Temporal Difference Learning Is Not All You Need

Huy Ha
huy.ha@columbia.edu

Sian Lee Kitt
sll2190@columbia.edu

Abstract: For decades, temporal-difference learning has emerged as the dominant computational framework for understanding dopamine's role in decision making in biological systems. Under this framework, dopamine response signals a reward prediction error for a scalar value prediction under a state, and drives learning in a model-free decision making paradigm. However, recent empirical evidence from biological systems have discovered discrepancies from the predictions made by TD learning. In this paper, we review works which reveal the role of dopamine in probabilistic inference and in model-based planning, which the classical scalar-value model-free TD learning perspective does not capture.

1 Introduction

The seminal works connecting dopamine neurons in biological systems to Temporal-Difference (TD) learning [1, 2, 3, 4] stipulated that dopamine served as a signal for a reward prediction error (RPE) which drives learning in biological systems. In this framework, learning occurs in the value function, whose scalar value predictions for states are updated to minimize RPEs, and the policy used for action selection is computed to maximize the expected future discounted returns under the value function. TD learning is a model-free algorithm, in that action selection does not involve any model of the transition dynamics or reward functions used for forward simulation of actions from the current state. Though this perspective has useful in explaining an immense amount of biological data, recent works have discovered the multiplicity and multi-faceted nature of decision making in the brain, some of which are unaccounted for in the classical TD learning perspective of dopamine's role.

In this paper, we review the TD learning perspective on dopamine and the predictions it makes about the behavior of its agents. Then, we present works that suggest dopamine's central role in model-based reinforcement learning, which is entirely unaccounted for in the model-free TD learning perspective. Next, we investigate works that argue for a decision-making framework that explicitly accounts for the uncertainty of acting in a complex and partially observable environment. After motivating the need for reasoning about uncertainty in the state, value, and policy, we explain dopamine's role in the probabilistic inference procedures required for acting optimally under such sources of uncertainty. Finally, we outline some open problems surrounding the role of dopamine, decision making, and reinforcement learning in the brain.

2 Background

Reinforcement learning is a computational framework for solving sequential decision-making tasks where an agent interacts with an environment to maximize a scalar reward signal. In this section, we review two mathematical frameworks that formalize the reinforcement learning task, Markov decision processes, and partially observable Markov decision processes. These frameworks introduces notation which is used for explaining Temporal-Difference learning (section 2.4) and dopamine's role in probabilistic computation (section 4.1). We will also introduce model-based reinforcement learning and compare it to model-free reinforcement learning, which is used for understanding dopamine's involvement in both of these decision making paradigms in the brain (section 3).

2.1 Markov Decision Processes

Formally, the task is described by a MDP $\mathcal{M}$ defined by the tuple $(\mathcal{S},\mathcal{A},T,\mathcal{R},\gamma)$. At timestep t, the agent observes states $s_t \in \mathcal{S}$, takes actions $a_t \in \mathcal{A}$ which leads to the next state $s_{t+1} \in \mathcal{S}$ according to the

transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, then observes reward r_t from the reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The goal of the RL is to learn the agent's policy function $\pi : \mathcal{S} \to \mathcal{A}$ in order to maximize the expected returns $E[\sum_{t=0}^{T} \gamma^t r_{t+1}]$. This framework also allows for stochastic dynamics by expressing the transition function as a distribution over next states $T(s_{t+1}|s_t,a_t) \in [0,1]$ and stochastic policies $\pi(a_t|s_t) \in [0,1]$.

2.2 Partially-Observable Markov Decision Processes

At the core of convergence guarantees of RL algorithms like Q learning[5, 6] is the Markovian assumption, also known as the memoryless-ness property, which asserts that the next state s_{t+1} is independent of all states before the current state s_t when conditioned on its current state s_t. However, when the states are not fully-observable, the Markovian assumption and thus the convergence for RL algorithms that rely on the assumption no longer holds.

Partially-observable Markov decision processes (POMDP) generalizes MDPs with the set of observations Ω and conditional observation probabilities O, such that in the state $s_t \in \mathcal{S}$ the agent receives an observation $o_{t+1} \in \Omega$ sampled from $O(o_{t+1}|s_t)$. From the perspective of the agent in a POMDP, its world is no longer Markovian from just its observations. However, for every POMDP $(\mathcal{S},\mathcal{A},T,\mathcal{R},\Omega,O,\gamma)$, one can formulate a MDP $(\mathcal{B},\mathcal{A},T',\mathcal{R},\gamma)$. Here, every belief state $b \in \mathcal{B}$ is a *distribution* over all states $s \in \mathcal{S}$ from the POMDP, and the agent updates its belief state with the belief transition function $T'(b_{t+1}|b_t,a_t)$.

This belief MDP is useful because though the world is not Markovian from the agent's observations in the original POMDP, it is Markovian in the agent's belief states in the belief MDP. Specifically, computing $T'(b_{t+1}|b_t, a_t)$ involves the posterior over states conditioned on an agent's previous interactions with the POMDP

$$b(s) = P(s|o) \propto O(o|s)P(s) \tag{1}$$

where $P(s)$ is the prior over states. Because the agent updates its beliefs optimally using Bayes rule, all its past interactions are encoded into its belief state, giving the belief MDP the memoryless-ness property.

2.3 Model-based versus model-free reinforcement learning

It's helpful to classify RL algorithms into model-based and model-free approaches (though recent algorithms have blurred the lines between these two categories [7]). Model-based algorithms learn a model of the environment, which is usually some approximation of the transition function T and rewards function $\mathcal{R}$, and uses this model for online planning with algorithms like dynamic programming (a.k.a Value iteration) or Monte-Carlo Tree Search [8]. Though using this model requires expensive online computation, it is more flexible and generalizes better. In contrast, model-free algorithms forgo learning a model of the world, and instead directly estimates the scalar returns with a value function $V(s)$ from states s, then uses this value function to compute a policy. To learn this value function $V(s)$, one typical uses a variant of Temporal-Difference learning (See section 2.4) to minimize the Bellman error.

$$V(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} P(s,a,s')V(s') \right) \tag{2}$$

In contrast to model-based algorithms, the policy is explicitly represented in model-free approaches, which reduces online computation costs. However, for any small changes in the dynamics of the environment, model-free approaches require significant experience to adapt, since the recursive nature of Bellman updates prevents any local change to the policy to be isolated from the rest of the policy. This makes model-free approaches such as TD learning less flexible than model-based approaches.

2.4 Dopamine and TD learning

Dopamine neurons of the ventral tegmental area (VTA) and substantia nigra have long been associated with the processing of reward stimuli [2, 3, 4]. In the experiment conducted by Schultz et al., researchers train a monkey to immediately reach for a lever when a light is illuminated. Once the lever is pulled the monkey is given fruit juice as a reward. The experiment can be considered as three phases. In the pre-training phase (when the monkey does not expect to see the reward), most dopamine neurons showed a spike in activity when the reward was delivered. After training, the primary reward no longer elicits a phasic response, and the onset of the predictive light (cue) now causes a phasic activation in dopamine cell output. In post-training trials where the reward is not delivered immediately after

the cue, dopamine neurons are depressed below their basal firing rate exactly at the time that the reward should have been delivered [10]. This well-timed decrease in spike output shows that the expected time of reward delivery based on the occurrence of the light is encoded in the fluctuations in dopamine activity as shown in Figure 1.

Dopamine neurons, therefore, are reliable detectors that emit a positive prediction error via an increased spike in dopamine if the experienced event is better than predicted, no signal if the event occurs as predicted, and a negative prediction error via a depressed spike in dopamine if an experienced event is worse than predicted. From this experiment, we begin to see connections to theories in computation established by Sutton [12] which suggests that dopamine neurons do not only report the occurrence of experienced events but also encode a deviation or error between the actual reward received and predictions of the time and magnitude of reward [10].

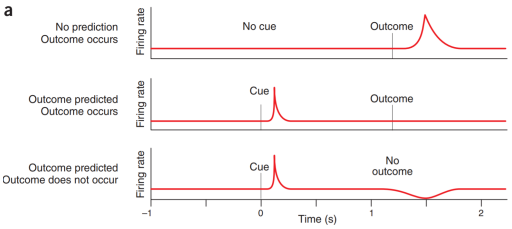

Figure 1: Dopaminergic Prediction Error. Dopamine neurons fire on unpredicted outcomes (top), predicted cues (middle and bottom, left), but not to predicted outcomes (middle, right). When expected outcomes do not occur (bottom, right), firing is suppressed. Figure taken from [11]

The TD reward prediction error (RPE) at time t is given by

$$\delta(s_t) = r(s_t) + \gamma V(s_{t+1}) - V(s_t) \tag{3}$$

The TD algorithm [13] has a provided a foundational framework for interpreting the activity of dopamine neurons. A core property of TD, supported by the temporally-precise responses of dopamine, is that it is model-free. The learned state values are an aggregated scalar representation of total future expected rewards and these scalar state values are learned through experience and saved for future use.

There are two main assumptions in TD learning [10, 1]. First, sensory cues can be used to predict a discounted sum of all future rewards $V(t)$

$$V(t) = \mathbb{E}[\gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + ...]$$

where $r(t)$ is the reward at time t and $\mathbb{E}[\cdot]$ denotes the expected value of the sum of future rewards up to the end of the trial. $\gamma \in [0,1]$ is a discount factor that makes rewards that arrive sooner more important than rewards that arrive later. Second, the presentation of future sensory cues and rewards depends only on the current cues and not previous sensory cues (i.e. Markov process).

TD provides a simple way to learn $V(t)$ using prediction errors. Consider the case study of the monkey and fruit juice once again. What happens on a trial when the light is followed by the reward after a shorter delay? As illustrated in figure 2, TD suggests that the cue triggers a discrete sequence of activity that represents sequential time points after the cue. At each time point, the summation of this weighted representation produces a scalar estimate of future value V, which dopamine neurons (DA in figure 2) compare to the obtained reward to compute a prediction error signal. The prediction error is then broadcasted widely (red) and used to modify the weights for neurons that were recently active (circles on arrows). When an unexpectedly early reward is delivered, the prediction error signals the difference in time-discounted value and modifies the weights for the part of the representation that is active when the prediction error is signaled.

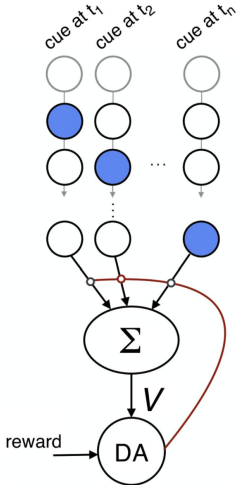

Figure 2: Temporal difference algorithm. Figure taken from [14]

3 Dopamine and model-based reinforcement learning

As defined by Sutton and Barto [15], model-based reinforcement learning are goal-directed algorithms with a planning component at its core. Here, planning means using an explicit representation of the MDP or a model of the world (i.e: transition and reward functions) to predict outcomes (i.e: future states and rewards) of various action sequences in action selection to reach a goal. Compared to model-free reinforcement learning, which requires many visits to states and exploration of actions before the

information is reflected in its explicit representation of the value and policy function, model-based algorithms are more sample-efficient. However, without an explicit representation of the policy, model-based approaches must perform extra computation at test time in using planning algorithms on its models. In this section, we review model-based learning in biological systems and dopamine's role in such learning processes.

3.1 Model-based learning in biological systems

Since Köhler's 1921 experiments [16], where chimpanzees could stack crates and use long poles get bananas that were initially out of reach, we have suspected that some form of model-based learning exists in non-human animals as well. This is because the chimps have never associated stacked crates with high value, and thus would never do so if it were using a model-free learning approach. In contrast, their physics intuition of objects and gravity, among other things, provides chimps with a model of the world, which they could use to forward simulate actions and states (i.e: plan) and "creatively" compose novel action sequences to achieve their goal.

3.2 Model-based or model-free?

From just observing the behavior of an agent making decisions to maximize its rewards, how can we tell whether it's learning and using a model for planning, as opposed to just learning values associated with states and actions?

A common approach to decipher between model-based and model-free learning is called sensory pre-conditioning and involves three stages. First, in a pre-conditioning phase lasting a few days, the animal learns a model of its environments in absence of any reward. This amounts to showing the animal the transitions that event A leads to event B repeatedly, where each event is implemented by presenting visual and auditory cues. Second, in a conditioning phase lasting another few days, the animal is conditioned into a high value for the event B, which just means flashing the lights and tones that correspond to event B, then giving the animal a dosage of flavored milk. Third, in the testing phase, a single probe test is executed, which means the animal is presented with event A and its behavior is observed. If the animal is using a model-based learning approach, then they should exhibit behavior that demonstrates an expectation of the rewards, since they know A leads to B leads to reward. In contrast, in the control events C and D, which also go through all three phases of sensory pre-conditioning, event C should not trigger such behaviors.

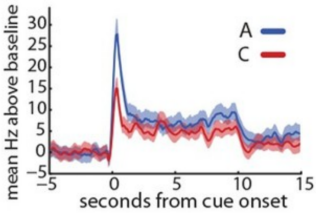

Figure 3: Dopamine neurons spiking to pre-conditioned cues, inconsistent with TD-learning predictions. Figures taken from [17]

If dopamine is only involved in TD-learning mechanisms in the brain, we do not expect the animals to display a dopamine spike at A, because it has never be associated with high value by itself. However, Sadacca et al. [17] observe that dopamine neurons in rats actual do spike to A, significantly more so than the baseline C (Fig. 3). Could it be that the rats are storing state, action, reward pairs in an experience replay buffer [18], and updating its learned value for A in between the conditioning and testing phase? Indeed, though TD-learning does not provide any mechanism for values to transfer between predictive cues after their experience, the experience replay buffer seems like a simple enough extension to incorporate this empirical result into its repertoire of behavior predictions. However, as we will see in the next section, this does not seem to be the end of the story.

3.3 Dopamine is necessary and sufficient for model-based associations

Now that we've developed the necessary background, let's investigate dopamine's role in model-based learning in biological systems. In this section, we will discuss how dopamine could "reactivate" model-based associative learning when such learning has been blocked, showing that dopamine is sufficient for this learning process. Then, we will explain how inhibiting dopamine neurons during pre-condition prevents associative model-learning, showing that dopamine is also necessary.

An animal's learning of the transition of event X to rewarding event B ($X \rightarrow B$) can be "blocked" if the animal has already learned that event A leads to event B ($A \rightarrow B$) and always presenting X to the animal in the presence of A ($AX \rightarrow B$). Indeed, this is a hallmark behavior of model-based learning,

because if event A already reliably predicts B and its rewards, then there is no need for learning event X. This is called blocking [19], a technique to show that reward prediction plays an important role in model-based associative learning.

Using blocking, we can create a condition where model-learning is absent, then optogenetically control dopamine spikes and see if the model-learning ability is regained. First, animals are preconditioned to learn $A \rightarrow B$, where the animal was rewarded sucrose pellets in B. Second, Sharpe et al. [20] optogenetically activates dopamine neurons at the time of B in the transition $AX \rightarrow B$, while doing nothing for $AY \rightarrow B$ such that learning of Y is blocked as usual, and the animal continues to get rewarded pellets in B. Finally, in the testing phase, a probe test can be done by presenting X and Y alone. The authors found that in the former case, activating dopamine neurons was sufficient to "unblock" the learning of X because the animals displayed reward expectation behavior (i.e: checking the food cup) more frequently than the baseline level, while Y was blocked as usual. What did the animal learn about X? To investigate whether X acquired value by itself or only predicts the outcome B, the authors use devaluation of rewards [21]. This amounts to presenting the animals with B and giving them the same sucrose pellets but injected with lithium chloride to induce nausea, such that the food reward associated with B has a low value. Then, with a probe test on X, the authors found that the rats were less likely to check the food cup. This suggests that optogenetically activating dopamine neurons unblocked learning that X predicts the specific sucrose pellets associated with B, and thus dopamine is sufficient for associative model learning.

How can we test whether dopamine is necessary for associative model learning? Sharpe et al. proposed to optogenetically inhibit dopamine neurons during event transitions, then seeing if the animal could still learn the transitions. Specifically, first, in the pre-conditioning phase, the transitions $A \rightarrow B$ and $C \rightarrow D$ are presented to the animal in absence of external rewards. In $A \rightarrow B$, dopamine neurons are inhibited, while $C \rightarrow D$ serves as the control. Second, in the conditioning phase, the animal is rewarded in events B and D, allowing it to associate rewards with B and D. Finally, in the testing phase, the probe tests on A and C, and any differences in the animals' behavior will tell us our answer. The authors found that while the animals in the control study displayed high rates of conditioned responding (see Section 3.2) as expected, inhibiting dopamine neurons resulted in a significant reduction in conditioned responding, suggesting that the transition $A \rightarrow B$ was not learned. This suggests that dopamine is also necessary for associative model learning, in that inhibiting dopamine neuron firing prevents it.

3.4 A place for model-based learning in a model-free world

The findings discussed in section 3.3 is quite surprising. For decades, dopamine has almost come analogous to RPE and TD-learning, yet here, empirical results show that this neuro-modulator is heavily involved in model-based learning mechanisms. Since the model-based and model-free learning paradigms are so different, could it be that dopamine is involved in both?

Killcross and Coutureau [22] shows that perhaps these two modes of decision making could be working together instead of in opposition. In the first stage, rats are trained to press a lever to get a high-value food reward. In the second stage, the food reward is devalued separately from the lever. The model-based approach will predict a corresponding decrease in the rate of conditioned responding since the lever leads to devalued foods. In contrast, the model-free approach will predict the same rate of conditioned responding, since the lever resulted in high-value rewards the last time the animal pressed the lever. Killcross and Coutureau found that both types of behaviors were present in rats, depending on how much training the animal received. Specifically, in the restricted training regime, the rates of conditioned responses significantly decreased after devaluation. However, after overtraining on the high-value-reward lever pushing task, the rats remained an equally high tendency to push the lever after the devaluation, suggesting that their behavior had become habits that are automatic, involuntary, or impulsive.

On the other hand, results from deep reinforcement learning research suggest that model-free learning could also underlie model-based learning like behavior. Here, rather than following Sutton and Barto's definition of model-based reinforcement learning which requires an explicit model in the planning, one could take a behaviorist approach that focuses not on the agent's inner workings but instead on the properties of the agent's decision making which appears to an outside observer as if the agent was using a model-based algorithm for action selection. Specifically, Guez et al. [23] proposed to represent the policy with generic high capacity function approximation architectures, such as convolutional neural networks and long-short term memory networks, trained end-to-end in a purely model-free

fashion. After training, the authors evaluate the policy and discovered that the agent demonstrated good generalization on combinatorial domains (i.e: procedural environments), sample efficiency, and performance improvement when allowed more thinking time that was tell-tale signs of model-based reinforcement learning algorithms. While there is still work in understanding how such properties could arise from a purely model-free algorithm, it will be interesting to see whether scaling up model-free learning in biological systems also give rise to model-based planning like behavior, as it could explain how dopamine takes part in both model-based and model-free systems at the same time.

Research in deep reinforcement learning has observed that model-based algorithms have superior empirical sample-efficiency over model-free approaches [24] and thus has been preferred for applications where data collection is expensive or impossible such as offline reinforcement learning [25, 26]. While some works that combine model-based and model-free approaches exist [7, 27], but none display decision-making behavior which transitions from model-based to model-free as the amount of training data as in [22]. Perhaps this motivates more work on algorithms that integrate these two paradigms for decision making that has been long recognized in biological agents [28] into artificial agents as well.

4 Probabilistic Computation with Dopamine

Though the classical formulation of RL from Section 2.1 has been fruitful for progress in solving toy problems, its formulation includes assumptions that fail to extend to the complexities of RL in the real world in three ways. First, the classical formulation assumes full observation of states, when in reality, agents operate on noisy, partial, and ambiguous observations that are only correlated with the ground truth underlying state. Second, exactly solving an MDP requires learning the exact value for each state, but for large state/observation spaces (i.e: combinatorial, continuous), agents with bounded space and time can't afford to memorize the value for all states. Thus, the approximation of the value function leads to uncertainty in state value estimation, which trickles down to uncertainty in computing the optimal policy. Third, and intricately linked to the last point, agents must balance between actions that lead to high value under its value function and actions that provide it with new potentially useful information about its environment. This problem is known as the exploration-exploitation problem, and agents must also reason about this action selection uncertainty.

These three challenges all point towards a probabilistic framework for reinforcement learning. While the TD learning perspective for understanding the role of dopamine has explained lots of empirical observations, we are beginning to see more experimental data that are unaccounted for under this perspective in isolation. In this section, we review empirical evidence for dopamine's involvement in probabilistic computations involve state uncertainty, value uncertainty, and policy uncertainty in biological systems.

4.1 State uncertainty and belief states

From section 2.2, we know that agents must operate in belief states instead of observations when faced with a task under partially-observability. These two different modes of operations give rise to different characteristic behavior and RPE predictions, which can be exploited to understand which mode an agent is working in.

Babayan et al. [29] designed a Pavlovian conditioning task for mice, where two different states s_1 and s_2 with low and high rewards respectively leads to the same observation (an odor). Thus, from just the partial observation, the mice need to infer which of the two states they are in and make predictions about the reward it expects to receive accordingly.

Without loss of generality, assume state s_1 yields a reward of 0, and s_2 yields reward 1. If the animal operates on only the partial observation o (which is identical in both states) and both states are experienced evenly, then its value prediction $V(o)$ will be the expectation of reward over its experience

$$V(o) = P(s = s_1)r(s_1) + P(s = s_2)r(s_2) = 0.5$$

Since its value prediction is a constant value, we expect to see its RPE response (equation 3) to be linear with respect to the actual reward received (figure 4, left).

However, if the animal is uncertain about which of the two states it is in, it would perform state inference by expressing its uncertainty as a belief state $b(s)$, a distribution over states. The posterior over the

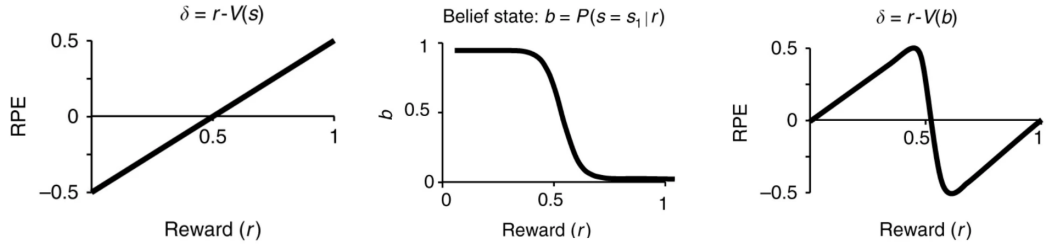

Figure 4: Operating on partial observations will predict a constant value, giving a monotonically increasing RPE as a function of actual rewards (left). However, operating on belief states will weight reward from a state by the probability of being that state (center), giving value predictions a non-monotic pattern (right) Figure taken from [29]

latent state s given its observed rewards r is computed using Bayes' rule

$$b(s) = P(s|r) = \frac{P(r|s)P(s)}{P(r)} \tag{4}$$

From this belief state, its value prediction $V(b)$ is a function of its belief state b, given by $V(b) = b(s_1)r(s_1) + b(s_2)r(s_2)$. Staring at equation 4, this would mean for observed rewards less than 0.5, the animal's belief would assign more mass to state s_1, and its value prediction $V(b)$ would be closer to 0. Similarly, for observed rewards greater than 0.5, the chances of being in s_2 are much higher, and its value prediction $V(b)$ would be closer to 1. Thus, if the animal is operating in belief states, we expect to see a non-monotonic pattern for its RPE as a function of reward received (figure 4, right).

After training the mice on states s_1 and s_2 with liquid rewards of $1\mu L$ and $10\mu L$ respectively, the authors measured their reward expectation with anticipatory licking rate and dopamine responses. To sample the mice's RPE function, they introduced rare intermediate rewards at intervals between $1\mu L$ and $10\mu L$. Their results showed a non-monotonic RPE curve that closely followed the predicted RPE curve which uses state inference, supporting the hypothesis that under partial observability, mice operate on belief states rather than partial observations.

If biological systems do indeed perform state inference when tasks call for it, then the TD learning update rule in equation 3 is incomplete. Instead, from a belief state RPE $\delta(b_t)$ at timestep t will provide a learning signal for all states $s \in \mathcal{S}$, scaled in proportion to their probability under b_t.

4.2 Reward uncertainty in backward-blocking

Consider an extension of the blocking problem (see section 3.3) called backward-blocking, where the order of A+ (+ being the reward) and AB+ phases are reversed. During AB+ training, there should be a positive error to drive learning about B, since A has not yet been paired with the reward on its own. Surprisingly training with A+ causes an association between B and the reward, which is not captured under the standard TD model. Gershman and Uchida [30] explain this association with a Bayesian treatment of the TD model called Kalman TD Learning [31].

In the experiment conducted by Gershman and Uchida the first phase consists of preconditioning the animal by pairing A serially with B. Since no reward is delivered the standard TD model does not predict any learning. However, the Kalman TD model will learn a positive covariance between A and B since the offset of A is associated with the onset of B. In the second phase, B is paired with the reward and in the third phase, the animal's response to A is probed. It was observed that the animal showed a conditioned response to A, even though A was never paired with the reward. The dopamine neurons respond to A more than to the control stimulus in the preconditioning phase. These findings are consistent with the Kalman TD model.

4.3 Exploration-exploitation under dopamine

The standard TD-model also fails to capture the exploration vs exploitation dilemma. Gershman and Uchida was inspired by the biological results which linked variation in prefrontal dopamine levels with

directed exploration and variation in striatal dopamine levels with random exploration Humphries et al. [32]. Therefore, they propose treating the policy as a latent variable that can be inferred conditional on a goal-oriented objective such as maximizing cumulative rewards. This methodology asserts that the dopamine response can be interpreted as estimated precision (inverted variance) of an inferred policy instead of RPEs. The precision corresponds to the agent's confidence that the policy it is currently following is optimal. Therefore, when an agent is in a certain state it will continue to exploit actions in this state if the dopamine response is positive since this is indicative of low variance and low uncertainty of the current policy. However, the agent will choose to explore other states if the dopamine gives no signal or the RPE becomes negative since this is indicative that the current policy is no longer optimal.

5 Discussion and Open Challenges

We have reviewed dopamine's role in decision-making processes in the brain which is not explained by the classical TD learning perspective. In addition to model-based learning and probabilistic computations, there are other open challenges of reconciling empirical results of dopamine in the brain with the reinforcement learning framework provided by artificial intelligence research.

For instance, artificial reinforcement learning agents operate on discrete-time steps, but how time is represented in biological agent's decision making processes is not well understand. Indeed, we know the predictive power of TD learning is significantly affected by timing noise. It is a complex task to differentiate whether neural prediction error signals correspond with the high degree of noise in behavioral timing [33]. Other works have also observed that the identity of the reward (as opposed to only the magnitude of the reward) also affects dopamine spikes in the brain, but reward in artificial intelligence research disregards reward source. Another example observes that dopamine is only involved in appetitive outcomes leading to positive rewards [34, 35, 36], but the neural mechanisms in the brain for aversive outcomes leading to negative rewards is still unknown. As we hinted in section 3.4, while we know that there are at least two reinforcement learning systems in the brain that makes different assumptions, computations, and thus gives rise to different decision-making behavior. We do not know how many other reinforcement learning algorithms might also be present in the brain, and whether they could be unified under one single master algorithm.

References

[1] P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, 16(5):1936–1947, 1996.

[2] R. A. Wise. Dopamine and reward: the anhedonia hypothesis 30 years on. *Neurotoxicity research*, 14(2-3):169–183, 2008.

[3] R. Romo and W. Schultz. Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *Journal of neurophysiology*, 63(3):592–606, 1990.

[4] W. Schultz, P. Apicella, E. Scarnati, and T. Ljungberg. Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of neuroscience*, 12(12):4595–4610, 1992.

[5] C. J. C. H. Watkins. Learning from delayed rewards. 1989.

[6] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[7] D. Hafner, T. P. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. *CoRR*, abs/1912.01603, 2019. URL http://arxiv.org/abs/1912.01603.

[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[9] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997. ISSN 00368075, 10959203. URL http://www.jstor.org/stable/2893707.

[10] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[11] G. Schoenbaum, G. R. Esber, and M. D. Iordanova. Dopamine signals mimic reward prediction errors. *Nature neuroscience*, 16(7):777–779, 2013.

[12] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1): 9–44, 1988.

[13] E. A. Ludvig, R. S. Sutton, and E. J. Kehoe. Evaluating the td model of classical conditioning. *Learning & behavior*, 40(3):305–319, 2012.

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

[15] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[16] W. Köhler. *Intelligenzprüfungen an Menschenaffen: Mit einem Anhang: Zur Psychologie des Schimpansen*, volume 134. Springer-Verlag, 2013.

[17] B. F. Sadacca, J. L. Jones, and G. Schoenbaum. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *Elife*, 5:e13665, 2016.

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

[19] L. J. Kamin. Attention-like processes in classical conditioning. 1967.

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

[21] A. Dickinson. Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 308(1135):67–78, 1985.

[22] S. Killcross and E. Coutureau. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral cortex*, 13(4):400–408, 2003.

[23] A. Guez, M. Mirza, K. Gregor, R. Kabra, S. Racanière, T. Weber, D. Raposo, A. Santoro, L. Orseau, T. Eccles, G. Wayne, D. Silver, and T. Lillicrap. An investigation of model-free planning, 2019.

[24] M. Janner, J. Fu, M. Zhang, and S. Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, pages 12519–12530, 2019.

[25] T. Yu, G. Thomas, L. Yu, S. Ermon, J. Zou, S. Levine, C. Finn, and T. Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

[26] R. Rafailov, T. Yu, A. Rajeswaran, and C. Finn. Offline reinforcement learning from images with latent space models, 2020.

[27] J. Buckman, D. Hafner, G. Tucker, E. Brevdo, and H. Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *Advances in Neural Information Processing Systems*, pages 8224–8234, 2018.

[28] B. W. Balleine, N. D. Daw, and J. P. O'Doherty. Multiple forms of value learning and the function of dopamine. In *Neuroeconomics*, pages 367–387. Elsevier, 2009.

[29] B. M. Babayan, N. Uchida, and S. J. Gershman. Belief state representation in the dopamine system. *Nature communications*, 9(1):1–10, 2018.

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

[31] S. J. Gershman. A unifying probabilistic view of associative learning. *PLoS Comput Biol*, 11(11): e1004567, 2015.

[32] M. D. Humphries, M. Khamassi, and K. Gurney. Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in neuroscience*, 6:9, 2012.

[33] Y. Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, 2009.

[34] J. Mirenowicz and W. Schultz. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, 379(6564):449–451, 1996.

[35] P. N. Tobler, A. Dickinson, and W. Schultz. Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience*, 23(32):10402–10410, 2003.

[36] M. A. Ungless, P. J. Magill, and J. P. Bolam. Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, 303(5666):2040–2042, 2004.