Vector-valued dopamine improves learning of continuous outputs in the striatum

Emil Wärnberg[1, 2] and Arvind Kumar[2]

[1]Dept. of Neuroscience, Karolinska Institutet, Stockholm, Sweden

[2]Division of Computational Science and Technology, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

Abstract

It is well established that midbrain dopaminergic neurons support reinforcement learning (RL) in the basal ganglia by transmitting a reward prediction error (RPE) to the striatum. In particular, different computational models and experiments have shown that a striatum-wide RPE signal can support RL over a small discrete set of actions (e.g. no/no-go, choose left/right). However, there is accumulating evidence that the basal ganglia functions not as a selector between predefined actions, but rather as a dynamical system with graded, continuous outputs. To reconcile this view with RL, there is a need to explain how dopamine could support learning of dynamic outputs, rather than discrete action values.

Inspired by the recent observations that besides RPE, the firing rates of midbrain dopaminergic neurons correlate with motor and cognitive variables, we propose a model in which dopamine signal in the striatum carries a vector-valued error feedback signal (a loss gradient) instead of a homogeneous scalar error (a loss). Using a recurrent network model of the basal ganglia, we show that such a vector-valued feedback signal results in an increased capacity to learn a multidimensional series of real-valued outputs. The corticostriatal plasticity rule we employed is based on Random Feedback Learning Online learning and is a fully local, ”three-factor” product of the presynaptic firing rate, a post-synaptic factor and the unique dopamine concentration perceived by each striatal neuron. Crucially, we demonstrate that under this plasticity rule, the improvement in learning does not require precise nigrostriatal synapses, but is compatible with random placement of varicosities and diffuse volume transmission of dopamine.

Introduction

The basal ganglia are thought to be the main locus of reinforcement learning (RL) in the brain (Niv and Langdon, 2016). In particular, dopamine modulated long-term plasticity in the corticostriatal synapses is crucial for learning and fine-tuning skilled movements based on environmental feedback (Perrin and Venance, 2019). Combined with the striking observation that midbrain dopaminergic cells transmit a reward prediction error (RPE) to the striatum (Schultz et al., 1997; Kim et al., 2020), this has inspired a plethora of computational models of the basal ganglia implementing various forms of RL. Notably however, virtually all these models assume the set of actions that can be selected — the *action space* in RL terminology — is small and discrete (see e.g. Humphries et al., 2006; Stewart et al., 2012; Berthet et al., 2012; Bolado-Gomez and Gurney, 2013; Gurney et al., 2015; Baston and Ursino, 2015; Bogacz et al., 2016; Berthet et al., 2016; Dunovan et al., 2019; Bahuguna et al., 2019; Pozzi et al., 2020). Practically, this means that each action can be exclusively represented by a disjoint group of striatal neurons, sometimes called *action channels* (Redgrave et al., 1999). At their core, in each of these models there is some sort of competition between the action channels so that the selected action (or likely selected in probabilistic models) corresponds to the channel with the highest activity. This is consistent with a global RPE transmitted by dopamine that reinforces or depresses the corticostriatal synapses of the active channel.

However, there is now accumulating evidence that action space of the basal ganglia is not small and discrete, but rather multidimensional and continuous (Yin, 2014; Barter et al., 2015; Rueda-Orozco and Robbe, 2015; Park et al., 2020; Dhawale et al., 2021). For a multidimensional output, a global RPE is not as effective at driving learning as it is for discrete action channels. For example, a mouse learning to reach for a food pellet may need to learn to control its paw in the x, y, and z directions. Intuitively, a single error, perhaps proportional to the final distance to the target, would be less efficient than having a three-dimensional error signal representing the error in the three directions. More formally, producing a continuous and multidimensional output requires the basal ganglia to learn a function approximation rather than tabular values (Sutton et al., 1999). In this formulation, a scalar RPE signal transmitted to the entire network corresponds to perturbation learning (Chen and Goldberg, 2020), which is well-known to be considerably less efficient than vector valued feedback, especially for a network with multiple layers (Lillicrap et al., 2020).

Therefore, we asked if it would be possible for dopamine to support function approximation learning in the basal ganglia by carrying a vector-valued feedback signal from the midbrain back to the striatum. Such a feedback signal would manifest in the VTA and SNc in terms of cell tunings to various task-related variables, consistent with recent observations (Fan et al., 2012; Howe and Dombeck, 2016; Engelhard et al., 2019). However, one apparent problem with dopamine transmitting a vector-valued error is that dopaminergic axons do not precisely target specific neurons in the striatum, but instead release dopamine from a large number of varicosities that can then diffuse over a short distance through extracellular space (Arbuthnott and Wickens, 2007), thereby mixing any individual error components. In principle one could imagine this problem being solved by representing each action dimension in a spatially compact region that receives its private dopamine channel (Hamid et al. (2021) and Lee et al. (2022) have proposed similar ideas). Firstly however, experimental evidence suggests both individual striatal projection neurons (Rueda-Orozco and Robbe, 2015; Klaus et al., 2017; Weglage et al., 2021) and individual dopaminergic neurons (Avvisati et al., 2022) have mixed tuning rather than responding to a single task variable, as would be predicted from a naive parallel organization. Secondly, although there is a coarse-grain somatotopic organisation of the striatum (Hunnicutt et al., 2016) as well as the substantia nigra (Foster et al., 2021), the axonal arborizations of individual SNc neurons are huge and cover large portions of the striatum (up to 5%; Matsuda et al., 2009). Therefore, a large set of isolated parallel channels without cross-talk appears unlikely.

In this work, rather than separating the entire basal ganglia into fine-grained parallel channels, we propose that the mixing of multiple error components can be undone downstream from the striatum. In particular, we propose that if the striatofugal projections were to be subject to similar systematic long-term plasticity as the corticostriatal projections, then we can make use of *feedback alignment* (Lillicrap et al., 2016) to have striatum learn continuous outputs efficiently. We demonstrate that learning in a recurrent neural network model of the basal ganglia is better with a stylized model of diffuse mixing of dopaminergic feedback compared to a model where there is a homogeneous, scalar dopamine concentration in the entire striatum. Thus, we connect two seemingly unrelated observations: heterogeneous dopamine response and the involvement of striatum in learning continuous actions.

Results

Network model

As a model of the basal ganglia learning a skilled movement (such as an animal reaching for a food pellet or pressing a lever), we constructed a task wherein a recurrent neural network must learn to repeatedly output a given trajectory in d-dimensional space. The output trajectory is defined as the activity of d readout neurons. In our idealized model of a small piece of basal ganglia, we take the readout population to be the either the internal globus pallidus (GPi) or the substantia nigra pars reticulata (SNr) (see Barter et al., 2015, for experimental support).

This striatum projects to the readout population (GPi/SNr) and receives excitatory inputs from two input populations: cortex and thalamus (Fig. 1A). The task of the network is to adjust the input and recurrent synaptic weights in the striatum so that the readout matches the desired d-dimensional target $T(t)$ as closely as possible (Fig. 1I).

In cortex and striatum, we model the sub-threshold membrane potential $V_i(t)$ of the ith neuron (or small group of neurons) as

$$\frac{\mathrm{d}V_i(t)}{\mathrm{d}t} = -\frac{1}{\tau_m}V_i(t) + \sum_j w_{ji} r_j(t) \tag{1}$$

where τ_m is the membrane time constant, w_{ji} is the signed synaptic weight from neuron i to neuron j and $r_j(t)$ is the firing rate of unit j given as:

$$r_j(t) = \phi\left(V_j(t)\right) = \frac{1}{1 + e^{-V_j(t)+b}} \tag{2}$$

where the term $b = 2$ shifts the sigmoid to the right so that the firing rates are sparser when the inputs are balanced and the membrane potentials fluctuate around 0 (Fig. 1C).

The purpose of our model is to demonstrate that mixing of dopamine is not detrimental to vector-valued feedback, not to capture every detail of basal ganglia. Nevertheless, to make sure the learning set-up is fair, we added a number of constraints to the model. First, all connections except the readout are sparse, i.e. only a fraction of pairs of neurons are allowed to connect (Fig. 1B). Second, we required the signs of the weights w_{ij} to match the sign of the projection (excitatory or inhibitory) throughout learning (Fig. 1A-B). For simplicity, we omitted the external globus pallidus (GPe) and subthalamic nucleus (STN) and modelled the indirect pathway as a direct excitatory projection from striatum to GPi/SNr. The sign of dopamine-driven plasticity was reversed for the striatal projection neurons in the indirect pathway (iSPNs; Fig. 1A). Because

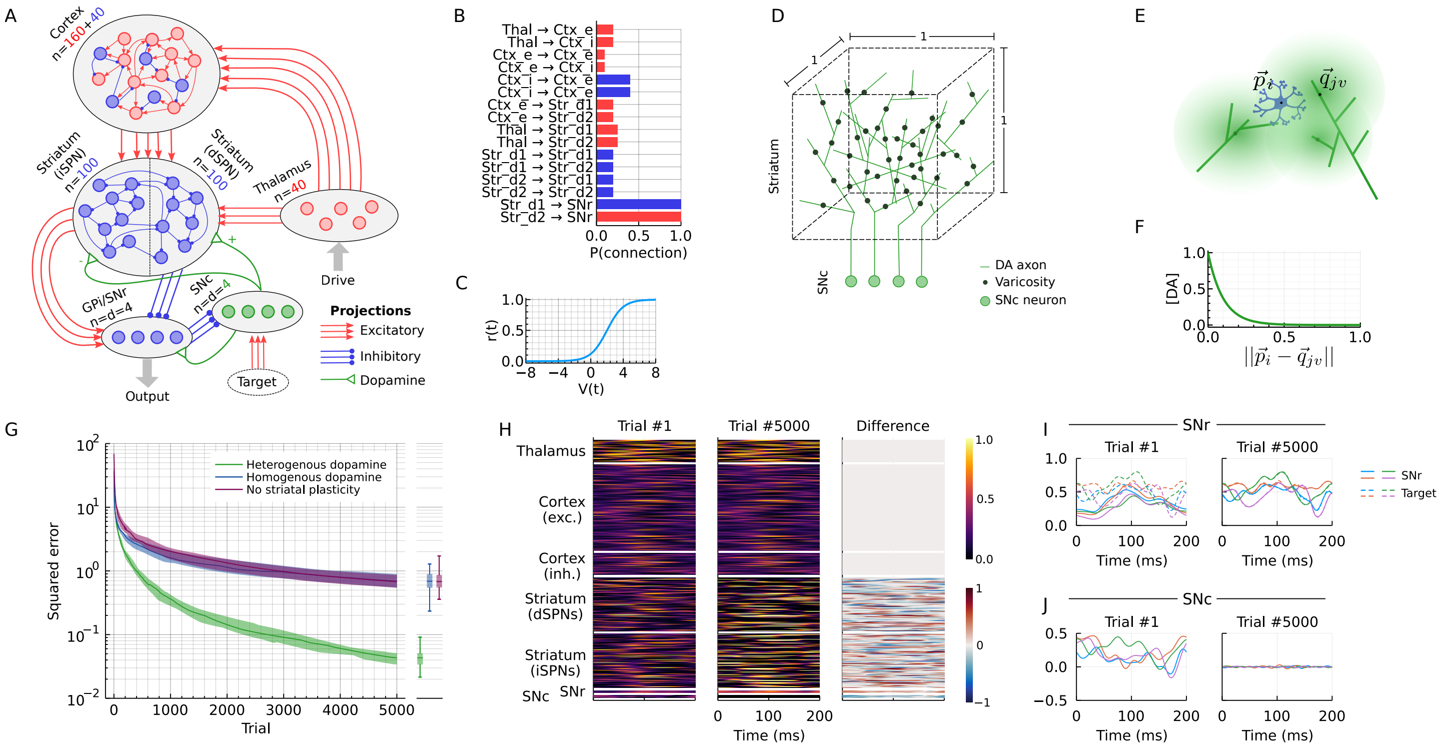

Figure 1: Heterogeneous dopamine improves learning in a RNN. (A) Network architecture. n indicates the number of units in each population. (B) Connection probability of a synapse for any pair of neurons in the network model. (C) Transfer function of each unit. D) Illustration of placement of varicosities in a unit cube. (E) Illustration of a striatal projection neurons (SPN) (located at $\vec{p_i} = [p^i_x, p^i_y, p^i_z]$) receiving a mix of dopamine from three close-by varicosities. (F) Relative dopamine concentration as a function of distance from the varicosity. (G) Convergence of the loss (L2) over trials for the model (heterogeneous dopamine) as well as two control models (homogeneous dopamine and no striatal plasticity). Each trial is one presentation of the pattern. Solid lines indicate the median of 25 runs; shaded areas show first and third quartiles. Boxplots show median, quartiles and min/max of the same 25 runs at the final (5000th) training trial. (H) Example activity in the network in the first and last trials of training. (I) Comparison of the 4-dimensional readout (i.e. SNr/GPi activity) with the target in the first and last trial of training. (J) The difference between the readout and the target (i.e. SNc activity) in the first and last trial of training.

of our focus on dopamine-dependent learning in the striatum, we only include the cortex as a reservoir of rich but task-aligned dynamics, and do not consider any learning that might take place in the cortex itself. However, we include dopamine-dependent plasticity in all basal ganglia synapses: corticostriatal, thalamostriatal, striatostriatal and striatofugal.

Derivation of the synaptic plasticity rule

To construct normatively appropriate learning rules for the plastic synapses, we note the task is to minimize the loss

$$\ell(t) = \frac{1}{2}\sum_{k=1}^{d}\left(r_k^{\mathrm{SNr}}(t) - T_k(t)\right)^2 \tag{3}$$

For the error to decrease over time, we would like the plasticity rule to change the weight of each corticostriatal synapse (w_{ij}^{CtxStr}) such that

$$\frac{\mathrm{d}w_{ji}^{\mathrm{CtxStr}}(t)}{\mathrm{d}t} \propto -\frac{\partial \ell(t)}{\partial w_{ji}^{\mathrm{CtxStr}}(t)} \tag{4}$$

In Supplementary Text , we show that by expanding this partial derivative with two simplifications, only considering ℓ only at t (i.e. not backwards or forward in time) and treating the firing rates of other SPNs as fixed (Murray, 2019), we arrive at the following plasticity rule:

$$\begin{aligned}
\frac{\mathrm{d}w_{ji}^{\mathrm{CtxStr}}(t)}{\mathrm{d}t} &= -\alpha\gamma_j(t)p_{ji}(t) && (5)\\
\tau^{\mathrm{Str}}\frac{\mathrm{d}p_{ji}}{\mathrm{d}t} &= -p_{ji}(t) + r_j^{\mathrm{Str}}(t)\left(1 - r_j^{\mathrm{Str}}(t)\right) r_i^{\mathrm{Ctx}}(t) && (6)\\
\gamma_j(t) &= \sum_{k=1}^{d}\epsilon_k(t)w_{kj}^{\mathrm{StrSNr}}(t) && (7)\\
\epsilon_k(t) &= \left(r_k^{\mathrm{SNr}}(t) - T_k(t)\right) r_k^{\mathrm{SNr}}(t)\left(1 - r_k^{\mathrm{SNr}}(t)\right) && (8)
\end{aligned}$$

The plasticity rules for the thalamostriatal and striatostrial synaptic weights are fully analogous. We interpret this plasticity rule in biological terms as follows. From Eq. 5, we see that the weight update depends on a neuron-specific factor γ_j and a synapse-specific factor p_{ji}. The latter is a low-pass filtered trace of a Hebbian-like product between pre- and postsynaptic firing. This could be identified as an eligibility trace (Izhikevich, 2007; Gurney et al., 2015) and we note it could be represented for example by the local concentration of calcium in the spine.

The eligibility trace p_{ji} is multiplied by a "third factor" γ_j. Experimental results suggest plasticity in corticostriatal synapses depends on three factors: presynaptic activity, postsynaptic activity, and dopamine (Fisher et al., 2017; Gerstner et al., 2018). Given that the two former are captured by p_{ji}, we would like to associate γ_j in Eq. 5 with dopamine. If we assume the number of dopaminergic neurons to be the same as the number of read-out neurons, we can assign ϵ_k to the firing rate r_k^{SNc}. Note that we can handle negative values of ϵ_k by loosely interpreting r_k^{SNc} as the deviation from some baseline firing rate. This leaves just one problem: the coefficients used to sum to contribution of the dopaminergic cells in Eq. 7 should be the *downstream* striatofugal weights w_{kj}^{StrSNr}, which are not available to the corticostriatal synapses. However, Murray (2019) showed that, if the readout weights w_{kj}^{StrSNr} themselves are plastic, we can replace w_{kj}^{StrSNr} in Eq. 7 with a random value at only a minor cost to the convergence of the loss, thanks to a phenomenon called *feedback alignment* (Lillicrap et al., 2016). Therefore, we next asked whether the dopaminergic nigrostriatal projection could form such a random feedback matrix.

Dopamine diffusion

To investigate if dopaminergic feedback could be used to communicate the third factor needed for the corticostriatal plasticity, we set up a stylized model of dopamine diffusion. We assumed each striatal neuron had a position $\vec{p_j} = [p_x^j, p_y^j, p_z^j]$ where $p_x^j, p_y^j, p_z^j \in [0, 1]$. Second, we assumed that each SNc neuron sent axonal projections that covered the entire cube, and that axonal arbor of each SNc neuron has $N^{\mathrm{var}} = 10$ varicosities randomly placed in the unit cube (Fig. 1D). Third, we assumed the dopamine released from each varicosity is proportional to the firing rate at the soma in the SNc and that the dopamine concentration decreases exponentially with distance from the varicosity. This gives the dopamine concentration $C_j(t)$ at striatal neuron j as

$$C_j(t) = \sum_{k=1}^{N^{\mathrm{SNc}}} \sum_{v=1}^{N^{\mathrm{Var}}} r_k^{\mathrm{SNc}}(t) \frac{1}{\lambda} e^{\frac{||\vec{p_j} - \vec{q_{kv}}||}{\lambda}} \tag{9}$$

where $\vec{q_{kv}} = [q_x^{kv}, q_y^{kv}, q_z^{kv}]$ is the position of the vth varicosity of SNc neuron k. λ controls the rate of decay with distance and was set to 0.1 (so that dopamine concentration decreases to $1/e \approx 37\%$ after diffusing a distance equivalent to 10% of the side of the cube). Under this model, there is an effective nigrostrital weight

$$d_{jk} = \sum_{v=1}^{N^{\mathrm{Var}}} \frac{1}{\lambda} e^{-\frac{||\vec{p_i} - \vec{q_{kv}}||}{\lambda}} \tag{10}$$

that does not vary with time, so we can write

$$C_i(t) = \sum_{k=1}^{N^{\mathrm{SNc}}} r_k^{\mathrm{SNc}}(t) d_{jk} \tag{11}$$

Note the similarity between Eqs. 11 and 7. If we introduce plasticity in the striatofugal projection, feedback alignment will cause $w_{kj}^{\mathrm{StrSNr}}(t) \to d_{jk}$. Therefore, we again start with

$$\frac{\mathrm{d}w_{kj}^{\mathrm{StrSNr}}(t)}{\mathrm{d}t} \propto -\frac{\partial \ell(t)}{\partial w_{kj}^{\mathrm{StrSNr}}(t)} \tag{12}$$

and arrive at

$$\frac{\mathrm{d}w_{kj}^{\mathrm{StrSNr}}(t)}{\mathrm{d}t} = -\beta \epsilon_k(t) r_j^{\mathrm{Str}}(t) \tag{13}$$

with $\epsilon_k(t) = r_k^{\mathrm{SNc}}(t)$ as before and $\beta = 10^{-3}$.

In summary, we set the corticostriatal plasticity update to

$$\frac{\mathrm{d}w_{ji}^{\mathrm{CtxStr}}(t)}{\mathrm{d}t} = -\alpha p_{ji}(t) \sum_{k=1}^{N^{\mathrm{SNc}}} r_k^{\mathrm{SNc}}(t) d_{jk} \tag{14}$$

where d_{jk} is given by Eq. 10 and $p_{ji}(t)$ is given by Eq. 6. We set $\alpha = -2.5 \cdot 10^{-2}$ for direct pathway neurons and $\alpha = 2.5 \cdot 10^{-2}$ for indirect pathway neurons to capture the different effects of D1 and D2 receptors. The minus sign helps $w_{kj}^{\mathrm{StrSNr}}(t) \to d_{jk}$ for direct pathway striatal neurons, because for these neurons $w_{kj}^{\mathrm{StrSNr}}(t) < 0$. We used analogous plasticity rules for thalamostriatal and striatostriatal connections.

Learning with vector valued dopamine feedback

Having set up the model, we simulated the dynamics for 5000 presentations of the target output (Fig. 1H). Each target was a 4-dimensional 200ms time series drawn from a Gaussian process (Fig. 1I; see Methods). In the first trial, the output does not match the target (Fig. 1I, left), but after 5000 trials the plasticity rules have driven the network to produce GPi/SNr output that closely matches with the targets (Fig. 1I, right). This is achieved both by plasticity in the readout population (GPi/SNr; Eq. 14) and plasticity in the striatum that adapts the SPN firing rates (Eq. 14; Fig. 1H).

We next asked how this learning depends on the nature of the dopamine feedback. It is well-known that for recurrent networks with rich dynamics, plasticity in the readout is sufficient to learn complex patterns (Jaeger and Haas, 2004; Maass and Markram, 2004). Therefore, we first compared our model to a reduced model that only had plasticity in the readout (striatofugal) projection according to Eq. 13, i.e. no plasticity in the striatum. We found that learning with dopamine feedback was faster (Fig. 1G). Next, we compared our model to a model in which each SPN at every timepoint receives the same dopamine feedback, i.e. the striatum receives a homogeneous, scalar dopamine signal. Strikingly, this model performs no better than the

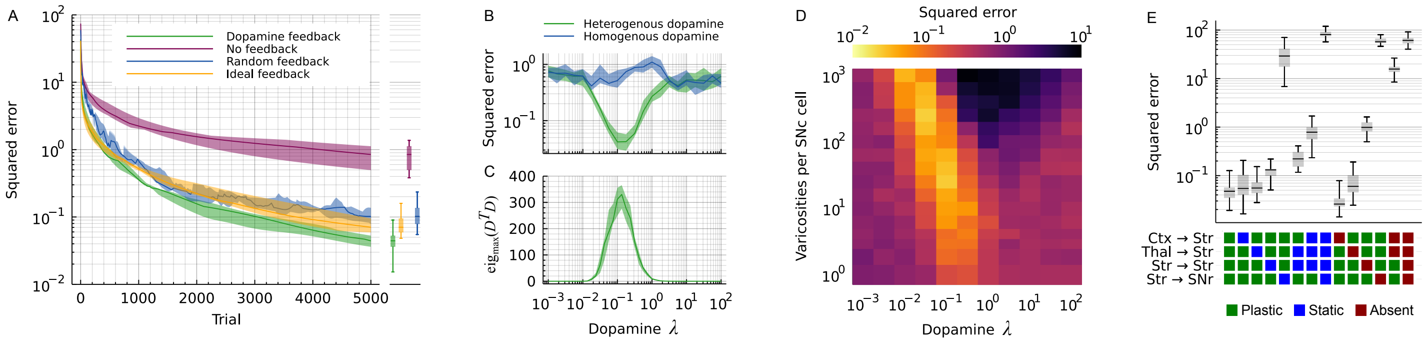

Figure 2: Learning improvement depends on feedback alignment. (A) Convergence of the loss (L2) for different types of feedback: dopamine feedback as described by our model, no feedback at all (effectively making all synapses except the striatofugal static), random feedback where the elements of the feedback matrix D are shuffled randomly and ideal feedback where $D = W^T_{\mathrm{StrSNr}}$. (B) Squared error after 5000 trials when some of the four plastic projections is impaired: either set to *static*, i.e. no plasticity, or removed from the model altogether (*absent*). (C) Squared error after 5000 trials as a function of the spatial constant of dopamine diffusion (λ). λ is measured in relation to the side of the cube in Fig. 1D. (D) Linear independence of the rows of D as a function of the dopamine spatial constant λ. (E) Squared loss after 5000 trials as a joint function of the dopamine spatial constant λ and the number of varicosities per SNc cell. See Supplementary Fig. S1D for the complete set of manipulations. All panels) errors are reported as the median over 25 trials. Shaded areas indicate the first and third quartiles of the 25 trials. Boxplots show median, first and third quartiles; whiskers indicate min and max.

reduced model that only has learning in the readout layer (Fig. 1G). The increase in learning performance persisted with different sizes of the striatal (Supplementary Fig. S1A) and the readout (Supplementary Fig. S1B) populations, as well as for faster and slower timescales of the target (τ_{task} in Eq. 19; Supplementary Fig. S1C). These observations show that vector-valued dopamine feedback is crucial for the improvement in learning.

The improvement in learning is because of feedback alignment

We next wanted to further explore the conditions during which vector-valued feedback improves learning of the targets. First, we considered two additional alternative models: (i) the feedback is random, i.e. we still have vector-valued feedback, but shuffle the coefficients d_{jk} of Eq. 11 and (ii) we use the "ideal" feedback $d_{jk} = w^{\mathrm{StrSNr}}_{kj}(t)$. Note that locking the feedback weights to the feedforward in the second model means the feedback matrix is time-dependent. Both of these models perform similarly as our dopamine model (Fig. 2A). This suggests that the main criterion for the feedback to be effective is that feedback matrix $D = [d_{jk}]$ is non-degenerate.

With this hypothesis in mind, we tested varying the spatial scale λ of dopamine diffusion (see Eq. 10). Note that this spatial scale is measured as a fraction of the side of the cube (Fig. 1D). When $\lambda \ll 1$, almost no dopamine reaches any SPN from the varicosities, and the network reverts to the *No feedback* control model (Fig. 2B). On the other hand, when $\lambda \gg 1$, the dopamine from each varicosity covers the entire cube so that all SPNs effectively receive the sum of the dopamine released anywhere. This causes the network to revert to the *Homogeneous dopamine* control model. In between these two extremes, where dopamine scale is intermediate, there is a sweet spot where each SPN receives dopamine corresponding to a unique random linear projection of the 4-dimensional error (Fig. 2C). In this regime, the benefit of the feedback is the largest (Fig. 2B). Finally, we also tried to vary the number of varicosities N^{Var} in Eq. 10 and found that with a larger number N^{Var}, a smaller spatial scale λ becomes viable (Fig. 2D). This is also consistent with the creation of a non-degenerate D.

To illustrate the importance of striatofugal plasticity for learning, we simulated the network model with "lesioned" basal ganglia projections by either removing the plasticity (*static*) or clamping them to 0 (*absent*). As expected, when the plasticity of the striatofugal projection was turned off, no feedback alignment could take place and the striatal plasticity could not contribute to learning the targets (Fig. 2E, Str→SNr). Turning off plasticity of the striatal projections (corticostriatal, thalamostriatal, and striatostriatal) on the other hand has a more moderate impact. This is because even with all of them fixed, we can still have echo-state-like learning in the striatofugal weights (see the *No feedback* null model in Fig. 1G and 2A). Similarly as when fixing the weights, removing either the cortical or the thalamic projection does not change the eventual error much, as both projections play similar and mostly interchangeable roles in our model, whereas removing both silences the striatum completely and hence gives a very large error (Fig. 2E). See Supplementary Fig. S1D for a systematic investigation how removing or blocking plasticity on different

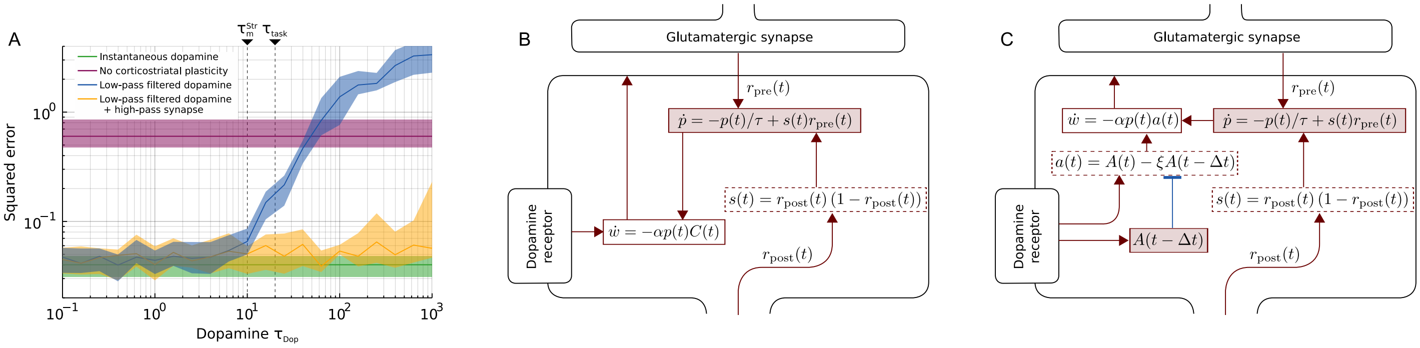

Figure 3: Slow dopamine dynamics. (A) When the dopamine signal is low-pass filtered with a time constant slower than the time constant of the target output, the error (blue) increases. For comparison, the errors of previous models with instantaneous (green) and absent (purple) dopamine are also shown (these do not depend on the τ_{Dop}). Finally, the orange line shows the error when using the extended synaptic dynamics shown in panel C. All errors are reported as the median over 25 trials. Shaded areas indicate first and third quartiles of the 25 trials. (B) The synaptic dynamics of the standard RFLO learning rule. (C) Adding an extra forward-inhibition motif to the synaptic dynamics to compensate for temporal smoothing of dopamine.

projections affect learning .

Fast synaptic dynamics can compensate for slow dopamine

So far we have assumed that dopamine is diffuse in space, but delivered instantly to the receiving SPNs. While this allows the synaptic weights to be updated correctly on every time step, it neglects the temporal dynamics of dopamine release, diffusion, reuptake, etc. A faithful quantitative model of these processes is beyond the scope of our abstract rate network, but it is nevertheless important to determine how dependent our dopamine-based learning rule is on the assumption of instantaneous dopamine release. We simplified all temporal dynamics of dopamine into a simple exponential low-pass filter with time constant τ_{DA}. That is, we changed the equation for dopamine concentration at striatal neuron j (Eq. 9) to

$$C_j'(t) = \int_0^t e^{\frac{t'-t}{\tau_{\mathrm{DA}}}} C_j(t')\mathrm{d}t' \tag{15}$$

The blue line in Fig. 3A shows the resulting error after 5000 trials for a range of values of

τ_{DA}. When τ_{DA} is much faster than the time constant of the task (here $\tau_{\mathrm{task}} = 20$ms), the error (Fig. 3A, blue line) is similar to the earlier, instantaneous dopamine model (Fig. 3A, green line). However, for slower τ_{DA} the error increases and even surpasses the null model with no striatal plasticity (Fig. 3A, purple line).

We next asked what would be needed to rescue the learning performance in face of slow dopamine dynamics. One possible solution would be that each synapse high-pass filters its local dopamine concentration. Such high-pass filtering can be done by assuming a feed-forward inhibition motif as the first step in the biochemical pathway triggered by dopamine (Fig. 3C). In the ideal case one biochemical node in the each synapse is tracking dopamine concentration one time step ago, and another calculates the difference

$$a(t) = A(t) - \xi A(t - \Delta t) \tag{16}$$

Furthermore, if we choose $A(t) = C_i'(t)/(1-\xi)$ and $\xi = e^{-\frac{\Delta t}{\tau_{\mathrm{DA}}}}$ we get $a(t) \approx C_i(t)$ and we can use the same RFLO learning rule as before but with $a(t)$ instead of $C_i(t)$ (compare Figs 3B and 3C).

We verified this idea by introducing the synapse model in Fig. 3C in all striatal synapses and then again plotting the error after 5000 trials. As predicted by the ideal choice of ξ, the error was consistently similar to instantaneous even for large τ_{DA} (Fig. 3A, yellow line).

Discussion

The broad, unspecific dopaminergic axonal projections have been argued to only allow for the transmission of a scalar homogeneous feedback signal (Chen and Goldberg, 2020). Here we provide a tenable counter-example of this view, even if speculative and highly idealized. We demonstrate that in a reduced model of a piece of basal ganglia that a heterogeneous, vector-valued feedback signal could in fact be transmitted by dopamine, even if the dopaminergic projections in the striatum are random. We have identified four key requirements for effective use of vector-valued dopamine, which also serve as predictions that can be verified experimentally:

1. At least one projection downstream of the striatum must be plastic.

2. The vector-valued error must be available to both the dopaminergic (here, SNc) and the readout (here, SNr) populations.

3. Striatal dopamine dynamics must be at least as fast as the targeted movements, or, alternatively, be high-pass filtered by feed-forward inhibition in the synaptic biochemical pathways.

4. The dopamine received by each SPN must be sufficiently independent, or, stated formally, the effective connectivity matrix arising from summing contribution of individual varicosities (Eq. 10) must not be degenerate.

Direct experimental evidence for any downstream plasticity in the basal ganglia (*Requirement 1*) is scarce, but in a recent study González-Rodríguez et al. (2021) showed that dopamine depletion in the SNr plays a larger role than striatal dopamine in producing motor deficits in Parkinson's disease. This is qualitatively consistent with the relative importance of striatonigral over corticostriatal plasticity in our model (Fig. 2F). However, we note that although we placed this plasticity in the striatofugal projection(s), it could in principle also be met by plasticity in the nigrothalamic or nigrocollicular projections.

A vector-valued error would likely appear as tunings to various motor and task variables in experimental animals (*Requirement 2*), especially in the phase before the animals are so overtrained their error is zero. Indeed, cells in the SNr (Fan et al., 2012; Barter et al., 2015; Tang et al., 2021) as well as the SNc (Howe and Dombeck, 2016; Dodson et al., 2016; Coddington and Dudman, 2018; Avvisati et al., 2022) respond to a plethora of behavioral and task variables. We deliberately excluded the details of how this error may be computed in the brain, but we speculate at least three possible algorithmic ways in which it could appear:

1. Bran region regions such as motor cortex or cerebellum could have a forward model of the world as well as the target, and thus, can directly compute the error and send it to the midbrain.

2. The brain could be wired as a set of hierarchical control loops, in which each loop provides the target for the level below (as proposed by Yin, 2014). Each such loop could stretch throughout the cortex and basal ganglia.

3. If the executed action has more variability than the command read out by the SNr, the *policy gradient theorem* (Sutton et al., 1999) states that the gradient for the update should be

$$r^{\mathrm{SNc}}(t) \propto \delta(t) \nabla \ln p \left(a(t) | r^{\mathrm{SNr}}(t) \right) \tag{17}$$

   where a is the vector-valued action taken, δ is the temporal difference (TD) error as predicted by a critic, and p is the probability density function of a. Note that this suggests that SNc cells should fire proportionally to both the TD error and to (the gradient of) some behavioural variables, which could explain why many SNc cells appear tuned to both (Parker et al., 2016; Lee et al., 2019). Lindsey and Litwin-Kumar (2022) have proposed dorsal striatum could make use of such a policy gradient, but nonetheless argue dopamine itself is a scalar proportional to the squared norm of the policy gradient.

There are several ways the brain could implement filters that allow extraction of faster fluctuations in dopamine concentration (*Requirement 3*). Our example in Fig. 3 is highly idealized and makes arguably unfair use of our idealized perfect exponential decay of the dopamine. In reality, the journey of dopamine molecule from a varicosity to a dopamine receptor depends on the local geometry and dopamine reuptake so that the dependence on both time and distance is most likely complicated and non-linear (although these effects might less pronounced at very short distances; Cragg and Rice, 2004; Liu et al., 2021). Nevertheless, evolution has had a good opportunity to tweak the biochemical pathways to compensate for these effects as far as permitted by the signal-to-noise ratio. Whether this is tenable in a realistic model of dopamine diffusion and biochemical cascades remains an open question, but we predict that there is at least one node in the biochemical cascade of dopamine-induced synaptic plasticity that is sensitive to fast fluctuations in local dopamine concentration.

The spatial frequency of the dopamine landscape in the striatum must be high enough so that even neighbouring SPNs do not sense the exact same dopamine concentration (*Requirement 4*). This can be achieved by having a short spatial constant of dopamine diffusion, and possibly compensating with a larger number of varicosities (Fig. 2D). Consistent with our model, Cragg and Rice (2004) estimated the diffusion distance of dopamine following release to a few microns.

The main goal of our work was to demonstrate that the broad and unspecific nigrostriatal dopaminergic projection can in principle transfer a usable vector-valued error to the striatum; our ambition was not to provide a complete biological account of the process. For this reason, there are many likely very important features of basal ganglia anatomy and physiology we did not include, for example dorsolateral/dorsomedial functional division in the striatum (Balleine and O'Doherty, 2010), the different roles of the matrix and the striosome (Bloem et al., 2017), axonally initiated dopaminergic release by cholinergic interneurons (Threlfell et al., 2012; Liu et al., 2022), saturating dopamine receptors (Liu et al., 2021), etc. Similarly, our primary goal was not to introduce a new algorithm for training recurrent neural networks; the network setup and plasticity rule is an application of the RFLO rule (Murray, 2019). Nevertheless, we show that the RFLO rule is applicable in a basal ganglia-like network with multiple inhibitory synapses and with our toy model of dopamine feedback, and propose vector-valued error feedback as a candidate functional role of dopamine.

Previous proposals for use of heterogeneous dopamine (Hamid et al., 2021; Lee et al., 2022) assume that the heterogeneous responses of dopaminergic cells are transmitted to the striatum through private parallel channels without any cross-talk. However, this not easily reconciled with functional and anatomical findings (see Introduction). Another proposed use of heterogeneous firing in the midbrain dopaminergic neurons is to support a distributional coding of value (Dabney et al., 2020). However, a distributional value code only explains different gains in the coding of the reward prediction error, not why the neurons respond to non-rewarded task variables. Nevertheless, it is entirely possible that the brain simultaneously employs a distributional value code (perhaps most strongly in the VTA) and a vector-valued error code (perhaps most strongly in the SNc).

In conclusion, we propose that the heterogeneous responses of dopamine cells seen by Fan et al. (2012); Howe and Dombeck (2016); Engelhard et al. (2019) and others represent a vector-valued error. By providing this type of error, the SNc supports the basal ganglia learning to select actions from a continuous action space in continuous time, thereby providing the animal with vital behavioral flexibility, control and adaptability.

Acknowledgements

We thank Dr. Moritz Weglage and Prof. Erik Fransén for helpful discussions. Partial funding from StratNeuro (to AK), Swedish Research Council (to AK) and Digital Futures (to AK) and Karolinska Institutet (KID doctoral funding to EW) is gratefully acknowledged.

References

Arbuthnott, G. W. and Wickens, J. (2007). Space, time and dopamine, *Trends in Neurosciences* **30**: 62–69.

Avvisati, R., Kaufmann, A.-K., Young, C. J., Portlock, G. E., Cancemi, S., Ponte Costa, R., Magill, P. J. and Dodson, P. D. (2022). Distributional coding of associative learning within projection-defined populations of midbrain dopamine neurons, *bioRxiv* .

Bahuguna, J., Weidel, P. and Morrison, A. (2019). Exploring the role of striatal D1 and D2 medium spiny neurons in action selection using a virtual robotic framework, *European Journal of Neuroscience* **49**(6): 737–753.

Balleine, B. W. and O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action, *Neuropsychopharmacology* **35**(1): 48–69.

Barter, J. W., Li, S., Sukharnikova, T., Rossi, M. A., Bartholomew, R. A. and Yin, H. H. (2015). Basal ganglia outputs map instantaneous position coordinates during behavior, *Journal of Neuroscience* **35**: 2703–2716.

Baston, C. and Ursino, M. (2015). A Biologically Inspired Computational Model of Basal Ganglia in Action Selection, *Computational Intelligence and Neuroscience* pp. 1–24.

Berthet, P., Hellgren-Kotaleski, J. and Lansner, A. (2012). Action selection performance of a reconfigurable basal ganglia inspired model with Hebbian–Bayesian Go-NoGo connectivity, *Frontiers in Behavioral Neuroscience* **6**.

Berthet, P., Lindahl, M., Tully, P. J., Hellgren-Kotaleski, J. and Lansner, A. (2016). Functional Relevance of Different Basal Ganglia Pathways Investigated in a Spiking Model with Reward Dependent Plasticity, *Frontiers in Neural Circuits* **10**.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2017). Julia: A fresh approach to numerical computing, *SIAM Review* **59**(1): 65–98.

Bloem, B., Huda, R., Sur, M. and Graybiel, A. M. (2017). Two-photon imaging in mice shows striosomes and matrix have overlapping but differential reinforcement-related responses, *eLife* **6**.

Bogacz, R., Martin Moraud, E., Abdi, A., Magill, P. J. and Baufreton, J. (2016). Properties of

Neurons in External Globus Pallidus Can Support Optimal Action Selection, *PLoS Computational Biology* **12**(7): 1–28.

Bolado-Gomez, R. and Gurney, K. (2013). A biologically plausible embodied model of action discovery, *Frontiers in Neurorobotics* **7**(MAR): 1–24.

Chen, R. and Goldberg, J. H. (2020). Actor-critic reinforcement learning in the songbird, *Current Opinion in Neurobiology* **65**: 1–9.

Coddington, L. T. and Dudman, J. T. (2018). The timing of action determines reward prediction signals in identified midbrain dopamine neurons, *Nature Neuroscience* **21**(11): 1563–1573.

Cragg, S. J. and Rice, M. E. (2004). DAncing past the DAT at a DA synapse, *Trends in Neurosciences* **27**(5): 270–277.

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

Dhawale, A. K., Wolff, S. B. E., Ko, R. and Ölveczky, B. P. (2021). The basal ganglia control the detailed kinematics of learned motor skills, *Nat. Neuro.* **24**: 1256.

Dodson, P. D., Dreyer, J. K., Jennings, K. A., Syed, E. C. J., Wade-Martins, R., Cragg, S. J., Bolam, J. P. and Magill, P. J. (2016). Representation of spontaneous movement by dopaminergic neurons is cell-type selective and disrupted in parkinsonism, *Proceedings of the National Academy of Sciences* **113**(15): E2180–E2188.

Dunovan, K., Vich, C., Clapp, M., Verstynen, T. and Rubin, J. (2019). Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making, *PLOS Computational Biology* **15**(5): e1006998.

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

Fan, D., Rossi, M. A. and Yin, H. H. (2012). Mechanisms of Action Selection and Timing in Substantia Nigra Neurons, *Journal of Neuroscience* **32**(16): 5534–5548.

Fisher, S. D., Robertson, P. B., Black, M. J., Redgrave, P., Sagar, M. A., Abraham, W. C. and Reynolds, J. N. (2017). Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo, *Nature Communications* **8**(1).

Foster, N. N., Barry, J., Korobkova, L., Garcia, L., Gao, L., Becerra, M., Sherafat, Y., Peng, B., Li, X., Choi, J.-H., Gou, L., Zingg, B., Azam, S., Lo, D., Khanjani, N., Zhang, B., Stanis, J., Bowman, I., Cotter, K., Cao, C., Yamashita, S., Tugangui, A., Li, A., Jiang, T., Jia, X., Feng, Z., Aquino, S., Mun, H.-S., Zhu, M., Santarelli, A., Benavidez, N. L., Song, M., Dan, G., Fayzullina, M., Ustrell, S., Boesen, T., Johnson, D. L., Xu, H., Bienkowski, M. S., Yang, X. W., Gong, H., Levine, M. S., Wickersham, I., Luo, Q., Hahn, J. D., Lim, B. K., Zhang, L. I., Cepeda, C., Hintiryan, H. and Dong, H.-W. (2021). The mouse cortico–basal ganglia–thalamic network, *Nature* **598**: 188–194.

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D. and Brea, J. (2018). Eligibility traces and plasticity on behavioral time scales: Experimental support of neohebbian three-factor learning rules, *Frontiers in Neural Circuits* **12**.

González-Rodríguez, P., Zampese, E., Stout, K. A., Guzman, J. N., Ilijic, E., Yang, B., Tkatch, T., Stavarache, M. A., Wokosin, D. L., Gao, L., Kaplitt, M. G., López-Barneo, J., Schumacker, P. T. and Surmeier, D. J. (2021). Disruption of mitochondrial complex i induces progressive parkinsonism, *Nature* p. 1476.

Gurney, K. N., Humphries, M. D. and Redgrave, P. (2015). A New Framework for Cortico-Striatal Plasticity: Behavioural Theory Meets In Vitro Data at the Reinforcement-Action Interface, *PLoS Biology* **13**(1): e1002034.

Hamid, A. A., Frank, M. J. and Moore, C. I. (2021). Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment, *Cell* **184**(10): 2733–2749.e16.

Howe, M. W. and Dombeck, D. A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward, *Nature* **535**(7613): 505.

Humphries, M. D., Stewart, R. D. and Gurney, K. N. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia, *Journal of Neuroscience* **26**(50): 12921–12942.

Hunnicutt, B. J., Jongbloets, B. C., Birdsong, W. T., Gertz, K. J., Zhong, H. and Mao, T. (2016). A comprehensive excitatory input map of the striatum reveals novel functional organization, *eLife* **5**: e19103.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of stdp and dopamine signaling, *Cerebral Cortex* **17**(10): 2443–2452.

Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, *Science* **304**(5667): 78–80.

Kim, H. G. R., Malik, A. N., Mikhael, J. G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., Gershman, S. J. and Uchida, N. (2020). A unified framework for dopamine signals across timescales, *Cell* **183**: 1600–1616.e25.

Klaus, A., Martins, G. J., Paixao, V. B., Zhou, P., Paninski, L. and Costa, R. M. (2017). The spatiotemporal organization of the striatum encodes action space, *Neuron* **95**: 1171–1180.e7.

Lee, R. S., Engelhard, B., Witten, I. B. and Daw, N. D. (2022). A vector reward prediction error model explains dopaminergic heterogeneity, *bioRxiv* p. 2022.02.28.482379.

Lee, R. S., Mattar, M. G., Parker, N. F., Witten, I. B. and Daw, N. D. (2019). Reward prediction error does not explain movement selectivity in dms-projecting dopamine neurons, *eLife* pp. 1–16.

Lillicrap, T. P., Cownden, D., Tweed, D. B. and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning, *Nat. Comm.* **7**(13276).

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. and Hinton, G. (2020). Backpropagation and the brain, *Nature Reviews Neuroscience* **21**: 335–346.

Lindsey, J. and Litwin-Kumar, A. (2022). Action-modulated midbrain dopamine activity arises from distributed control policies, *arXiv* pp. 1–17.

Liu, C., Cai, X., Ritzau-Jost, A., Kramer, P. F., Li, Y., Khaliq, Z. M., Hallermann, S. and Kaeser, P. S. (2022). An action potential initiation mechanism in distal axons for the control of dopamine release, *Science* **375**: 1387–1385.

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

Maass, W. and Markram, H. (2004). On the computational power of circuits of spiking neurons, *Journal of Computer and System Sciences* **69**(4): 593–616.

Matsuda, W., Furuta, T., Nakamura, K. C., Hioki, H., Fujiyama, F., Arai, R. and Kaneko, T. (2009). Single nigrostriatal dopaminergic neurons form widely spread and highly dense axonal arborizations in the neostriatum, *Journal of Neuroscience* **29**: 444–453.

Murray, J. M. (2019). Local online learning in recurrent networks with random feedback, *eLife* **8**: 1–25.

Niv, Y. and Langdon, A. (2016). Reinforcement learning with marr, *Current Opinion in Behavioral Sciences* **11**: 67–73.

Park, J., Coddington, L. T. and Dudman, J. T. (2020). Basal ganglia circuits for action specification, *Ann. Rev. Neuro.* **43**(1): 485.

Parker, N. F., Cameron, C. M., Taliaferro, J. P., Lee, J., Choi, J. Y., Davidson, T. J., Daw, N. D. and Witten, I. B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target, *Nature Neuroscience* **19**(6): 845–854.

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

Pozzi, I., Bohte, S. and Roelfsema, P. (2020). Attention-gated brain propagation: How the brain can implement reward-based error backpropagation, *in* H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (eds), *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., pp. 2516–2526.

Redgrave, P., Prescott, T. and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem?, *Neuroscience* **89**: 1009–1023.

Rueda-Orozco, P. E. and Robbe, D. (2015). The striatum multiplexes contextual and kinematic information to constrain motor habits execution, *Nat. Neurosci.* **18**(3): 453.

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

Stewart, T. C., Bekolay, T. and Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia, *Frontiers in Neuroscience* **6**(JAN): 1–14.

Sutton, R. S., McAllester, D., Singh, S. and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation, *in* S. Solla, T. Leen and K. Müller (eds), *Advances in Neural Information Processing Systems*, Vol. 12, MIT Press, pp. 1057–1063.

Tang, Y., Yang, H., Chen, X., Zhang, Z., Yao, X., Yin, X. and Guo, Z. V. (2021). Opposing regulation of short-term memory by basal ganglia direct and indirect pathways that are coactive during behavior, *bioRxiv* .

Threlfell, S., Lalic, T., Platt, N. J., Jennings, K. A., Deisseroth, K. and Cragg, S. J. (2012). Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons, *Neuron* **75**: 58–64.

Weglage, M., Wärnberg, E., Lazaridis, I., Calvigioni, D., Tzortzi, O. and Meletis, K. (2021). Complete representation of action space and value in all dorsal striatal pathways, *Cell Reports* **36**(4): 109437.

Yin, H. H. (2014). Action, time and the basal ganglia, *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**(1637).

Methods

The dynamics of neurons, network structure and learning rule is already described in the Results section. Here we describe only the technical details needed to run the simulations.

Network simulations

The network was simulated in a custom simulator written in Julia (Bezanson et al., 2017). The dynamics were simulated with forward-Euler with dt=1ms.

Simulations consisted of multiple trials concatenated after each other without any reset of the network in between. The current time in the current trial was signaled to the network by setting the thalamic firing rates to

$$r_i^{\mathrm{Thal}}(t) = \phi\left(A_i \cos\frac{2\pi t}{T} + B_i \sin\frac{2\pi t}{T}\right) \tag{18}$$

where $T = 200\mathrm{ms}$ is the duration of a single trial, and A_i and B_i are constants drawn randomly from a circle with radius 4 (i.e. $A_i^2 + B_i^2 = 4^2$ for all i). ϕ is the logistic transfer function (Eq. 2).

Initializing the weights

For each pair of cells in each projection, there was a fixed probability (Fig. 1B) of a synapse being inserted. If a synapse was inserted, its weight was drawn from a uniform distribution $[0, w_{\max}/\sqrt{N_{\mathrm{post}}}]$ (see Table S1) and then multiplied by -1 for the inhibitory projections. The weights in Table S1 were chosen for the network to have close to chaotic trajectories before training.

After all synapses were created, the sum of the weight of all incoming synapses was calculated for each neuron. If this sum was greater than 0, all the inhibitory synapses were slightly increased so that the new sum was exactly 0. Conversely, if the sum was less than 0, all the excitatory weights were slightly increased to reach sum 0. This ensured that each neuron had roughly balanced excitation and inhibition, which in turn created rich dynamics from the start.

Target signals

The targets were drawn from a Gaussian process with mean 0.5 and variance given by

$$\sigma(t_1, t_2) = \sqrt{0.15}\exp\left(-\frac{\delta(t_1, t_2)^2}{\tau_{\text{task}}^2}\right) \tag{19}$$

where $\delta(t_1, t_2)$ is the smallest difference between t_1 and t_2 when including wrap-around, i.e.

$$\delta(t_1, t_2) = \min(|t_1 - t_2|, |t_1 - t_2 + T|, |t_1 - t_2 - T|) \tag{20}$$

where $T = 200\text{ms}$ is the duration of a single trial. The periodic kernel is to avoid discontinuities when running consecutive trials without resetting the network. For all experiments in the main figures, $\tau_{\text{task}} = 20\text{ms}$.

pre	post	$w_{\max}$
Cortex (exc)	Cortex(exc)	25
Cortex (exc)	Cortex(inh)	25
Cortex (inh)	Cortex(exc)	25
Cortex (inh)	Cortex(inh)	25
Cortex (exc)	Striatum (dSPN)	25
Cortex (exc)	Striatum (iSPN)	25
Striatum (dSPN)	Striatum (dSPN)	25
Striatum (dSPN)	Striatum (iSPN)	25
Striatum (iSPN)	Striatum (dSPN)	25
Striatum (iSPN)	Striatum (iSPN)	25
Striatum (dSPN)	SNr/GPi	5
Striatum (iSPN)	SNr/GPi	5
Thalamus	Cortex (exc)	50
Thalamus	Cortex (inh)	50
Thalamus	Striatum (dSPN)	30
Thalamus	Striatum (iSPN)	30

Table S1: Initial synaptic strength. Note that weights for inhibitory connections were multiplied by -1 after they are drawn.

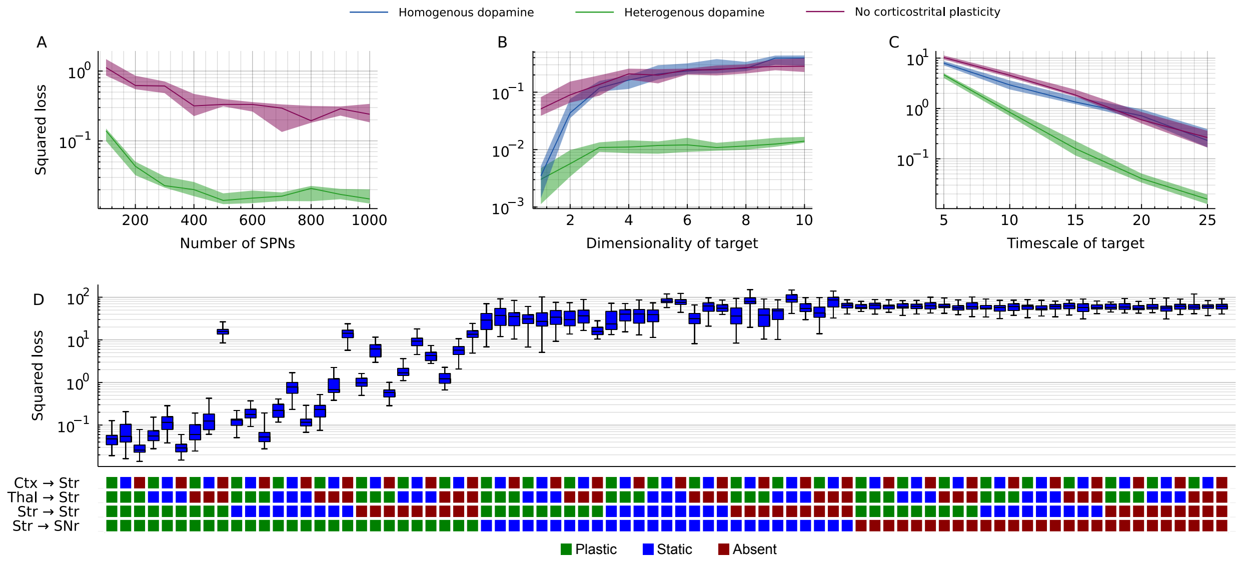

Figure S1: Influence of each projection to the loss. (A) Mean squared error after 5000 training trials for different sizes of the striatal population. (B) Mean squared error for increasing values of $N^{\mathrm{SNr}} = N^{\mathrm{SNc}} = d$. (C) Mean squared error for increasing values of τ_{task}. (D) The network model was run with impairments to some of the connections. Either the plasticity was removed so that the synaptic weights of the connection were fixed to their starting weights ("static"), or the synapses were removed altogether ("absent"). Squared error is measured as the error after 5000 trials. Boxplots show median and quartiles across 25 runs; whiskers indicate min and max of the 25 runs.

Derivation of the plasticity rule

We begin by restating the network equations. First, we write out the three inputs (cortex, thalamus and recurrent striatum) as

$$\begin{aligned} I_j(t') &= \sum_{i=1}^{N^{\mathrm{Ctx}}} w_{ji}^{\mathrm{CtxStr}}(t') r_i^{\mathrm{Ctx}}(t') + \sum_{m=1}^{N^{\mathrm{Thal}}} w_{jm}^{\mathrm{ThalStr}}(t') r_m^{\mathrm{Thal}}(t') \qquad \text{(S1)} \\ &\quad + \sum_{j'=1}^{N^{\mathrm{Str}}} w_{jj'}^{\mathrm{StrStr}}(t') r_{j'}^{\mathrm{Str}}(t') \qquad \text{(S2)} \end{aligned}$$

Then we rewrite the membrane equation as an integral

$$\begin{aligned} V_j^{\mathrm{Str}}(t) &= \int_0^t \exp\left(\frac{t'-t}{\tau^{Str}}\right) I_j(t')\mathrm{d}t' \qquad \text{(S3)} \\ r_j^{\mathrm{Str}}(t) &= \phi\left(V_j^{\mathrm{Str}}(t)\right) \qquad \text{(S4)} \end{aligned}$$

Since we treat SNr as a read-out layer, we let $\tau^{\mathrm{SNr}} \to 0$ so that

$$\begin{aligned} V_k^{\mathrm{SNr}}(t) &= B^{\mathrm{SNr}} + \sum_{j=1}^{N^{\mathrm{Str}}} w_{jk}^{StrSNr}(t) r_j^{\mathrm{Str}}(t) \qquad \text{(S5)} \\ r_k^{\mathrm{SNr}}(t) &= \phi\left(V_k^{\mathrm{SNr}}(t)\right) \qquad \text{(S6)} \end{aligned}$$

The instantaneous loss is

$$\ell(t) = \frac{1}{2}\sum_{k=1}^{d}\left(r_k^{\mathrm{SNr}}(t) - T_k(t)\right)^2 \qquad \text{(S7)}$$

where $T(t) \in \mathbb{R}^d$ is the target output at time t. To greedily (i.e. not considering earlier or later losses) minimize $\ell(t)$ (see Murray, 2019) we want to have

$$\frac{\mathrm{d}w_{ji}^{\mathrm{CtxStr}}(t)}{\mathrm{d}t} \propto -\frac{\partial \ell(t)}{\partial w_{ji}^{Ctx}(t)} \qquad \text{(S8)}$$

Expanding this taking a partial derivative where $r_{j'}(t)$ is fixed for $j' \neq j$:

$$\begin{aligned}
-\frac{\partial \ell(t)}{\partial w_{ji}^{\text{CtxStr}}(t)} &= -\sum_{k=1}^{d} \left(r_k^{\text{SNr}}(t) - T_k(t)\right) \frac{\partial r_k^{\text{SNr}}(t)}{\partial w_{ji}^{\text{CtxStr}}(t)} && \text{(S9)}\\
&= -\sum_{k=1}^{d} \underbrace{\left(r_k^{\text{SNr}}(t) - T_k(t)\right) \phi'\left(V_k^{\text{SNr}}(t)\right)}_{\epsilon_k(t)} \frac{\partial v_k^{\text{SNr}}(t)}{\partial w_{ji}^{\text{CtxStr}}(t)} && \text{(S10)}\\
&= -\underbrace{\sum_{k=1}^{d} \epsilon_k(t) w_{kj}^{\text{StrSNr}}(t)}_{\gamma_j(t)} \frac{\partial r_j^{\text{Str}}(t)}{\partial w_{ji}^{\text{CtxStr}}(t)} && \text{(S11)}\\
&= -\gamma_j(t)\phi'\left(V_j^{\text{Str}}(t)\right) \frac{\partial v_j^{\text{Str}}(t)}{\partial w_{ji}^{\text{CtxStr}}(t)} && \text{(S12)}\\
&= -\gamma_j(t) \int_0^t \exp\left(\frac{t'-t}{\tau^{\text{Str}}}\right) \phi'\left(V_j^{\text{Str}}(t')\right) \frac{\partial I_j(t')}{\partial w_{ji}^{\text{CtxStr}}(t)} \mathrm{d}t' && \text{(S13)}\\
&= -\gamma_j(t) \underbrace{\int_0^t \exp\left(\frac{t'-t}{\tau^{\text{Str}}}\right) \phi'\left(V_j^{\text{Str}}(t')\right) r_i^{\text{Ctx}}(t')\mathrm{d}t'}_{p_{ji}(t)} && \text{(S14)}\\
&= -\gamma_j(t) p_{ji}(t) && \text{(S15)}
\end{aligned}$$

Furthermore, note that

$$\begin{aligned}
\phi'(V) &= -\frac{-e^{-V+b}}{\left(1+e^{-V+b}\right)^2} = \frac{1}{1+e^{-V+b}} \frac{\left(1+e^{-V+b}\right)-1}{1+e^{-V+b}} && \text{(S16)}\\
&= \frac{1}{1+e^{-V+b}}\left(1 - \frac{1}{1+e^{-V+b}}\right) = \phi(V)\left(1-\phi(V)\right) && \text{(S17)}
\end{aligned}$$

In summary, we get the following update rule for the corticostriatal synaptic weights

$$\begin{aligned}
\epsilon_k(t) &= \left(r_k^{\text{SNr}}(t) - T_k(t)\right) r_k^{\text{SNr}}(t) \left(1 - r_k^{\text{SNr}}(t)\right) && \text{(S18)}\\
\gamma_j(t) &= \sum_{k=1}^{d} \epsilon_k(t) w_{kj}(t) && \text{(S19)}\\
\tau^{\text{Str}} \frac{\mathrm{d}p_{ji}}{\mathrm{d}t} &= -p_{ji}(t) + r_j^{\text{Str}}(t)\left(1 - r_j^{\text{Str}}(t)\right) r_i^{\text{Ctx}}(t) && \text{(S20)}\\
\frac{\mathrm{d}w_{ji}^{\text{CtxStr}}(t)}{\mathrm{d}t} &= -\alpha\gamma_j(t)p_{ji}(t) && \text{(S21)}
\end{aligned}$$

The update rules for thalamostriatal (w^{ThalStr}) and striatostriatal (w^{StrStr}) have the same form and are derived in the same way. The striatofugal weight plasticity is

$$\frac{\mathrm{d}w_{kj}(t)}{\mathrm{d}t} = -\beta\epsilon_k(t)r_j^{\mathrm{Str}}(t) \tag{S22}$$

with the same ϵ_k as above.