# Action-modulated midbrain dopamine activity arises from distributed control policies

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Animal behavior is driven by multiple brain regions working in parallel with distinct control policies. We present a biologically plausible model of off-policy reinforcement learning in the basal ganglia, which enables learning in such an architecture. The model accounts for action-related modulation of dopamine activity that is not captured by previous models that implement on-policy algorithms. In particular, the model predicts that dopamine activity signals a combination of reward prediction error (as in classic models) and "action surprise," a measure of how unexpected an action is relative to the basal ganglia's current policy. In the presence of the action surprise term, the model implements an approximate form of $Q$-learning. On benchmark navigation and reaching tasks, we show empirically that this model is capable of learning from data driven completely or in part by other policies (e.g. from other brain regions). By contrast, models without the action surprise term suffer in the presence of additional policies, and are incapable of learning at all from behavior that is completely externally driven. The model provides a computational account for numerous experimental findings about dopamine activity that cannot be explained by classic models of reinforcement learning in the basal ganglia. These include differing levels of action surprise signals in dorsal and ventral striatum, decreasing amounts of movement-modulated dopamine activity with practice, and representations of action initiation and kinematics in dopamine activity. It also provides further predictions that can be tested with recordings of striatal dopamine activity. [1]

## 1 Introduction

An extensive body of work has sought to account for the function of the basal ganglia using the computational framework of reinforcement learning (RL), in particular RL algorithms for action selection and value learning [29, 43, 28, 57]. Striatal neurons within the basal ganglia integrate diverse inputs, including projections from across the cerebral cortex. The activity of these neurons, in particular neurons in the dorsal striatum, plays a key role in action selection [45]. On the other hand, neurons in the ventral striatum have been shown to encode the learned value of stimuli [10, 42, 44]. The phasic activity of midbrain dopamine neurons projecting to the striatum gates plasticity at cortico-striatal synapses [62, 9, 11] and may therefore modulate the learning of actions and values. Indeed, in many RL algorithms, the learning of state-to-action mappings ("policies") and value estimates is modulated by a scalar factor known as the "advantage" or "reward prediction error" (RPE), which measures deviations in attained reward from expectations based on learned value estimates. Numerous experiments have shown that striatal dopamine activity encodes an RPE-like signal [49, 42, 25].

[1] Source code for our experiments to be provided upon publication.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Collectively, these findings suggest a model in which the basal ganglia implements an online actor-critic RL algorithm, where ventral and dorsal striatal subregions play the roles of critic and actor, respectively, and dopaminergic activity encodes the advantage (RPE) signal [29]. This model has been extended in a variety of ways to incorporate more biological detail [55, 56, 11, 8, 54], ideas from model-based RL [15], distributional RL [13], and meta-RL [60].

Despite these promising links, there remain challenges to the view of basal ganglia as implementing an actor-critic RL algorithm. In this work, we address two such challenges. First, dopamine activity in the basal ganglia is observed to encode other information beside RPEs [18], particularly signals relating to movement initiation and vigor [12, 66]. Second, the classic actor-critic model of the basal ganglia is an on-policy RL algorithm—it is designed to learn from experiences driven by the basal ganglia's policy. However, motor control in the brain is distributed, with multiple brain regions including the motor cortex and cerebellum exerting influence on behavior [19, 63, 3, 51, 6]. We show that classic actor-critic models of the basal ganglia fail in this scenario and argue that the ability to learn from off-policy experience is essential to models of basal ganglia learning.

We present a biologically plausible model of off-policy RL in continuous action spaces. Our model differs from the classic actor-critic model of the basal ganglia by adding an additional action-related term to dopamine activity. We call this term "action surprise," as it measures deviation of the action an agent takes in a particular state from the typical action that the output of the basal ganglia would select in that state. We show mathematically that, with this addition, the algorithm implements an approximate form of $Q$-learning in continuous action spaces proposed by [22]. Using simulations, we show that action surprise is essential to effective learning when controllers other than the basal ganglia also contribute to behavior. Thus, action-related activity need not be understood as an independent function of midbrain dopamine neurons separate from their role in learning, but rather as a necessary component of an algorithm that enables off-policy RL. This action surprise model accounts for several experimental findings about movement-related dopamine activity in the basal ganglia. Finally, we describe predictions of our model regarding dopamine activity.

## 2 Background: Connection between policy gradient algorithms and cortico-striatal plasticity

Here we briefly introduce the connection between on-policy policy gradient algorithms and synaptic plasticity rules in the striatum. In what follows, we assume familiarity with standard notation and concepts in RL; for a brief review, see Appendix A.1. Policy gradient algorithms update the parameters $\boldsymbol{\theta}_P$ of the policy according to

$$\Delta\boldsymbol{\theta}_P \propto \nabla_{\boldsymbol{\theta}_P} \log \pi(\mathbf{a}_t|\mathbf{s}_t)\delta_t, \tag{1}$$

where $\delta_t$ is an estimate of the advantage function $A(\mathbf{s}_t, \mathbf{a}_t) = Q(\mathbf{s}_t, \mathbf{a}_t) - V(\mathbf{s}_t)$.

Throughout this work we assume that policies $\pi(\mathbf{a}|\mathbf{s})$ are parameterized by Gaussian distributions $a_i \sim \mathcal{N}(\mu_i(\mathbf{s}), \sigma_i^2)$. For simplicity, we will assume $\sigma_i = \sigma$ for all $i$; however, our results are easily extended to the case of heterogeneous $\sigma_i$. Letting $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ be the parameters of $\boldsymbol{\mu}(\mathbf{s})$, we have

$$\Delta\boldsymbol{\theta}_{\boldsymbol{\mu}} \propto \frac{1}{\sigma^2}\delta_t\left(\mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t)\right)\nabla_{\boldsymbol{\theta}_{\boldsymbol{\mu}}}\boldsymbol{\mu}(\mathbf{s}_t). \tag{2}$$

Supposing our policy is parameterized by a linear map from a feature representation of $\mathbf{s}_t$, $\boldsymbol{\mu}(\mathbf{s}_t) = \mathbf{W}_{\boldsymbol{\mu}}\boldsymbol{\phi}(\mathbf{s}_t)$, this becomes:

$$\Delta\mathbf{W}_{\boldsymbol{\mu}} \propto \frac{1}{\sigma^2}\delta_t\left(\mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t)\right)\boldsymbol{\phi}(\mathbf{s}_t)^\top. \tag{3}$$

This is a "three-factor" learning rule [20] for synapse $W_{ij}$ obtained by multiplying presynaptic activity $\phi_j(\mathbf{s}_t)$, a postsynaptic term $(\mathbf{a}_t)_i - \mu_i(\mathbf{s}_t)$, measuring the deviation of the sampled action from the typical action in this state, and a third factor $\delta_t$. RL models of the basal ganglia assume that $(\mathbf{a}_t)_i - \mu_i(\mathbf{s}_t)$ is available to the postsynaptic neuron and that $\delta_t$ is signaled by dopamine release in the striatum [57, 43]. Experiments have indeed shown that a coincidence of dopamine release and pre and post-synaptic neural activity enables plasticity at cortico-striatal synapses [62, 9, 11]. Biological implementations of this learning rule are discussed further in Appendix A.2.

In actor-critic models the $\delta_t$ factor is

$$\delta_t = r_{t+1} + \gamma\hat{V}(\mathbf{s}_{t+1}) - \hat{V}(\mathbf{s}_t), \tag{4}$$

often referred to as "reward prediction error" (RPE). Here, $\hat{V}$ is an estimate of the value function output by a "critic" network, separate from the policy network, which learns its parameters $\boldsymbol{\theta}_V$ using temporal difference (TD) learning:

$$\Delta\boldsymbol{\theta}_V \propto \delta_t \nabla_{\boldsymbol{\theta}_V} \hat{V}(\mathbf{s}_t). \tag{5}$$

In models of the basal ganglia, the ventral striatum is often assigned the role of the critic, as it is implicated in value-learning but less so in controlling actions [42, 43]. Importantly, TD learning uses the same quantity $\delta_t$ for learning the value function as for policy learning. Hence, a scalar $\delta_t$ signal measuring RPE and broadcast across the striatum supports learning for both the actor and critic. RPE captures key experimental features of striatal dopamine activity—responses to unexpected reward or reward-predictive cues, no significant response to expected reward, and suppression in response to unexpected lack of reward [49, 42, 25].

# 3 Off-policy RL through action surprise signals in dopamine activity

The actor-critic algorithm described above is an on-policy algorithm. Hence, the algorithm may perform suboptimally, or fail altogether, if the actions used for learning are not (always) sampled from the learned policy. Off-policy algorithms are often used to enable learning from a replay buffer of past experiences, expert demonstrations, and/or a separate exploration policy [2]. In a biological context, we argue that the fact that other brain regions can exert control over behavior independently of the basal ganglia further motivates off-policy RL. Indeed, we will show empirically that standard on-policy algorithms suffer when other controllers exert (partial) control over an agent's behavior.

## 3.1 Action-sensitive dopamine activity arises from parameterized $Q$-learning

While there are a variety of approaches to off-policy RL, many require learning an estimate of the $Q$-function $Q(\mathbf{s}, \mathbf{a})$ rather than $V(\mathbf{s})$. $Q$-learning iteratively minimizes the following loss function:

$$\mathcal{L} = \left\| y_{t+1} - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t) \right\|^2, \tag{6}$$

$$y_{t+1} = r_{t+1} + \gamma \max_{\mathbf{a}} \hat{Q}(\mathbf{s}_{t+1}, \mathbf{a}). \tag{7}$$

Computing the quantity $\max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$ directly is intractable in high-dimensional, continuous action spaces. A variety of approaches to this problem have been proposed. Here we focus on an approach adopted by [22], which involves restricting the form of the $Q$-function estimate to a family of functions whose maximum is easy to compute (we discuss other techniques for continuous Q-learning and off-policy RL, which we argue are less biologically realistic, in Appendix A.3).

We parameterize the $Q$-function as follows:

$$\hat{Q}(\mathbf{s}, \mathbf{a}) = \hat{V}(\mathbf{s}) - \frac{1}{\sigma^2} \left\| \mathbf{a} - \boldsymbol{\mu}(\mathbf{s}) \right\|^2. \tag{8}$$

For now we treat the the scaling factor $\sigma$ as a fixed hyperparameter; however, it can also be learned online (see Appendix A.4). This parameterization is a special case of the one proposed in [22].

A primary insight of our work is the observation that, under the parameterization of Eq. 8, gradient updates of the loss function of Eq. 6 yield a biologically plausible actor-critic algorithm with action-sensitive dopamine activity. In particular, taking the gradient of the $Q$-learning loss function with respect to the parameters $\boldsymbol{\theta}_V$ and $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ of $\hat{V}$ and $\boldsymbol{\mu}$, respectively, yields the following learning updates:

$$\Delta\boldsymbol{\theta}_V \propto \delta_t^+ \nabla_{\boldsymbol{\theta}_V} \hat{V}(\mathbf{s}_t), \tag{9}$$

$$\Delta\boldsymbol{\theta}_{\boldsymbol{\mu}} \propto \frac{1}{\sigma^2} \delta_t^+ (\mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t)) \nabla_{\boldsymbol{\theta}_{\boldsymbol{\mu}}} \boldsymbol{\mu}(\mathbf{s}_t), \tag{10}$$

where

$$\delta_t^+ = r_{t+1} + \gamma \hat{V}(s_{t+1}) - \left[ \hat{V}(\mathbf{s}_t) - \frac{1}{\sigma^2} \left\| \mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t) \right\|^2 \right] \tag{11}$$

$$= \delta_t + \frac{1}{\sigma^2} \left\| \mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t) \right\|^2. \tag{12}$$

These update equations are the same as those of the on-policy advantage actor-critic algorithm, but with one additional term added to the dopaminergic signal. Now this signal $\delta^+$ represents a the sum of classic RPE $\delta$ and $\|\mathbf{a} - \boldsymbol{\mu}(\mathbf{s})\|^2$, which measures the deviation of the sampled action from the action most likely to be chosen by the actor. We call this term "action surprise." It can be interpreted as a reduction in the predicted value $Q(\mathbf{s}_t, \mathbf{a}_t)$ of actions that deviate from the basal ganglia's policy.

We note that our model is agnostic to whether action surprise is encoded by the same neurons as RPE. It may be encoded by separate neurons as long as they release dopamine in the same areas as RPE-signaling dopamine. We also note that the update equation (Eq. 9) for the actor uses the action $\mathbf{a}_t$ taken by the agent. Biologically, this requires an efferent copy of the agent's action (taking into account the influence of other controllers) be sent to the striatal projection neurons representing action in the basal ganglia. This architecture is consistent with the presence of pathways from motor cortex, thalamus, and cerebellum to striatal projection neurons [36, 24, 5, 35]. A schematic of the connections involved in our model is depicted in Fig. 1.

It is easy to misinterpret the effect of the dopaminergic action surprise as encouraging the basal ganglia's policy to imitate those of other controllers. This is in fact not the case. Later, we show empirically that this model robustly enables off-policy learning even in the presence of poorly performing external controllers. Below, however, we show how imitation learning can be implemented.

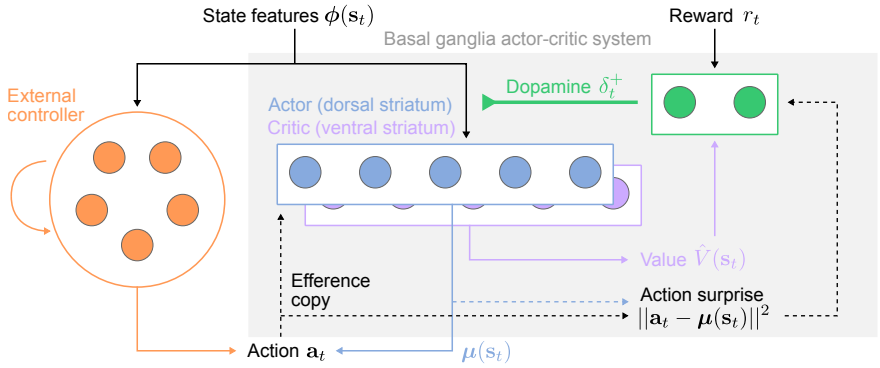


Figure 1: Schematic of model architecture. Actions are driven in parallel by an actor-critic architecture in the basal ganglia (gray region) and an external controller, representing other brain regions (orange). Both systems receive state information via a feature representation $\phi(\mathbf{s}_t)$ (e.g. cortico-striatal projection neuron activity). The actor-critic module performs RL with weight updates modulated by $\delta_t^+$, which is a combination of both RPE and action surprise. Dashed lines indicate architectural features of the action surprise model that are not present in the classic actor-critic model.

### 3.2 Differential action surprise signals in dorsal and ventral striatum

The formulation above predicts that the same dopamine signal $\delta^+$ is broadcast to both the actor and critic. However, the ratio of movement-related to reward-related modulation of dopamine activity appears to vary across the striatum, with more movement-related activity in dorsal regions [26]. In our modeling framework, this corresponds to the action surprise term being weighted more strongly in the actor (dorsal striatum) than the critic (ventral striatum). Such an asymmetry is produced by adding an additional term to the weight updates of the actor, proportional to

$$\|\mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t)\|^2 (\mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t)) \nabla_{\boldsymbol{\theta}_\mu} \boldsymbol{\mu}(\mathbf{s}_t). \tag{13}$$

This update is aligned with the gradient with respect to $\boldsymbol{\theta}_\mu$ of the following loss function:

$$\mathcal{L}_{\text{sup}} = \mathbb{E}\left[\|\mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t)\|^2\right], \tag{14}$$

with the learning rate itself governed by the magnitude of action surprise. Hence, these updates encourage the basal ganglia policy to imitate the agent's behavioral policy on steps with large action surprise. Heuristically, steps with large action surprise are more likely to have been driven by an external controller. An additional action surprise contribution to actor-modulating dopamine activity

can therefore be interpreted as an approximate supervised learning signal encouraging imitation of the external controller's policy. We show in simulations that this effect helps the basal ganglia more rapidly consolidate expert policies of other controllers.

# 4 Experimental Setup

***Tasks***. We simulated two simple continuous control tasks to demonstrate the role of the action surprise term in off-policy learning. See Fig. 2A for illustrations.

*Open-field navigation:* An agent is positioned in a two-dimensional square environment with continuous coordinates. On each trial, the initial position of the agent and a goal location are randomly sampled. The agent controls its $x$ and $y$ acceleration.

*Two-joint:* An agent controls a two-joint arm with two equal-length segments. The arm position and a target location are randomly sampled at the beginning of each trial. The agent outputs torques at each of the joints in order to move the most peripheral point of the arm toward the target.

In both tasks the cost incurred by the agent at each time step is proportional to the inverse of its squared distance from the goal location $\mathbf{g}$ at each time step, plus penalties for its squared velocity and acceleration (summed across joints in the two-joint arm case)

$$r_t = -\left( \|\mathbf{x}(t) - \mathbf{g}\|^2 + \alpha_1 \left\| \frac{d\mathbf{x}(t)}{dt} \right\|^2 + \alpha_2 \left\| \frac{d^2\mathbf{x}(t)}{dt^2} \right\|^2 \right) \tag{15}$$

We also experimented with a version of the loss function in which rewards for reaching the goal location are sparse and binary, rather than continuous, which gave similar results (see Appendix A.7).

***Model architecture***. For both tasks we used a neural network architecture with a single hidden layer. The state inputs to the network consisted of the agent's (angular) position and (angular) velocity, as well as the position of the target location, a total of six scalar variables. Each variable was represented as a one-hot vector by discretizing its domain into 10 equally spaced bins. These vectors were concatenated and used as the network input. The hidden layer had size $256$ and used ReLU nonlinearities. Input weights to the hidden layer were fixed at random (Kaiming uniform) initializations. We fixed these weights to avoid the complexity of biological implementations of backpropagation, since this is not the focus of our work (however, we note that the action surprise model can be applied to deep networks by backpropagating gradients through additional layers). Thus, the hidden layer activations served as a fixed feature representation $\boldsymbol{\phi}(\mathbf{s})$ of the environment state, and all learning occurs in output weights $\mathbf{W}_{\boldsymbol{\mu}}$ and $\mathbf{w}_V$ which output actions $\boldsymbol{\mu} = \mathbf{W}_{\boldsymbol{\mu}}\boldsymbol{\phi}(\mathbf{s})$ and value estimates $\hat{V} = \mathbf{w}_V \cdot \boldsymbol{\phi}(\mathbf{s})$, respectively. We refer to this network as the basal ganglia network.

***External controllers and training protocol***. To model the influence of other brain regions on behavior, we introduced an additional neural network that also exerted control over the agent's actions (Fig. 1, orange). To vary the performance of this controller's policy, we trained the controller on the task with backpropagation for different numbers of steps: 0 (random policy), 2,000 (intermediate policy), or 100,000 (expert policy). We also varied the degree to which the external controller influences behavior. In the fully on-policy case, the external controller is not used. In the fully off-policy case, the external controller entirely drives the behavior. In the case of partial control, each action is sampled from the basal ganglia's network or the external controller with probability 0.5 each. Additionally, we implemented an alternative partial control mechanism, in which the average of the basal ganglia network output and the external controller output is used.

Throughout training, Gaussian exploration noise is added to the output of both networks at each step. In all cases, for the action surprise model, the action $\mathbf{a}_t$ used for the actor update in Eq. 9 is the action taken by the agent, taking into account the contributions of both controllers and the exploration noise.

***On-policy baseline models***. We compared the action surprise model with two baselines in which dopamine neurons signal pure RPE. The first is described by Eqs. 2 and 5, where the action used the actor update is simply the action $\mathbf{a}_t$ taken by the agent, as in the action surprise model. As discussed above, this approach is potentially unstable in the presence of other controllers due to the mismatch between the behavioral policy from which actions are sampled and the basal ganglia network's policy. As a second, potentially more competitive baseline, we instead used the output of the basal ganglia network alone, including exploration noise, in place of $\mathbf{a}_t$ in Eq. 2. We refer to this as the "no efferent

copy" model since the basal ganglia is blind to the action ultimately taken by the agent. In this version, the influence of the external controller is effectively treated as part of the environment. We note that this approach has no ability even in principle to learn from fully off-policy data, but can learn when the basal ganglia network exerts partial control over actions.

***Hyperparameters***. For all models we optimized the learning rate and magnitude of exploration noise as hyperparameters. For the action surprise model we optimized the coefficient $\frac{1}{\sigma^2}$ of the action surprise term as a hyperparameter. To ensure strong baselines we allowed the actor and critic learning rates to be optimized separately for the RPE-only models. See Appendix A.5 for details.

## 5 Experimental Results

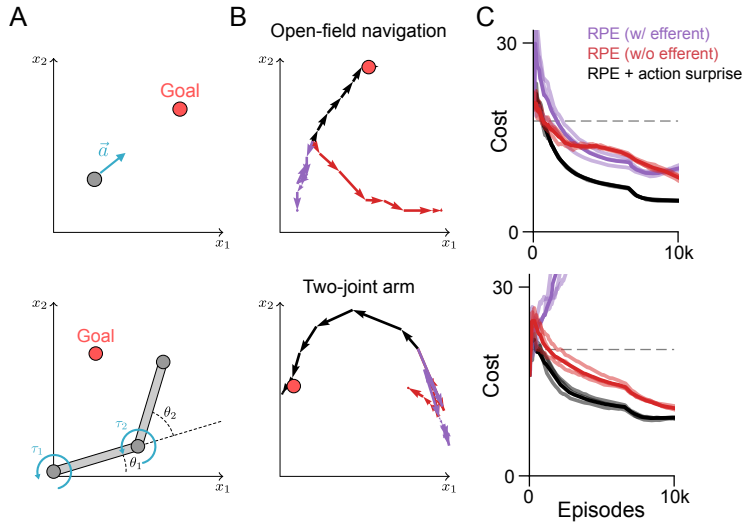


Figure 2: **A**: Depictions of simulated tasks. **B**: Example trials in which the action surprise model successfully reaches the target but the RPE-only models fail to. **C**: Performance of the three models when the basal ganglia shares control of behavior with an external controller. Dark lines indicate mean performance over three runs, and faint individual lines (in this case difficult to see due to tight overlap) indicate individual runs. Performance shown reflects policy of the basal ganglia network and external controller. Dashed line indicates the performance of the external controller alone.

We first analyzed performance for the case of actions partially driven by the basal ganglia network and partially driven by the external controller, intended as a representative model of behavioral control distributed across brain regions. Fig. 2C shows that the basal ganglia policy attained superior learning speed and performance than RPE-only models (shown here is a representative case using the intermediate-level external controller). Fig. 2B shows example trials in which this performance advantage manifests very clearly, with the action surprise model successfully reaching the target while the other models fail completely. The action surprise model remained advantageous when we varied the expertise of the external controller (Fig. 3A). The same was true when we assessed combined performance of the basal ganglia network and the external controller (Fig. 3B), rather than performance of the basal ganglia network by itself. Notably, the RPE-only model with efferent copy failed to learn in some cases. We found similar results when the contributions of the basal ganglia network and external controller were averaged rather than combined by sampling (Appendix A.6).

We next examined the fully off-policy case. The RPE-only model with no efferent copy failed to learn from off-policy data in all cases, as expected, as the action term used in the update for this model is uncorrelated with the actions taken by the external controller. Interestingly, the RPE-only model with efferent copy also failed (catastrophically) to learn in all cases. The action surprise model, by contrast, learned both tasks regardless of the quality of the external controller's policy (Fig. 3C).

A potential concern is that the action surprise model's improvement for off-policy learning comes at the expense of on-policy learning performance. However, we found that even in the case of fully

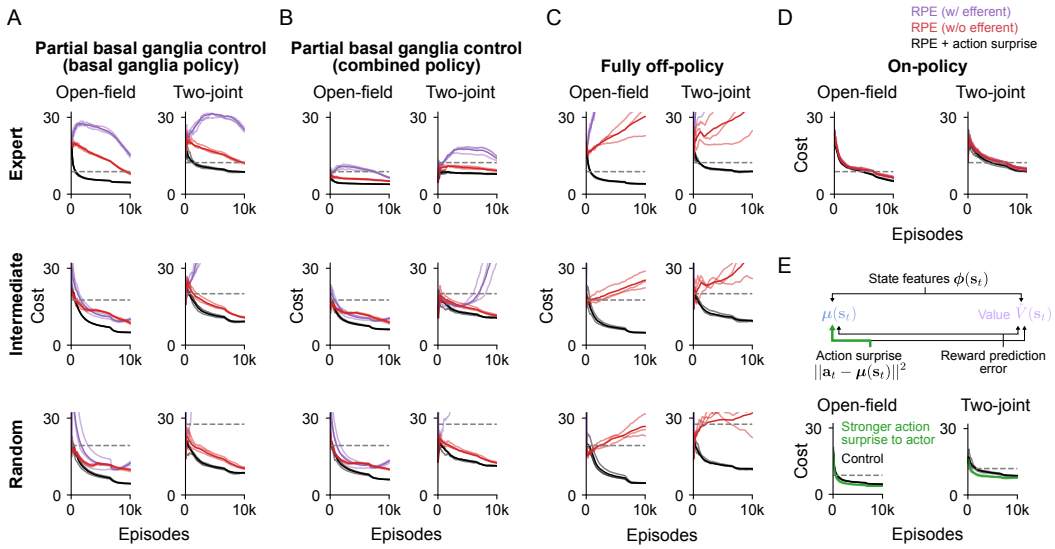


Figure 3: **A**: Performance of the three models, as in Fig. 2C, shown for external controllers with varying levels of expertise (random, intermediate, expert; trained via backpropagation on the task for different numbers of steps). Middle row corresponds to Fig. 2C. **B**: Same as A, but showing the performance of the combined policy of the basal ganglia network and external controller. **C**: Same as A, but in the case where all behavior is driven by the external controller during learning. **D**: Performance of the RPE and RPE+action surprise models for fully on-policy learning on the two tasks. **E**: Top: Schematic of model in which the contribution of the action surprise term to dopamine activity is weighted more strongly in the actor than in the critic (green arrow). Bottom: comparison of this model to the original action surprise model in the case of partial basal ganglia control and fully off-policy learning with the expert controller (black traces are the same as in the top row of panel E). In this example the action surprise term is weighted 8 times as strongly in the actor than the critic.

on-policy learning, with no external controller, the action surprise model matched and indeed slightly outperformed the RPE-only model (Fig. 3D).

We also tested the variant of the action surprise model in which an additional update (Eq. 13) is applied to the actor, corresponding to a higher action surprise coefficient in the actor (dorsal striatum) than in the critic (ventral striatum). We found that this addition improves learning in the presence of expert external controllers (Fig. 3E), consistent with accelerated consolidation of the expert policy.

## 6 Biological implications of the action surprise model

The action surprise model explains several features of midbrain dopamine activity and also makes several testable predictions, which we outline below.

***Nonspecific encoding of action***. A central property of the model is that the action surprise term $\|\mathbf{a}_t - \boldsymbol{\mu}(\mathbf{s}_t)\|^2$ is not action-specific—it reflects only scalar information about the agent's action, even in high-dimensional action spaces, and does not distinguish between two equally surprising actions. This contrasts with the representation in striatal projection neurons forming the output of the actor network, which specify movement commands. Indeed, experimental recordings have found detailed encoding of kinematics in striatal projection neurons [16] but only coarse movement-related signals in dopamine activity that do not reliably distinguish between movement types [12, 37]. This distinction is consistent with our model and inconsistent with models that explain action-modulated dopamine activity in terms of specific motor commands.

***Decrease in action-modulated dopamine activity with learning***. The action surprise model predicts that movement-related dopamine activity is lower when the basal ganglia's policy more closely matches the agent's actions. Before the basal ganglia has learned an effective policy, actions driven by an expert external controller will typically provoke large action surprise. Once the basal ganglia

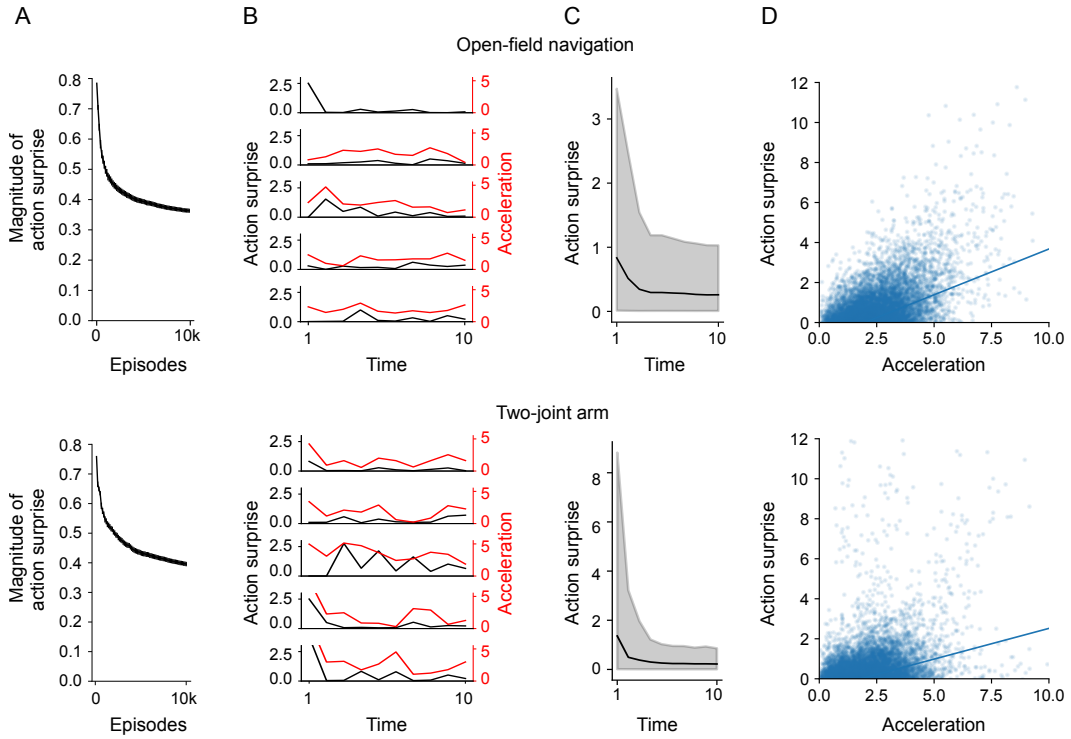


Figure 4: **A**: Mean and standard deviation of action surprise term in the dopamine response over the course of learning, in the case where the basal ganglia shares partial control with an expert controller. In all these examples $\frac{1}{\sigma^2}$ was set to 0.125 and and the exploration noise to 1.0. **B**: Traces of the action surprise signal (black) and the the agent's acceleration (red) – or rotational acceleration averaged across joints in the two-joint arm case – for five randomly sampled example trials. **C**: Mean and 95% confidence interval (across trials) of action surprise as a function of the time within a trial. **D**: Scatter plots (showing individual steps, across many episodes) of the action surprise term in the model dopamine response vs. magnitude of agent acceleration.

network has learned an expert policy itself, action surprise will typically be lower. Thus, we predict a reduction in the magnitude of action surprise signals over the course of learning. We observed this phenomenon in both simulated tasks (Fig. 4A). This prediction is supported by a finding that dopamine activity coinciding with lever-press movements decreases with repeated task practice [12].

***Context-dependent correlation between dopamine activity and acceleration***. Several studies have shown that striatal dopamine activity correlates with movement speed and acceleration [59, 48, 4]. Others [26, 12, 37] have shown that activity in movement-responsive midbrain dopamine neurons is modulated primarily movement at initiation (and in some cases at movement offset) but less so during ongoing movements. Our model gives rise to a complex relationship between movement and action-driven dopamine signals, with the two appearing correlated on some trials but not others (Fig. 4B). On average, action-driven dopamine activity is highest at trial onset (Fig. 4C), and it is positively correlated with the agent's acceleration (Fig. 4D). Our results suggest that the representations of movement initiation and kinematic information attributed to dopamine activity in prior work may actually emerge as byproducts of an action surprise computation.

***Neural circuits underlying the computation of action surprise***. The action surprise model requires information about the agent's action be available to midbrain dopamine neurons for computation of the action surprise term. This is consistent with the presence of pathways from motor cortex and cerebellum to midbrain dopamine neurons [61, 38]. It is also consistent with perturbation experiments finding a causal effect of cortical perturbations and lesions on striatal dopamine activity [32, 53, 1, 7]. We note that the specific form of the action surprise term in our model, $\|\mathbf{a} - \boldsymbol{\mu}(\mathbf{s})\|^2$, is somewhat arbitrary, and other measures of distance between $\mathbf{a}$ and $\boldsymbol{\mu}(\mathbf{s})$ may be equally suitable. We also note that our model is agnostic as to whether the action surprise and RPE components of dopamine activity

are represented in the same individual neurons. Empirically, RPE-signalling and movement-signaling dopamine neurons appear to comprise distinct but partially overlapping populations in the same striatal subregions [12], though data on this question is limited.

***Considerations for future experimental work***. We intend this work to motivate experimental work that can support, challenge, or refine our model. Doing so requires careful experimental design. First, the effects of action-related dopamine activity on learning are easily overlooked in the context of simple tasks in which animals learn to associate cues or actions with immediate reward. In such tasks, standard RPE-based models may be adequate to achieve good performance. Second, disentangling "action surprise" as we define it from other related quantities (initiation, acceleration, motivation, etc.) requires a rich action space, while many animal RL studies involve a small, discrete set of actions (e.g. binary choice). Hence, motor tasks that require sequences of movements involving continuous control [30, 40] are most appropriate for our purposes. Testing predictions of the model also involves disentangling the behavioral contributions of the basal ganglia and other brain regions and analyzing these in concert with simultaneously recorded dopamine activity and behavior. While the technology exists to conduct such simultaneous recordings, no such experiment has been performed. Tests of the model would also benefit from the ability to causally modulate the surprisingness of actions, for instance using catch trials with different target movements.

# 7 Discussion

Our work demonstrates a new link between action-related midbrain dopamine activity and reinforcement learning in the basal ganglia. While each of these topics has received extensive treatment in the neuroscience literature, they have typically been studied separately. Midbrain dopamine activity is known to causally affect movement initiation and invigoration in real time [12, 46]. Consequently, the function of such activity is often regarded as a motivational signal [64, 41], or as a contributor to action selection [17], separate from dopamine's role in reinforcement learning. However, action and RPE-related dopamine release coincide in the same striatal subregions, often even in the same neurons, and with comparable magnitude [59, 48, 12, 37, 18]. Thus, it is likely that both action-related and RPE-related dopamine activity impact cortico-striatal synaptic plasticity and learning. Our results show that such an influence is a necessary component of an architecture capable of off-policy learning. We note that our proposal does not preclude alternative roles for dopamine activity. Indeed, action-related dopamine signals are observed to both precede and follow movement [12]. Early responses may participate in motivation and action selection while lingering responses responses (extending into the critical plasticity window for cortico-striatal synapses [65]) may aid in learning. Whether similar signals can achieve both functions is an important direction for future modeling work. Prior work [17] has suggested that dopaminergic action selection signals may correspond to imagined RPEs of simulated action outcomes according to a learned model of the environment, which may correlate strongly with action surprise if actions are partially controlled by a model-based system.

Our model suggests that distributed control of behavior by many regions plays an important role in learning in the basal ganglia. A number of experiments have observed differential contributions from multiple brain regions during learning. For instance, some motor skills are observed to recruit the motor cortex early in learning before being consolidated into the basal ganglia [30, 27]. These results suggest that regions other than the basal ganglia may be more adept at flexibly adapting to novel tasks, while the basal ganglia specializes in consolidating well-practiced skills. Other studies have observed parallel contributions of model-based and model-free reinforcement learning strategies in the same task [52, 21], revealing arbitration mechanisms leverage the advantages of each [33, 31]. The flexibility afforded by off-policy RL algorithms enables the basal ganglia to benefit from complementary learning and control strategies adopted by other neural circuits. Exploring how off-policy RL algorithms can best leverage these diverse sources of expertise is a fruitful avenue for extensions to our model. For instance, action surprise signals may be modulated by the confidence of external controllers in order to more efficiently learn from expert behavior.

Our work also has implications for reinforcement learning. The introduction of a biologically plausible off-policy reinforcement learning algorithm, involving local learning rules and a single modulatory factor, enables the deployment of off-policy RL on plasticity-enabled neuromorphic hardware [14]. Moreover, we anticipate that insights into how neural circuits learn from off-policy behavior governed by distributed and diverse controllers will provide inspiration for RL algorithms.

# References


[1] Martín F Adrover, Jung Hoon Shin, Cesar Quiroz, Sergi Ferré, Julia C Lemos, and Veronica A Alvarez. Prefrontal cortex-driven dopamine signals in the striatum show unique spatial and pharmacological properties. *Journal of Neuroscience*, 40(39):7510–7522, 2020.

[2] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[3] F Gregory Ashby, Benjamin O Turner, and Jon C Horvitz. Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in cognitive sciences*, 14(5):208–215, 2010.

[4] Joseph W Barter, Suellen Li, Dongye Lu, Ryan A Bartholomew, Mark A Rossi, Charles T Shoemaker, Daniel Salas-Meza, Erin Gaidis, and Henry H Yin. Beyond reward prediction errors: the role of dopamine in movement kinematics. *Frontiers in integrative neuroscience*, 9:39, 2015.

[5] Andreea C Bostan, Richard P Dum, and Peter L Strick. The basal ganglia communicate with the cerebellum. *Proceedings of the national academy of sciences*, 107(18):8452–8456, 2010.

[6] Andreea C Bostan and Peter L Strick. The basal ganglia and the cerebellum: nodes in an integrated network. *Nature Reviews Neuroscience*, 19(6):338–350, 2018.

[7] Michael G Boyeson and Dennis M Feeney. Striatal dopamine after cortical injury. *Experimental neurology*, 89(2):479–483, 1985.

[8] Joshua Brown, Daniel Bullock, and Stephen Grossberg. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience*, 19(23):10502–10511, 1999.

[9] Paolo Calabresi, Paolo Gubellini, Diego Centonze, Barbara Picconi, Giorgio Bernardi, Karima Chergui, Per Svenningsson, Allen A Fienberg, and Paul Greengard. Dopamine and camp-regulated phosphoprotein 32 kda controls both striatal long-term depression and long-term potentiation, opposing forms of synaptic plasticity. *Journal of Neuroscience*, 20(22):8443–8451, 2000.

[10] Rudolf N Cardinal, John A Parkinson, Jeremy Hall, and Barry J Everitt. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience & Biobehavioral Reviews*, 26(3):321–352, 2002.

[11] José L Contreras-Vidal and Wolfram Schultz. A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of computational neuroscience*, 6(3):191–214, 1999.

[12] Joaquim Alves Da Silva, Fatuel Tecuapetla, Vitor Paixão, and Rui M Costa. Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*, 554(7691):244–248, 2018.

[13] Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.

[14] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1):82–99, 2018.

[15] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

[16] Ashesh K Dhawale, Steffen BE Wolff, Raymond Ko, and Bence P Ölveczky. The basal ganglia control the detailed kinematics of learned motor skills. *Nature neuroscience*, 24(9):1256–1269, 2021.

[17] Kenji Doya. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural networks*, 12(7-8):961–974, 1999.

[18] Ben Engelhard, Joel Finkelstein, Julia Cox, Weston Fleming, Hee Jae Jang, Sharon Ornelas, Sue Ann Koay, Stephan Y Thiberge, Nathaniel D Daw, David W Tank, et al. Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons. *Nature*, 570(7762):509–513, 2019.

[19] Cornelia Exner, Janka Koschack, and Eva Irle. The differential role of premotor frontal cortex and basal ganglia in motor sequence learning: evidence from focal basal ganglia lesions. *Learning & Memory*, 9(6):376–386, 2002.

[20] Nicolas Frémaux and Wulfram Gerstner. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in neural circuits*, 9:85, 2016.

[21] Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, 2010.

[22] Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International conference on machine learning*, pages 2829–2838. PMLR, 2016.

[23] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[24] Suzanne N Haber. Corticostriatal circuitry. *Dialogues in clinical neuroscience*, 2022.

[25] James C Houk and James L Adams. 13 a model of how the basal ganglia generate and use neural signals that. *Models of information processing in the basal ganglia*, page 249, 1995.

[26] Mark W Howe and Daniel A Dombeck. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature*, 535(7613):505–510, 2016.

[27] Eun Jung Hwang, Jeffrey E Dahlen, Yvonne Yuling Hu, Karina Aguilar, Bin Yu, Madan Mukundan, Akinori Mitani, and Takaki Komiyama. Disengagement of motor cortex from movement control during long-term learning. *Science advances*, 5(10):eaay0001, 2019.

[28] Makoto Ito and Kenji Doya. Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current opinion in neurobiology*, 21(3):368–373, 2011.

[29] Daphna Joel, Yael Niv, and Eytan Ruppin. Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural networks*, 15(4-6):535–547, 2002.

[30] Risa Kawai, Timothy Markman, Rajesh Poddar, Raymond Ko, Antoniu L Fantana, Ashesh K Dhawale, Adam R Kampff, and Bence P Ölveczky. Motor cortex is required for learning but not for executing a motor skill. *Neuron*, 86(3):800–812, 2015.

[31] Dongjae Kim, Geon Yeong Park, Sang Wan Lee, et al. Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nature communications*, 10(1):1–14, 2019.

[32] Polina Kosillo, Yan-Feng Zhang, Sarah Threlfell, and Stephanie J Cragg. Cortical control of striatal dopamine transmission via striatal cholinergic interneurons. *Cerebral cortex*, 26(11):4160–4169, 2016.

[33] Sang Wan Lee, Shinsuke Shimojo, and John P O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014.

[34] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[35] Nikolaus R McFarland and Suzanne N Haber. Convergent inputs from thalamic motor nuclei and frontal cortical areas to the dorsal striatum in the primate. *Journal of Neuroscience*, 20(10):3798–3813, 2000.

[36] Sarah Melzer, Mariana Gil, David E Koser, Magdalena Michael, Kee Wui Huang, and Hannah Monyer. Distinct corticostriatal gabaergic neurons modulate striatal output neurons and motor activity. *Cell reports*, 19(5):1045–1055, 2017.

[37] Marcelo D Mendonça, Joaquim Alves da Silva, Ledia F Hernandez, Ivan Castela, José Obeso, and Rui M Costa. Transient dopamine neuron activity precedes and encodes the vigor of contralateral movements. *bioRxiv*, 2021.

[38] William Menegas, Joseph F Bergan, Sachie K Ogawa, Yoh Isogai, Kannan Umadevi Venkataraju, Pavel Osten, Naoshige Uchida, and Mitsuko Watabe-Uchida. Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *elife*, 4:e10032, 2015.

[39] Nicolas Meuleau, Leonid Peshkin, and Kee-Eung Kim. Exploration in gradient-based reinforcement learning. 2001.

[40] Kevin GC Mizes, Jack Lindsey, G Sean Escola, and Bence P Ölveczky. Similar striatal activity exerts different control over automatic and flexible motor sequences. *bioRxiv*, 2022.

[41] Ali Mohebi, Jeffrey R Pettibone, Arif A Hamid, Jenny-Marie T Wong, Leah T Vinson, Tommaso Patriarchi, Lin Tian, Robert T Kennedy, and Joshua D Berke. Dissociable dopamine dynamics for learning and motivation. *Nature*, 570(7759):65–70, 2019.

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

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

[44] John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, 304(5669):452–454, 2004.

[45] Mark G Packard and Barbara J Knowlton. Learning and memory functions of the basal ganglia. *Annual review of neuroscience*, 25(1):563–593, 2002.

[46] Babita Panigrahi, Kathleen A Martin, Yi Li, Austin R Graves, Alison Vollmer, Lars Olson, Brett D Mensh, Alla Y Karpova, and Joshua T Dudman. Dopamine is required for the neural representation and control of movement vigor. *Cell*, 162(6):1418–1430, 2015.

[47] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[48] Corey B Puryear, Min Jung Kim, and Sheri JY Mizumori. Conjunctive encoding of movement and reward by ventral tegmental area neurons in the freely navigating rodent. *Behavioral neuroscience*, 124(2):234, 2010.

[49] Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[50] Weixing Shen, Marc Flajolet, Paul Greengard, and D James Surmeier. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, 321(5890):848–851, 2008.

[51] Maria Caterina Silveri. Contribution of the cerebellum and the basal ganglia to language production: Speech, word fluency, and sentence construction—evidence from pathology. *The Cerebellum*, 20(2):282–294, 2021.

[52] Peter Smittenaar, Thomas HB FitzGerald, Vincenzo Romei, Nicholas D Wright, and Raymond J Dolan. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron*, 80(4):914–919, 2013.

[53] Antonio P Strafella, Tomáš Paus, Maria Fraraccio, and Alain Dagher. Striatal dopamine release induced by repetitive transcranial magnetic stimulation of the human motor cortex. *Brain*, 126(12):2609–2615, 2003.

[54] Roland E Suri, J Bargas, and MA Arbib. Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, 103(1):65–85, 2001.

[55] Roland E Suri and Wolfram Schultz. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental brain research*, 121(3):350–354, 1998.

[56] Roland E Suri and Wolfram Schultz. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91(3):871–890, 1999.

[57] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[58] Eiji Uchibe. Cooperative and competitive reinforcement and imitation learning for a mixture of heterogeneous learning modules. *Frontiers in neurorobotics*, 12:61, 2018.

[59] Dong V Wang and Joe Z Tsien. Conjunctive processing of locomotor signals by the ventral tegmental area neuronal population. *PloS one*, 6(1):e16528, 2011.

[60] Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868, 2018.

[61] Mitsuko Watabe-Uchida, Lisa Zhu, Sachie K Ogawa, Archana Vamanrao, and Naoshige Uchida. Whole-brain mapping of direct inputs to midbrain dopamine neurons. *Neuron*, 74(5):858–873, 2012.

[62] JR Wickens, AJ Begg, and GW Arbuthnott. Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. *Neuroscience*, 70(1):1–5, 1996.

[63] Dirk Wildgruber, Hermann Ackermann, and Wolfgang Grodd. Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fmri. *Neuroimage*, 13(1):101–109, 2001.

[64] Roy A Wise. Dopamine, learning and motivation. *Nature reviews neuroscience*, 5(6):483–494, 2004.

[65] Sho Yagishita, Akiko Hayashi-Takagi, Graham CR Ellis-Davies, Hidetoshi Urakubo, Shin Ishii, and Haruo Kasai. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1620, 2014.

[66] Alexandre Zénon, Sophie Devesse, and Etienne Olivier. Dopamine manipulation affects response vigor independently of opportunity cost. *Journal of Neuroscience*, 36(37):9516–9525, 2016.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is not primarily intended to advance the state of the art in RL or any particular engineering application, but rather to provide basic scientific insights about computation in the brain. As such, we regard the potential for negative societal impacts of this work to be limited.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Section 3.1, in particular Equations (6) and (8)
   (b) Did you include complete proofs of all theoretical results? [Yes] Section 3.1
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Code will be released upon publication.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code will be provided as a URL upon publication.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]