal is to learn in parallel the distributional statistics—typically quantiles (generalizations of the median) or expectiles (generalizations of the mean). Distributional RL algorithms have reached state-of-the-art performance on Atari games and other challenging benchmarks. Here we illustrate how the distributional statistics can be learned using generalizations of TD. By using an asymmetric learning rate depending on the sign of the TD error ($_+$ and for positive and negative prediction errors, respectively), the TD update will learn different quantiles (if TD error is binarized) or expectiles (if TD error is untransformed) depending on the ratio of the learning rates (Lowet et al., 2020). A spectrum of asymmetric learning rates spread across separate learning channels is sufficient to collectively approximate the full return distribution.

Evidence for distributional RL in dopamine neurons comes from analyzing responses to rewards whose size is randomly sampled from a distribution. The reversal point (the reward value for which the TD error switches from negative to positive) varies across neurons and scales with the ratio of the inferred learning rates for positive and negative prediction errors (Dabney et al., 2020; Lowet et al., 2020). Striatum and prefrontal cortex, where the value functions are thought to be represented, also carry representations consistent with distributional RL (Muller et al., 2024; Lowet et al., 2024).

Theoretical work has proposed different implementations of distributional RL in dopamine neurons and their projections (Tano et al., 2020; Louie, 2022), but they converge on some similar ideas. Differences in sensitivity to reward across dopamine neurons allow them to convey at the population level a learning signal that would support distributional RL in the striatum. However, many questions remain both at the algorithmic level and its possible neural implementation. A key hurdle is to understand how the dopaminergic signal, which is diffuse, can be used in specific ways by different neurons (see Liu et al., 2021, and Chapter 13). Machine learning implementations of distributional RL have strictly separated loops that compute the value functions and their associated TD error in parallel, but share a feature network. In the mammalian brain, it is hard to conceive an implementation where the loops are similarly separated (though some parallelism has been proposed; see Alexander et al., 1986; Lau et al., 2017). Instead, theoretical work suggests that the dimensionality of the learning signal might allow for the learning to take place despite overlap. Although the distributional TD error is vectorized, it is still relatively low-dimensional in comparison to the space of possible states and to the number of neurons in the target area (e.g., striatum). Similarly, although dopamine release is not constrained to a synapse, it is still relatively local. This locality, coupled with the relative dimensionality and the dendritic structure of MSNs, can support learning from a vectorized signal (Wärnberg and Kumar, 2023). Experimental evidence for this proposal is still needed, but it removes a key hurdle to the possibility that the basal

ganglia implement distributional RL.

5.2.2 Learning at multiple timescales

Another way to generalize the expected return definition is through a spectrum of discount factors. This allows agents to learn at a mixture of timescales. As mentioned in Section 2, the standard definition assumes exponential discounting, but intertemporal choice behavior appears more consistent with hyperbolic discounting, where the value of a reward t steps in the future is down-weighted by $d(t) = \frac{1}{1+kt}$, where k is a discount parameter. Mixing a spectrum of exponentially discounted value functions can create non-exponential discounting (Sozou, 1998; Kurth-Nelson and Redish, 2009).[5] Similarly to distributional RL, several value functions with different discount factors can be learned in parallel and then combined to obtain a deep RL architecture that discounts hyperbolically (Fedus et al., 2019).

The discount factor of a single dopamine neuron can be inferred by measuring the TD error response to cues predicting delay at different horizons. Experiments have discovered dopamine neurons that exhibit a diversity of discount factors (Kobayashi and Schultz, 2008; Masset et al., 2023; Sousa et al., 2023), which implies that the circuits of the basal ganglia should in principle be able to implement multi-timescale RL (Tano et al., 2020; Fedus et al., 2019). This learning at multiple timescales could be used either to adapt discounting to the statistics of the environment or to implement scale-invariance in the learning process (Howard et al., 2023; Momennejad, 2024). In addition to this diversity of timescales at the single neuron level, there appears to be a gradient in the average timescale represented across striatal areas (Tanaka et al., 2004; Enomoto et al., 2020; Mohebi et al., 2024).

5.2.3 Generalized value functions and the successor representation

The previous subsections have retained the underlying "cumulant" of the return (reward) while generalizing what kinds of predictions are made about this cumulant and how its value is represented. Another point of departure is to generalize the cumulant itself, yielding different kinds of *generalized value functions* (Sutton et al., 2011; Schlegel et al., 2021). One influential version of this idea is to replace rewards with a set of state indicator functions, $\phi_i(s) = \mathbb{I}[s = i]$. Notice that we have deliberately overloaded the basis function notation from Section 3; this invites us to think about other kinds of features as candidate cumulants (see below). The expected discounted state visitation defines a generalized value function known as the *successor representation* (SR; Dayan, 1993; Gershman, 2018):

$$M(s; i) = \mathbb{E}\left[\sum_{t=0}^{1} \gamma^t \phi_i(s_t) \middle| s_0 = s\right]: \tag{19}$$

We can think of the SR as a kind of "predictive map" that represents each state in terms of its successor states. A useful property of the SR is that state values are linear functions of M:

$$V(s) = \sum_i M(s; i)\mathbb{E}[r_j i]; \tag{20}$$

[5] This turns out to be closely related to the real-valued Laplace transform commonly used in signal processing (Fedus et al., 2019; Tano et al., 2020).

where $\mathsf{E}[rji]$ is the expected reward in state i. Intuitively, calculating value boils down to taking a sum of immediate rewards in each state weighted by how often the agent expects to visit those states in the near future. Another important property of the SR is that it obeys a Bellman equation similar to Eq. 3, which means that an agent can use TD learning to estimate it, with a separate learning channel for each successor state.

The SR can be generalized to arbitrary basis functions (not just state indicators); the columns of M are then referred to as *successor features*. This feature representation turns out to have a wide variety of applications in AI, including multi-task RL, efficient exploration, and multi-agent cooperation (see Carvalho et al., 2024, for a review). As with the SR, successor features can be estimated using TD learning. If we extend the TD error hypothesis of dopamine to this setting, then we again arrive at a vectorized conceptualization of dopamine, where (in the most straightforward implementation) each component of the error vector corresponds to a successor feature (Gardner et al., 2018).[6]

Evidence for this "generalized prediction error" hypothesis comes from studies showing activation of dopamine neurons in response to unexpected changes in sensory features. For example, Takahashi et al. (2017) showed that abrupt changes in reward identity (holding reward magnitude fixed) induced a dopamine transient. Also consistent with the hypothesis are studies showing that optogenetic manipulations of dopamine affect stimulus-stimulus learning in the absence of overt reward (Sharpe et al., 2017; Chang et al., 2017).

More broadly, the generalized prediction error hypothesis provides one way of understanding why there is so much heterogeneity in the tuning of dopamine neurons. Several proposals have been made that decompose the learning signal across dopamine neurons in terms of different targets (Figure 2b). Different TD errors signaled by dopamine neurons may be components of a multidimensional error signal (Fiorillo et al., 2013; Watabe-Uchida and Uchida, 2018; Engelhard et al., 2019; Cox and Witten, 2019). For example, there are dopamine neurons that encode prediction errors for threat (Menegas et al., 2018), action (Greenstreet et al., 2022), and social information (Solié et al., 2022). These models also lead to decision, reward and action modulated responses in dopamine neurons through distinct dopamine neurons contributing to learning either the individual features of the value approximation or decomposing the reward into subcomponents (Lee et al., 2022; Greenstreet et al., 2022; Millidge et al., 2023). Additionally, in a model-based framework, a reward-like signal can also be internally generated by the world model and information about the structure of the environment can have intrinsic value (Bromberg-Martin and Hikosaka, 2009; Bromberg-Martin and Monosov, 2020). For example, the appearance or disappearance of a barrier in a maze can lead to prediction error signal depending on how it changed the length of the path to reward (Krausz et al., 2023).

An architecture that learns several predictions in parallel (Figure 2c) will have to arbitrate between them as eventually the agent has to take one action (Doya et al., 2002). A version of this challenge is studied in multi-agent RL algorithms (MARL) in which several agents act together to achieve a common goal. In this setting, each agent has access to a partial view of the environment and algorithms differ in the amount of coordination and information available across agents (Zhang et al., 2021). This would correspond to different dopamine neurons contributing to learning through competing sub-agents evaluating different aspects of reward (or of cumulants) based

[6]Based on decoding analyses applied to populations of dopamine neurons, Stalnaker et al. (2019) have argued that sensory feature information is available at the population level but not at the single cell level. This suggests that errors might be signaled by a population code.

on different state inputs (e.g. task state, motor action, etc). This class of models has been used to explain action modulation of dopamine responses, where evaluation of reward and production of action can interfere through a contribution to the action of other sub-agents with which they do not fully share information (Lindsey and Litwin-Kumar, 2022; Cruz et al., 2022).

5.2.4 What are the features?

Within AI, most models have moved away from bespoke feature representations towards deep learning architectures which learn the feature representations end-to-end. Theoretical work has shown that the structure of the feature representations is key in controlling the dynamics of value and policy learning (Patel et al., 2023; Bordelon et al., 2024). An intriguing open question is whether these distributed TD errors are being used to drive representation learning in upstream regions, comparable to the way that auxiliary tasks are used in deep RL to guide representation learning.

Taking the example of the representation of time, the usual assumption in neuroscience is that each stimulus is decomposed into a set of temporally distributed features. For example, the aforementioned complete serial compound (Figure 2a) representation assigns each feature to a post-stimulus time bin. The stimulus can thus be thought of as activating a cascade of "time cells" (neurons tuned to particular time intervals relative to stimulus onset), as observed experimentally in a number of brain areas (Paton and Buonomano, 2018). Variations of this idea have been studied extensively, including models in which the basis functions are heterogeneous across post-stimulus time (Ludvig et al., 2008; Gershman et al., 2014) and adaptive (Mikhael and Gershman, 2019). In these models, the structure of time representations affects the predicted TD error, which in turn can alter the structure of the representations. Jakob et al. (2022) showed that trial-by-trial changes in dopamine affect time representation on subsequent trials, consistent with the representation learning hypothesis.

A similar observation is likely to govern the representations of the value (or generalized value) functions thought to be encoded in striatum (Doya, 2008). Value appears to be robustly encoded at the population level in the striatum but is distributed across single neurons (Samejima et al., 2005; Yamada et al., 2021; Shin et al., 2021; Lowet et al., 2024). Furthermore, dopamine also controls plasticity in the hippocampus (Tsetsenis et al., 2023) and many cortical areas (Macedo-Lima and Remage-Healey, 2021) whose representations also change with reward experience, hinting at the plausibility of dopamine-mediated changes in representations. Understanding how the distributed single neuron feature representations affect and are affected by diverse error signals conveyed by dopamine neurons will be an important avenue for future work.

6 Looking ahead

In this chapter, we have briefly summarized some of the key ideas behind RL, focusing on connections between recent AI algorithms and dopamine physiology. These connections aid us in understanding some of the complexities in the empirical data, while still retaining some of the core explanatory principles introduced in the original TD theory of dopamine. Our attention was still relatively narrow: we mainly discussed data from studies of dopamine projections to striatum, neglecting the rich work on dopamine signaling in prefrontal cortex (see Ott and Nieder,

2019, for a review) and its role in motor control (see Coddington and Dudman, 2019, and chapter 29).

Even within the circuits we have discussed, many challenges remain. The majority of the experimental work has either studied dopamine neurons in the midbrain, or neurons in target areas, but the lack of work integrating both leaves many question open, given the tight theoretical links between the TD error and value/policy representations. As the experimental tools progress and these experiments become more accessible, we hope to be able to start developing a more complete view of dopamine-mediated RL, including lingering questions about credit assignment (Jeong et al., 2022), motivation (Wise, 2004; Hamid et al., 2016), and attention (Kutlu et al., 2021, 2022). We expect that insights from AI will continue to be pivotal in addressing these questions.

References

Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9:357–381.

Amo, R., Matias, S., Yamanaka, A., Tanaka, K. F., Uchida, N., and Watabe-Uchida, M. (2022). A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nature Neuroscience*, 25:1082–1092.

Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., et al. (2020). What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*.

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

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 834–846.

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

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR.

Bellemare, M. G., Dabney, W., and Rowland, M. (2023). *Distributional reinforcement learning*. MIT Press.

Berns, G. S., Laibson, D., and Loewenstein, G. (2007). Intertemporal choice–toward an integrative framework. *Trends in Cognitive Sciences*, 11:482–488.

Bertsekas, D. (2019). *Reinforcement learning and optimal control*. Athena Scientific.

Blanco-Pozo, M., Akam, T., and Walton, M. E. (2024). Dopamine-independent effect of rewards on choices through hidden-state inference. *Nature Neuroscience*, pages 1–12.

Bordelon, B., Masset, P., Kuo, H., and Pehlevan, C. (2024). Loss dynamics of temporal difference reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., and Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23:408–422.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., and Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*, 107(4):603–616.

Bromberg-Martin, E. S. and Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, 63:119–126.

Bromberg-Martin, E. S. and Monosov, I. E. (2020). Neural circuitry of information seeking. *Current Opinion in Behavioral Sciences*, 35:62–70.

Carvalho, W., Tomov, M. S., de Cothi, W., Barry, C., and Gershman, S. J. (2024). Predictive representations: building blocks of intelligence. *arXiv preprint arXiv:2402.06590*.

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

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

Coddington, L. T. and Dudman, J. T. (2019). Learning from action: reconsidering movement signaling in midbrain dopamine neuron activity. *Neuron*, 104(1):63–77.

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

Cox, J. and Witten, I. B. (2019). Striatal circuits for reward learning and decision-making. *Nature Reviews Neuroscience*, 20(8):482–494.

Cruz, B. F., Guiomar, G., Soares, S., Motiwala, A., Machens, C. K., and Paton, J. J. (2022). Action suppression reveals opponent parallel control via striatal circuits. *Nature*, 607:521–526.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675.

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural computation*, 18(7):1637–1677.

Daw, N. D. and Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, 14:2567–2583.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624.

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, 11:410–416.

Doya, K., Samejima, K., Katagiri, K.-i., and Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, 14:1347–1369.

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H. J., Ornelas, S., Koay, S. A., Thiberge, S. Y., Daw, N. D., Tank, D. W., et al. (2019). Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons. *Nature*, 570(7762):509–513.

Enomoto, K., Matsumoto, N., Inokawa, H., Kimura, M., and Yamada, H. (2020). Topographic distinction in long-term value signals between presumed dopamine neurons and presumed striatal projection neurons in behaving monkeys. *Scientific Reports*, 10(1):8912.

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

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

Fedus, W., Gelada, C., Bengio, Y., Bellemare, M. G., and Larochelle, H. (2019). Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*.

Fiorillo, C. D., Newsome, W. T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nature Neuroscience*, 11:966–973.

Fiorillo, C. D., Yun, S. R., and Song, M. R. (2013). Diversity and homogeneity in responses of midbrain dopamine neurons. *Journal of Neuroscience*, 33:4693–4709.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, 17:51–72.

Freeze, B. S., Kravitz, A. V., Hammack, N., Berke, J. D., and Kreitzer, A. C. (2013). Control of basal ganglia output by direct and indirect pathway projection neurons. *Journal of Neuroscience*, 33(47):18531–18539.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, 62:1230–1233.

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

Gershman, S. J. (2018). The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, 38(33):7193–7200.

Gershman, S. J. (2024). What have we learned about artificial intelligence from studying the brain? *Biological Cybernetics*, pages 1–5.

Gershman, S. J., Moustafa, A. A., and Ludvig, E. A. (2014). Time representation in reinforcement learning models of the basal ganglia. *Frontiers in Computational Neuroscience*, 7:194.

Gershman, S. J. and Ölveczky, B. P. (2020). The neurobiology of deep reinforcement learning. *Current Biology*, 30:R629–R632.

Gershman, S. J. and Uchida, N. (2019). Believing in dopamine. *Nature Reviews Neuroscience*, 20(11):703–714.

Greenstreet, F., Vergara, H. M., Pati, S., Schwarz, L., Wisdom, M., Marbach, F., Johansson, Y., Rollik, L., Moskovitz, T., Clopath, C., et al. (2022). Action prediction error: a value-free dopaminergic teaching signal that drives stable learning. *BiorXiv*, pages 2022–09.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR.

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

Hattori, R., Hedrick, N. G., Jain, A., Chen, S., You, H., Hattori, M., Choi, J.-H., Lim, B. K., Yasuda, R., and Komiyama, T. (2023). Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience*, 26:2182–2191.

Hennig, J. A., Romero Pinto, S. A., Yamaguchi, T., Linderman, S. W., Uchida, N., and Gershman, S. J. (2023). Emergence of belief-like representations through reinforcement learning. *PLOS Computational Biology*, 19:e1011067.

Holland, P. C. (2000). Trial and intertrial durations in appetitive conditioning in rats. *Animal Learning & Behavior*, 28:121–135.

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

Howard, M. W., Esfahani, Z. G., Le, B., and Sederberg, P. B. (2023). Foundations of a temporal rl. *ArXiv*.

Ito, M. and Doya, K. (2015). Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed-and free-choice tasks. *Journal of Neuroscience*, 35:3499–3514.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*.

Jakob, A., Mikhael, J., Hamilos, A., Assad, J., and Gershman, S. (2022). Dopamine mediates the bidirectional update of interval timing. *Behavioral Neuroscience*, 136:445–452.

Jeong, H., Taylor, A., Floeder, J. R., Lohmann, M., Mihalas, S., Wu, B., Zhou, M., Burke, D. A., and Namboodiri, V. M. K. (2022). Mesolimbic dopamine release conveys causal associations. *Science*, 378:eabq6740.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, 15:535–547.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134.

Kim, H., Sul, J. H., Huh, N., Lee, D., and Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *Journal of Neuroscience*, 29:14701–14712.

Kim, H. R., Malik, A. N., Mikhael, J. G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., Gershman, S. J., et al. (2020). A unified framework for dopamine signals across timescales. *Cell*, 183(6):1600–1616.

Kirkpatrick, K. and Church, R. M. (2000). Independent effects of stimulus and cycle duration in conditioning: The role of timing processes. *Animal Learning & Behavior*, 28:373–388.

Klopf, A. H. (1972). *Brain Function and Adaptive Systems: A Heterostatic Theory*. Air Force Cambridge Research Laboratories, Air Force Systems Command.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27:712–719.

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

Krausz, T. A., Comrie, A. E., Kahn, A. E., Frank, L. M., Daw, N. D., and Berke, J. D. (2023). Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron*, 111(21):3465–3478.

Kravitz, A. V., Tye, L. D., and Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience*, 15:816–818.

Kurth-Nelson, Z. and Redish, A. D. (2009). Temporal-difference reinforcement learning with distributed representations. *PLoS One*, 4(10):e7362.

Kutlu, M. G., Zachry, J. E., Melugin, P. R., Cajigas, S. A., Chevee, M. F., Kelly, S. J., Kutlu, B., Tian, L., Siciliano, C. A., and Calipari, E. S. (2021). Dopamine release in the nucleus accumbens core signals perceived saliency. *Current Biology*, 31:4748–4761.

Kutlu, M. G., Zachry, J. E., Melugin, P. R., Tat, J., Cajigas, S., Isiktas, A. U., Patel, D. D., Siciliano, C. A., Schoenbaum, G., Sharpe, M. J., et al. (2022). Dopamine signaling in the nucleus accumbens core mediates latent inhibition. *Nature Neuroscience*, 25:1071–1081.

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

Lau, B. and Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, 58:451–463.

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

Lee, J. and Sabatini, B. L. (2021). Striatal indirect pathway mediates exploration via collicular competition. *Nature*, 599:645–649.

Lee, R. S., Engelhard, B., Witten, I. B., and Daw, N. D. (2022). A vector reward prediction error model explains dopaminergic heterogeneity. *bioRxiv*, pages 2022–02.

Lee, T. S. and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America*, 20:1434–1448.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lindsey, J. and Litwin-Kumar, A. (2022). Action-modulated midbrain dopamine activity arises from distributed control policies. *Advances in Neural Information Processing Systems*, 35:5535–5548.

Lindsey, J., Markowitz, J. E., Datta, S. R., and Litwin-Kumar, A. (2024). Dynamics of striatal action selection and reinforcement learning. *bioRxiv*.

Liu, C., Goel, P., and Kaeser, P. S. (2021). Spatial and temporal scales of dopamine transmission. *Nature Reviews Neuroscience*, 22(6):345–358.

Louie, K. (2022). Asymmetric and adaptive reward coding via normalized reinforcement learning. *PLoS computational biology*, 18(7):e1010350.

Lowet, A. S., Zheng, Q., Matias, S., Drugowitsch, J., and Uchida, N. (2020). Distributional reinforcement learning in the brain. *Trends in neurosciences*, 43:980–997.

Lowet, A. S., Zheng, Q., Meng, M., Matias, S., Drugowitsch, J., and Uchida, N. (2024). An opponent striatal circuit for distributional reinforcement learning. *bioRxiv*, pages 2024–01.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural computation*, 20(12):3034–3054.

Lyle, C., Rowland, M., Ostrovski, G., and Dabney, W. (2021). On the effect of auxiliary tasks on representation dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 1–9. PMLR.

Macedo-Lima, M. and Remage-Healey, L. (2021). Dopamine modulation of motor and sensory cortical plasticity among vertebrates. *Integrative and Comparative Biology*, 61:316–336.

Markowitz, J. E., Gillis, W. F., Beron, C. C., Neufeld, S. Q., Robertson, K., Bhagat, N. D., Peterson, R. E., Peterson, E., Hyun, M., Linderman, S. W., et al. (2018). The striatum organizes 3d behavior via moment-to-moment action selection. *Cell*, 174:44–58.

Masset, P., Tano, P., Kim, H. R., Malik, A. N., Pouget, A., and Uchida, N. (2023). Multi-timescale reinforcement learning in the brain. *bioRxiv*, pages 2023–11.

Menegas, W., Akiti, K., Amo, R., Uchida, N., and Watabe-Uchida, M. (2018). Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nature Neuroscience*, 21:1421–1430.

Mikhael, J. G. and Gershman, S. J. (2019). Adapting the flow of time with dopamine. *Journal of Neurophysiology*, 121:1748–1760.

Mikhael, J. G. and Gershman, S. J. (2022). Impulsivity and risk-seeking as Bayesian inference under dopaminergic control. *Neuropsychopharmacology*, 47:465–476.

Mikhael, J. G., Kim, H. R., Uchida, N., and Gershman, S. J. (2022). The role of state uncertainty in the dynamics of dopamine. *Current Biology*, 32(5):1077–1087.

Millidge, B. G., Song, Y., Lak, A., Walton, M. E., and Bogacz, R. (2023). Reward-bases: dopaminergic mechanisms for adaptive acquisition of multiple reward types. *BioRxiv*, pages 2023–05.

Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J. W., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., Nazi, A., et al. (2021). A graph placement methodology for fast chip design. *Nature*, 594:207–212.

Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., et al. (2016). Learning to navigate in complex environments. In *International Conference on Learning Representations*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Mohebi, A., Wei, W., Pelattini, L., Kim, K., and Berke, J. D. (2024). Dopamine transients follow a striatal gradient of reward time horizons. *Nature Neuroscience*, pages 1–10.

Möller, M. and Bogacz, R. (2019). Learning the payoffs and costs of actions. *PLoS Computational Biology*, 15:e1006285.

Momennejad, I. (2024). Memory, space, and planning: Multiscale predictive representations. *arXiv preprint arXiv:2401.09491*.

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

Muller, T. H., Butler, J. L., Veselic, S., Miranda, B., Wallis, J. D., Dayan, P., Behrens, T. E., Kurth-Nelson, Z., and Kennerley, S. W. (2024). Distributional reinforcement learning in prefrontal cortex. *Nature Neuroscience*, pages 1–6.

Ni, T., Eysenbach, B., and Salakhutdinov, R. (2022). Recurrent model-free RL can be a strong baseline for many pomdps. In *International Conference on Machine Learning*, pages 16691–16723. PMLR.

Ni, T., Eysenbach, B., Seyedsalehi, E., Ma, M., Gehring, C., Mahajan, A., and Bacon, P.-L. (2024). Bridging state and history representations: Understanding self-predictive rl. *arXiv preprint arXiv:2401.08898*.

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

Nour, M. M., Dahoun, T., Schwartenbeck, P., Adams, R. A., FitzGerald, T. H., Coello, C., Wall, M. B., Dolan, R. J., and Howes, O. D. (2018). Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. *Proceedings of the National Academy of Sciences*, 115(43):E10167–E10176.

Ott, T. and Nieder, A. (2019). Dopamine and cognitive control in prefrontal cortex. *Trends in Cognitive Sciences*, 23:213–234.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., Gross, C. E., and Boraud, T. (2007). Shaping of motor responses by incentive values through the basal ganglia. *Journal of Neuroscience*, 27(5):1176–1183.

Patel, N., Lee, S., Mannelli, S. S., Goldt, S., and Saxe, A. M. (2023). The RL perceptron: Dynamics of policy learning in high dimensions. In *ICLR 2023 Workshop on Physics for Machine Learning*.

Paton, J. J. and Buonomano, D. V. (2018). The neural basis of timing: distributed mechanisms for diverse functions. *Neuron*, 98:687–705.

Pavlov, I. (1927). *Conditioned Reflexes*. Oxford University Press.

Perrin, E. and Venance, L. (2019). Bridging the gap between striatal plasticity and learning. *Current Opinion in Neurobiology*, 54:104–112.

Pinto, S. R. and Uchida, N. (2023). Tonic dopamine and biases in value learning linked through a biologically inspired reinforcement learning model. *bioRxiv*.

Rao, R. P. (2010). Decision making under uncertainty: a neural model based on partially observable markov decision processes. *Frontiers in Computational Neuroscience*, 4:146.

Reynolds, J. N. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15:507–521.

Roy, N., Gordon, G., and Thrun, S. (2005). Finding approximate POMDP solutions through belief compression. *Journal of Artificial Intelligence Research*, 23:1–40.

Salinas-Hernández, X. I., Vogel, P., Betz, S., Kalisch, R., Sigurdsson, T., and Duvarci, S. (2018). Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes. *Elife*, 7:e38818.

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, 310(5752):1337–1340.

Schlegel, M., Jacobsen, A., Abbas, Z., Patterson, A., White, A., and White, M. (2021). General value function networks. *Journal of Artificial Intelligence Research*, 70:497–543.

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

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., Niv, Y., and Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature neuroscience*, 20(5):735–742.

Shen, W., Flajolet, M., Greengard, P., and Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, 321:848–851.

Shin, E. J., Jang, Y., Kim, S., Kim, H., Cai, X., Lee, H., Sul, J. H., Lee, S.-H., Chung, Y., Lee, D., et al. (2021). Robust and distributed neural representation of action values. *Elife*, 10:e53045.

Sippy, T. and Tritsch, N. X. (2023). Unraveling the dynamics of dopamine release and its actions on target cells. *Trends in Neurosciences*, 46:228–239.

Skinner, B. (1938). *The Behavior of Organisms*. Appleton-Century.

Sohn, H. and Narain, D. (2021). Neural implementations of Bayesian inference. *Current Opinion in Neurobiology*, 70:121–129.

Solié, C., Girard, B., Righetti, B., Tapparel, M., and Bellone, C. (2022). VTA dopamine neuron activity encodes social interaction and promotes reinforcement learning through social prediction error. *Nature Neuroscience*, 25:86–97.

Sousa, M., Bujalski, P., Cruz, B. F., Louie, K., McNamee, D., and Paton, J. J. (2023). Dopamine neurons encode a multidimensional probabilistic map of future reward. *bioRxiv*, pages 2023–11.

Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1409):2015–2020.

Stalnaker, T. A., Howard, J. D., Takahashi, Y. K., Gershman, S. J., Kahnt, T., and Schoenbaum, G. (2019). Dopamine neuron ensembles signal the content of sensory prediction errors. *Elife*, 8:e49315.

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

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

Surmeier, D. J., Ding, J., Day, M., Wang, Z., and Shen, W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences*, 30:228–235.

Sutton, R. and Barto, A. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, 88:135–170.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., and Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 761–768.

Szepesvári, C. (2022). *Algorithms for Reinforcement Learning*. Springer Nature.

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

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7:887–893.

Tano, P., Dayan, P., and Pouget, A. (2020). A local temporal difference code for distributional reinforcement learning. *Advances in Neural Information Processing systems*, 33:13662–13673.

Thorndike, E. L. (1898). *Animal Intelligence*. The Macmillan Company.

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

Tsao, A., Yousefzadeh, S. A., Meck, W. H., Moser, M.-B., and Moser, E. I. (2022). The neural bases for timing of durations. *Nature Reviews Neuroscience*, 23:646–665.

Tsetsenis, T., Broussard, J. I., and Dani, J. A. (2023). Dopaminergic regulation of hippocampal plasticity, learning, and memory. *Frontiers in Behavioral Neuroscience*, 16:1092420.

Tsutsui-Kimura, I., Matsumoto, H., Akiti, K., Yamada, M. M., Uchida, N., and Watabe-Uchida, M. (2020). Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. *Elife*, 9:e62390.

van Elzelingen, W., Goedhoop, J., Warnaar, P., Denys, D., Arbab, T., and Willuhn, I. (2022). A unidirectional but not uniform striatal landscape of dopamine signaling for motivational stimuli. *Proceedings of the National Academy of Sciences*, 119:e2117270119.

Vanderveldt, A., Oliveira, L., and Green, L. (2016). Delay discounting: Pigeon, rat, human—does it matter? *Journal of Experimental Psychology: Animal learning and cognition*, 42(2):141.

Veeriah, V., Hessel, M., Xu, Z., Rajendran, J., Lewis, R. L., Oh, J., van Hasselt, H. P., Silver, D., and Singh, S. (2019). Discovery of useful questions as auxiliary tasks. *Advances in Neural Information Processing Systems*, 32.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21:860–868.

Wärnberg, E. and Kumar, A. (2023). Feasibility of dopamine as a vector-valued feedback signal in the basal ganglia. *Proceedings of the National Academy of Sciences*, 120(32):e2221994120.

Watabe-Uchida, M. and Uchida, N. (2018). Multiple dopamine systems: weal and woe of dopamine. In *Cold Spring Harbor symposia on quantitative biology*, volume 83, pages 83–95. Cold Spring Harbor Laboratory Press.

Widrow, B., Gupta, N. K., and Maitra, S. (1973). Punish/reward: Learning with a critic in adaptive threshold systems. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 455–465.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, 5:483–494.

Xie, Y., Huang, L., Corona, A., Pagliaro, A. H., and Shea, S. D. (2023). A dopaminergic reward prediction error signal shapes maternal behavior in mice. *Neuron*, 111:557–570.

Yamada, H., Imaizumi, Y., and Matsumoto, M. (2021). Neural population dynamics underlying expected value computation. *Journal of Neuroscience*, 41(8):1684–1698.

Zhang, K., Yang, Z., and Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384.