Reinforcement Learning in a Spiking Neural Model of Striatum Plasticity

Álvaro González-Redondo, Jesús Garrido, Francisco Naveros Arrabal, Jeanette Hellgren Kotaleski, Sten Grillner, Eduardo Ros


bKungliga Tekniska Högskolan, SE-100 44, Stockholm, Sweden

Abstract

The basal ganglia (BG), and more specifically the striatum, have long been proposed to play an essential role in action-selection based on a reinforcement learning (RL) paradigm. However, some recent findings, such as striatal spike-timing-dependent plasticity (STDP) or striatal lateral connectivity, require further research and modelling as their respective roles are still not well understood. Theoretical models of spiking neurons with homeostatic mechanisms, lateral connectivity, and reward-modulated STDP have demonstrated a remarkable capability to learn sensorial patterns that statistically correlate with a rewarding signal. In this article, we implement a functional and biologically inspired network model of the striatum, where learning is based on a previously proposed learning rule called spike-timing-dependent eligibility (STDE), which captures important experimental features in the striatum. The proposed computational model can recognize complex input patterns and consistently choose rewarded actions to respond to such sensorial inputs. Moreover, we assess the role different neuronal and network features, such as homeostatic mechanisms and lateral inhibitory connections, play in action-selection with the proposed model. The homeostatic mechanisms make learning more robust (in terms of suitable parameters) and facilitate recovery after rewarding policy swapping, while lateral inhibitory connections are important when multiple input patterns are associated with the same rewarded action. Finally, according to our simulations, the optimal delay between the action and the dopaminergic feedback is obtained around 300ms, as demonstrated in previous studies of RL and in biological studies.

Keywords: Striatum, Reinforcement learning, Spiking neural network, Dopamine, Eligibility trace, Spike-timing-dependent plasticity

1. Introduction

Animals learn to choose actions among many options by trial and error, thanks to the feedback provided by sparse and delayed rewards. Reinforcement learning (RL) serves as a theoretical framework for an agent, a system that acts based on received feedback, to learn to map situations to actions. This state-action mapping aims to maximize the performance of actions, mainly (but not exclusively) considering how rewarding or punishing the consequences of the actions are (Sutton et al., 1992). The basal ganglia (BG), a group of forebrain nuclei, are posited to play a critical role in action-selection based on RL (Grillner et al., 2005; Graybiel, 1998; Hikosaka et al., 2000).
However, the roles of recent findings, such as striatal spike-timing-dependent plasticity (STDP) models and striatal asymmetrical lateral connectivity, remain unclear. Investigating these interactions could improve our comprehension of the BG’s role in RL, potentially leading to the development of more efficient bio-inspired reinforcement learning agents.

This study aims to explore the impact of homeostatic mechanisms and asymmetric lateral inhibitory connections on action-selection in the striatum. We use the RL framework to gain insights into the neural basis of decision-making and contribute to more biologically plausible basal ganglia models. Our model stands out from previous models in several ways: it does not require a critic or extra circuitry for a temporal difference signal, thereby simplifying the model and reducing computational complexity; additionally, it employs a spiking neural network with spike-time pattern representation that adapts well to varying pattern complexities in the pattern classification layer.

We propose a functional, biologically inspired striatum network model that incorporates dopamine-modulated spike-timing-dependent eligibility (STDE, Gurney et al. (2015)) and asymmetric lateral connectivity (Burke et al., 2017). This model improves upon existing striatum models by integrating homeostatic mechanisms, asymmetric lateral inhibitory connections, and the STDE learning rule, capturing essential experimental features found in the striatum.

In this article, we present a model that effectively processes complex input patterns in the context of reinforcement learning. We conduct multiple analyses to assess the interaction between the learning rule, homeostatic mechanisms, and lateral inhibitory connectivity patterns. By incorporating these elements, we strive to develop a comprehensive and biologically plausible striatum model that offers valuable insights. Our study examines the individual and combined effects of these factors, shedding light on the unique topology of the striatum network and its role in reinforcement learning tasks.

The main contributions and findings of this work are:

- A functional and biologically inspired network model of the striatum that integrates dopamine-modulated STDE, homeostatic mechanisms, and asymmetric lateral inhibitory connectivity, providing a more comprehensive and biologically plausible representation of the striatum’s function.
- Analysis of the role of homeostatic mechanisms in making learning more robust and facilitating recovery after rewarding policy swapping.
- Investigation of the importance of lateral inhibitory connections when multiple input patterns are associated with the same rewarded action.
- The use of a spiking neural network with spike-time pattern representation that scales well with different pattern complexity, making the model suitable for a wide range of reinforcement learning tasks.
- Demonstration that the optimal delay between action and dopaminergic feedback occurs around 300 ms, which is consistent with previous reinforcement learning and biological studies.
- A model that does not require a critic, simplifying the learning process and reducing the need for additional circuitry.

1.1. Basal Ganglia Circuitry and Striatal Connectivity in Decision Making

The BG network is composed of several structures, grouped in inputs [being the striatum the best known, and populated by medium spiny neurons (MSN)], intermediate layers [the
external segment of the globus pallidus (GPe), and the substantia nigra pars compacta (SNc) and output [substantia nigra pars reticulata (SNr)]. The information flows segregated through the BG circuits (DeLong et al., 1985; Parent and Hazrati, 1995). It has been proposed that the BG process a large number of cognitive streams or channels in parallel (Gurney et al., 2001), each of them representing a feasible action to be performed (Suryanarayana et al., 2019). According to recent research, this segregation through the entire cortico-BG-thalamic loop shows a very high specificity, down to almost neuron-to-neuron level (Hunicutt et al., 2016; Foster et al., 2021). Thus, it seems feasible to impact behavior at different levels of detail. However, with the current biological evidence it is not exactly known how the activation of a channel maps to the corresponding behavior and we just assume here that these channels involve a decision making process.

The striatum, as the primary input of the basal ganglia, connects to the SNr via direct and indirect pathways, which are traditionally thought to promote and inhibit behavior, respectively. Each pathway crosses the striatum through different subpopulations of MSNs, expressing dopamine receptors D1 for the direct pathway and D2 for the indirect pathway. Recent genetic and optical studies on striatal circuits have allowed for testing classical ideas about the functioning of this system, but new models are needed to better understand the role of the striatum in learning and decision-making (Cox and Witten, 2019).

1.2. Spiking Neural Networks: Learning, Reward Modulation, and Striatal Connectivity

In recent decades, the use of biologically plausible computational models composed of spiking neurons able to learn a target function has demonstrated being increasingly successful (Taherkhani et al., 2020; Tavanaei et al., 2019). These models use discrete-time events (spikes) to compute and transmit information. As the specific timing of spikes carry relevant information in many biological contexts, these models are useful to understand how the brain computes at the neuronal description level. Combined with the use of local learning rules, these models can be implemented in highly efficient, low-power, neuromorphic hardware (Rajendran et al., 2019). Within this framework, learning from past experiences can be achieved using the STDP learning rule, a synaptic model featuring weight adaptation that has been observed in both biological systems (Levy and Steward, 1983) and the BG (Fino and Venance, 2010). The STDP also was demonstrated to be competitive in unsupervised learning of complex pattern recognition tasks (Masquelier et al., 2009; Garrido et al., 2016). The complexity of the patterns comes from their statistically equivalent activity level and from being immersed within a noisy stream of hundreds or thousands of inputs. These studies shown that an oscillatory stream of inputs reaching a population of spiking neurons enables a target post-synaptic neuron equipped with STDP to detect and recognize the presence of repetitive current patterns (Masquelier et al., 2009). The added oscillatory drive performs a current-to-phase conversion: the neurons that receive the most potent static current will fire the first during the oscillation cycle. This mechanism locks the phase of the spike time, facilitating the recognition of the previously presented patterns.

However, STDP-based learning systems tend to use statistical correlations to strengthen synaptic connections, resulting in the selection of the most frequent patterns at the expense of the most rewarding (Garrido et al., 2016). Thus, the STDP rule can be modified to drive the learning of patterns that statistically correlate with a reward signal (Izhikevich, 2007; Legenstein et al., 2008). In biological systems, unexpected rewards signal relevant stimuli during learning by releasing dopamine (DA). More
specifically, the reward signal is linked to the phasic modulation of dopaminergic neurons in the SNc and ventral tegmental area (Schultz, 2010), that sends reinforcement signals to the striatal neurons. These rewards do not need to happen instantly after the relevant stimulus; they can be delayed seconds, resulting in the distal reward and temporal credit assignment problems. In Izhikevich (2007); Legenstein et al. (2008), the authors suggest a reward-modulated STDP rule that enables a neuron to detect rewarded input patterns lasting milliseconds, even if the reward is delayed by seconds, by using the so-called eligibility trace. Also, based on the eligibility trace, Gurney et al. (2015) developed a synaptic learning rule called Spike-Timing-Dependent Eligibility (STDE) based on physiological data that captures many features found in the biological MSN of the basal ganglia. This model is more flexible than the previous STDP-like rules as different learning kernels can be used depending on the amount and type (reward or punishment) of reinforcement received. Although the authors did not include some important BG features like the GPe nucleus or a cortico-striatal loop, their model successfully learned to select an action channel driven by stronger cortical input, based only on the timing of the input and the reward signal.

Another relevant feature of the striatum is its connectivity. Burke et al. (2017) proposed a model of asymmetric lateral connectivity in the striatum that tries to explain how different clusters of striatal neurons interact and which role they play in information processing. This model accounts for the in vivo phenomenon of co-activation of sub-populations of D1 or D2 MSNs, which seems paradoxical as each subpopulation projects to behaviorally opposite pathways (direct and indirect, respectively). This structured connectivity pattern is determined by lateral inhibition between neurons that belong to the same channel and between neurons within different channels but accounting for the same receptor type (D1 or D2). The authors also include asymmetrical connections with more intensive intra-channel inhibition from D2 to D1 neurons than in the opposite direction. This pattern resulted in synchronized phase-dependent activation between MSN D1 and D2 neuron groups that belong to different channels.

1.3. Contribution

All the previous ideas are important pieces of the process of goal-oriented learning but further research is required as their respective roles and how they complement each other are still not well understood. The combination of the STDE rule within a network with asymmetrically structured lateral inhibition has not been studied before, and some relevant conclusions emerge from this specific study. In this article, we design and study a functional and biologically inspired model of the striatum. Our approach is based on spike time representation of complex input patterns and integrates dopamine modulated STDE and asymmetric lateral connectivity, among other mechanisms. This model learns to select the most rewarding action to complex input stimuli through RL. The proposed model has been demonstrated to be capable of recognizing input patterns relevant for the task and consistently choosing rewarded actions in response to that input. We performed numerous analyses to measure and better understand the interaction between the learning rule with homeostatic mechanisms and the lateral inhibitory connectivity patterns. By measuring the single and combined effects of these factors in the learning process, we want to shed light on how the particular topology of the striatum network facilitates the resolution of RL tasks.

2. Methods

Aiming to implement a RL framework in a biologically plausible striatum model, we started designing a task where the agent has to learn how
to map different input patterns into actions based on the reward signal delivered by the environment. We implemented a network model of the striatum capable of learning this task. This system behaves like a RL agent and can solve action-selection tasks.

The methods section is structured as follows: we first define the neuron and synapse models, input pattern generation, and networks structures used in our experiments. Then we describe the experimental design used with the network model and how we measure its learning capability. In Supplementary Materials we also explain both a previous experiment and a simpler model we made to test the viability of the combination of oscillatory inputs, STDE and homeostatic rules that we employed in the final network model.

2.1. Computational models

2.1.1. Neuron models

We used conductance-based versions of the Leaky-Integrate and Fire (LIF) neuron model (Gerstner and Kistler, 2002) as it is computationally efficient and captures certain biological plausibility. We use this model in every layer of the network, but with different parameters. We classify the neuron types according to the layer they belong to: cortical neurons for the input, striatal neurons (divided in two subpopulations according to which DA receptor express, D1 or D2) for the learning layer, and action neurons for the output. There is also a dopaminergic neuron that receives the rewards and punishments. The parameters used for each type were manually tuned to obtain reasonable firing rates. For the cortical neurons we used a number of spikes per input cycle (with 8 cycles per second) close to Masquelier et al. (2009) and Garrido et al. (2016) (see details about the input protocol in section 2.1.2). For the striatal neurons, we tuned the parameters to obtain a mean firing rate of around one spike per second to be within biological ranges (Miller et al., 2008) but with activity peaks of two or three spikes per input cycle (16-24 spikes per second). The action neurons (an integrative population that outputs the agent’s behavior) are tuned to fire every input cycle if they receive enough stimulation from its channel (at least two more spikes from D1 neurons than D2 neurons each cycle). The dopamine neuron was tuned to have a firing range from 50 to 350 spikes per second, with these unrealistic values chosen for performance (instead of simulating a bigger dopaminergic population). The parameters used for each neuron type are shown at supplementary table [1].

2.1.2. Input and oscillatory drive

In the input generation procedure (Masquelier et al., 2009; Garrido et al., 2016) we consider a trial as a segment of time of the simulation where we present some input stimuli to the network. The length of each trial is taken from a uniform random distribution between 100 and 500ms. An input stimulus represents a combination of 2000 input current values conveyed one-to-one to a set of cortical neurons of the same size (Fig. [8A]). An input pattern is a combination of current values which target precisely the same cortical neurons every time the input pattern is presented for the entire simulation. For every time bin, one or no pattern is presented. Only half of the cortical neurons (1000) are pattern-specific when presenting a specific pattern, while the other half receives random current values. Two thousand current-based LIF cortical neurons transform the input current levels into spike activity. These neurons have a firing rate between 8 to 40 spikes per second due to the sum of the input current values (ranged from 87% to 110% of the cortical neuron rheobase currents) and an oscillatory drive at 8Hz feeding these neurons (with an amplitude of 15% of the rheobase current of the cortical neurons). This oscillatory drive turns the input encoding from analogical signal to phase-of-firing coding (Masquelier et al., 2009) by locking the phase of the cortical spikes within the oscillatory drive, as shown in Fig. [8B]. By using these parameters, the cortical neurons fire between 1 and 5 spikes per cycle.

2.1.3. Spike-Timing-Dependent Eligibility (STDE) learning rule

We implemented a version of the STDE learning rule (Gurney et al., 2015), a phenomenological model of synaptic plasticity. This rule is similar to STDP, but the kernel constants are DA-dependant (that is, different values are defined for low DA and high DA values, and interpolated for DA values in-between, as shown in Fig. [6] and Supplementary Fig. [9Aii and Aii]). STDE is derived from in vitro data and predicts changes in direct and indirect pathways during the learning and extinction of sin-
gle actions. Throughout, we used the following parameters and procedures unless we specified otherwise. The kernel shape is defined by the parameters $k_{SPK}^{DA}$ with $SPK \in \{+,-\}$ being the spike order pre-post for applying $k_{DA}^{+}$ and post-pre for applying $k_{DA}^{-}$, respectively, and $DA \in \{hi,lo\}$ being the high- or low-DA cases, resulting in four parameters in total: $k_{hi}^{+}$, $k_{lo}^{+}$, $k_{hi}^{-}$ and $k_{lo}^{-}$. We obtained these learning kernel constant values by hand-tuning for both MSN D1 and D2 cases (see Supplementary Fig. 9 and supplementary table 2). As in the classic STDP learning rule, the weight variation in STDE is calculated for every pair of pre- and post-synaptic spikes and decays exponentially with the time difference between the spikes (Fig. 1). We use time constants $\tau = 32$ ms and the weights values are clipped to $[0,0.075]$.

Our implementation of STDE uses elegibility traces that decay exponentially to store the potential weight changes, similarly to Izhikevich (2007). Following Gurney et al. (2015) we have two different elegibility traces per synapse, $c^{+}$ and $c^{-}$ for spike pairs with positive and negative timing respectively, updated for every pair of pre- and post-synaptic spikes at times $t_j$ and $t_i$ as in equations (1) and (2):

\[
\delta c^{+} = (\alpha k_{hi}^{+} + \overline{\alpha} k_{lo}^{+}) \cdot e^{\frac{t_j-t_i}{\zeta_{eli}}} \text{ if } t_j \leq t_i \tag{1}
\]

\[
\delta c^{-} = (\alpha k_{hi}^{-} + \overline{\alpha} k_{lo}^{-}) \cdot e^{-\frac{t_j-t_i}{\zeta_{eli}}} \text{ if } t_j > t_i \tag{2}
\]

with $\overline{\alpha} = 1 - \alpha$, $\alpha$ been a value dependent of DA that we define in equation 3 and $\zeta_{eli}$ been the eligibility trace time constant with a value twice the length of the mean reward delay. Overall plastic change at a single synapse is then the sum of contributions from both $c^{+}$ and $c^{-}$, scaled by a learning rate factor $\eta = 0.002$.

The level of DA in the system is determined by one neuron that fires at high (and unrealistic) rates for computational simplicity, representing a population of neurons from the SNC. This neuron fires spontaneously at a baseline frequency of 200Hz. The environment (i.e., the application of rewarding policies during the experiment) injects positive (or negative) current in the dopaminergic neuron when rewards (or punishments) are applied to the model, resulting in the firing rate of this neuron ranging between 50Hz and 350Hz. All plastic synapses share a global DA level $d$ that decays exponentially with temporal constant $\tau_{da} = 20$ ms. For each spike emit-
Our implementation of STDE uses the linear mixing function \( \alpha \) in equation (3), clipped to \([0, 1]\), to smoothly morph between kernels with low and high DA:

\[
\alpha = \frac{d - d_{\min}}{d_{\max} - d_{\min}}
\]

where \( d_{\min} \) and \( d_{\max} \) are the minimum and maximum values of DA considered. We use this equation for computational efficiency instead of the Naka-Rushton function used in Gurney et al. (2015) (the authors also noted that this is not a requirement, as long as the mixing function was increasingly monotonic and saturating). The function is bounded to values of DA firing rate between 50 and 350Hz, with the baseline at 200Hz.

2.1.4. Homeostatic mechanisms

During learning, in some cases, the neurons can stop firing indefinitely due to a learning history leading to the wrong parameters. Neuron activity can also die by sudden changes in the reward policy, leaving the state of the synaptic weights ill (not representing any stimuli and not getting enough input to fire by chance). To recover neurons from this state, we added two different homeostatic mechanisms, one at the synaptic level and one at the neuron level. Although one or the other is enough to avoid the ill-states, we saw in our tests that we recovered faster and more reliably by using both.

The synapses implementing the STDE included a non-Hebbian strengthening in response to every pre-synaptic spike. For each arriving spike, the synaptic weight increases by \( C_{\text{pre}} = \eta \cdot 4 \cdot 10^{-4} \). This non-Hebbian strengthening is added to enable the recovery of low-bounded synapses (e.g., after a rewarding policy switch). Although the rewarding policy does not change in the network experiment, this homeostatic mechanism also benefits the complete network model learning (more details in section 5.2.2 and Supplementary Fig. 14).

In order to avoid neurons to become permanently silent during learning, we include adaptive thresholds to our neuron models based on Galindo et al. (2020) according to the following equation:

\[
\frac{dV_{\text{th}}}{dt} = -\frac{V_{\text{th}} - E_{\text{leak}}}{\tau_{\text{th}}} \tag{4}
\]

where \( V_{\text{th}} \) represents the firing threshold at the current time, \( E_{\text{leak}} \) is the resting potential of the neuron, and \( \tau_{\text{th}} \) is the adaptive threshold time constant. According to equation (4) in the absence of action potentials, the threshold progressively decreases towards the resting potential, facilitating neuron firing. When the neuron spikes, the firing threshold increases a fixed step proportional to the constant \( C_{\text{th}} \) as indicated in equation (5) making neuron firing more sparse.

\[
\delta V_{\text{th}} = \frac{C_{\text{th}}}{\tau_{\text{th}}} \tag{5}
\]

2.1.5. Striatum network model

The network model of the striatum (Fig. 3A) contains two channels (channel A and channel B, each one representing a possible action). Every channel contains two same-sized subpopulations (D1 and D2 neurons, respectively) of striatal-like neurons (in total, 16 neurons per channel) and one so-called action neuron that integrates excitatory activity from D1 neurons and inhibitory activity from D2 neurons. This design simplifies the biological substrate in which all MSN are inhibitory, but we implemented the network computation by considering the net effect of each neuron type on behavior. Biological MSN D1 neurons inhibit SNr, which promotes behavior, and MSN D2 neurons inhibit GPe, which, in turn, inhibit SNr with the total effect of decreasing behavior (Fig. 3A).

Our striatum model implements lateral inhibition within each MSN D1 population, within each MSN D2 population, between MSN D1 and MSN D2 populations within the same channel, and between the MSN populations associated with different action channels. Inspired by Burke et al. (2017), we used an asymmetrical structured pattern of connectivity (Fig. 5E in Burke et al. (2017), and adapted here in Fig. 3). Following this connectivity pattern, we added lateral inhibition between neurons that belong to the same channel and between those that belong to different channels but use the same dopaminergic receptor D1 or D2 (with stronger inhibition from D2 to D1 neurons than in the opposite direction). Since the small size of the network under study and the small weight of the D1 to D2 MSN connections, the overall contribution of these connections was negligible, so we decided not to include them in our simulations as we see no significant impact on previous simulations.

The environment generates the reinforcement signal based on comparing the chosen and the expected action and then delivers it to the dopamin-
Figure 2: Connectivity pattern used for the lateral inhibition, inspired on Burke et al. (2017). Two channels (action A and action B) are shown, each with two populations of D1 and D2 MSN.

Dopaminergic neuron. Rewards are excitatory, and punishments are inhibitory inputs to this neuron. The dopaminergic modulatory signal is global and delivered to every STDE connection from cortical layer to striatal neurons (Fig. 3A). It is important to note that this model does not implement a critic (commonly used in actor-critic frameworks (Sutton et al., 1992)), so there is no reward prediction error signal.

2.2. Experimental design

We first validated the proposed learning mechanisms with a simpler network model of only one neuron and a easier experimental task, as can be seen in Supplementary Methods 5.1 and Supplementary Results 5.2.

The action-selection task used to test the model (Fig. 3B) works as follows: the agent has two possible actions to choose, A or B. An action is selected if the activity balance of its D1 and D2 neurons is biased to D1 in two spikes at least in one cycle (making the corresponding action neuron spike). The agent can do none, both, or any of them at a time. The input stream contains five different non-overlapping input patterns, each one presented 16% of the time (80% in total). The policy used to give rewards (excitation) and punishments (inhibition) to the agent (dopaminergic neuron) is the following. When pattern 1 or 2 is present, the agent is rewarded if action A is selected (action A neuron fires during the pattern presentation and action B neuron does not fire) but punished if action B is selected. When pattern 3 or 4 is present, the agent is rewarded if action B is selected but punished if action A is selected. When pattern 5 is present, the agent is punished if it selects action A or B. This policy applies no punishment or reward to the agent during noisy inputs, whatever the action taken is. In case of spiking both action neurons during a reinforced input, the network is punished.

2.3. Performance measurement

In the action-selection task we measure the performance of the models by calculating the percentage of correct action choices (i.e. the learning accuracy). This measure is widely used in classification problems when the objective is to describe the accuracy of a final map process (Stehman, 1997). To do so, for each pattern presentation we store the rewarded (expected) action in response to the presented pattern, and the finally selected (chosen) one during that pattern presentation. We only consider in the calculation those trials in which some reward or punishment can be delivered, ignoring those intervals with no repeating patterns conveyed to the inputs (only noisy inputs). We consider that an action has been taken if the corresponding action neuron has spiked at least once during the pattern presentation. Conversely, we consider that no action has been taken if none of the action neurons spikes during the same duration. In order to obtain an estimation of the temporal evolution of the accuracy we use a rolling mean of the last 100 values.

3. Results and discussion

We did extensive testing of the learning mechanisms we proposed. Some of these results demonstrate that the combination of STDE learning rule and homeostatic mechanisms allow learning (and re-learning) of rewarded patterns, or that there is no effect of the reward delay and the frequency of the input pattern on the learning process, among others. However, as they are not the main concern for this article, they are placed in the Supplementary Results 5.2 section for further examination.

The main results and discussion are structured as follows: we first show the general behavior of the network. Then we study the effect...
of the lateral connectivity pattern on the performance and the way neurons are processing information. Finally, we put our results in context by comparing our model with previously proposed models in the literature.

3.1. General network behavior

During the simulation of the action-selection task, each action group neuron becomes overall active in response to the presentation of the associated patterns as shown in the raster plots (Fig. 3C and D) and the activity balance for the action neuron groups (Fig. 3E), producing mainly dopaminergic rewarding (Fig. 3F). The action accuracy reveals steady-state performance after 200 seconds of simulations (Fig. 3G). According to these results, our combination of STDE learning rule (Gurney et al., 2015) with homeostatic mechanisms and an oscillatory input signal in a cortico-striatal model learns to accurately select the most rewarding action.

The way our network learns to associate the corresponding input stimulus with sub-populations of D1 and D2 neurons in channel A or channel B is the following: If the agent takes the right action for a specific input pattern, the environment delivers a reward with some delay (high DA level in Fig. 3F). This reward potentiates the synapses between the cortical layer and the action-associated D1 sub-population, resulting in more frequent firing. On the other hand, if the agent takes a wrong action, then it receives a punishment sometime later (low DA level in Fig. 3F). This punishment weakens the synapses from the cortical layer to the action-associated D1 sub-population while strengthening the corresponding synapses to the D2 (inhibitory) sub-population of the same channel. This learning process makes the agent stick to the rewarded action and switch to a different one when punished. For the specific case when the environment punishes any action during a stimulus presentation, both D2 sub-populations increase their activity, and both action neurons remain silent.

The proposed model shows how combining two complementary dopamine-based STDE learning rules (Fig. 1) can facilitate the association between sensorial cortical inputs and rewarded actions with arbitrary rewarding policies. Previously, the STDE rule had been shown to be capable of learning to select an action channel driven by stronger cortical input (Gurney et al., 2015), and here we show that this rule can also be used to learn inputs defined by the specific timing of their spikes (as all the inputs have the same average firing rate). This represents a higher complexity task and illustrates how STDE can be efficiently used for spike time pattern representation.

The model also is completely bioplausible, as all the mechanisms used have been described in biological systems: DA induces bidirectional, timing-dependent plasticity at MSNs glutamatergic synapses (Shen et al., 2008, in vitro pyramidal neural recordings are consistent with simulations of adaptive spike threshold neurons, and they lead to better stimulus discrimination than would be achieved otherwise (Huang et al., 2016), and rat hippocampal pyramidal neurons in vitro can use rate-to-phase transform (McLelland and Paulsen, 2009). Detailed discussion on the role of the homeostatic mechanisms can be found in Supplementary Materials.

3.2. Effect of lateral inhibition patterns and task complexity

Once we have demonstrated how the striatal network can support RL, we wondered to what extent the connectivity pattern of the lateral inhibition in the striatum could impact the learning capabilities. So that we extensively explored different versions of connectivity.

We first study if there is any relationship between the connectivity pattern and difficulty of the task. We organized the lateral inhibitory connections in two groups: intrachannel (inhibitory connections from D2 MSNs to D1 MSNs within the same channel) and
Figure 3: Cortico-striatal network solving a RL task. 

A. Structure of the network. See section 2.1.5 for a detailed explanation.

B-F. The activity of the network during the last 5 seconds of simulation. Background color indicates the reward policy (yellowish colors, action A is rewarded and B is punished; bluish colors, action B is rewarded and A is punished; grey, any action is punished).

B. Input pattern conveyed to the cortical layer.

C. Raster plot of the channel-A action neurons. Yellow dots represent MSN D1 spikes, and orange dots are MSN D2 spikes.

D. Raster plot of channel B. Cyan dots represent MSN D1 spikes, and dark blue dots are MSN D2 spikes.

E. Action neuron firing rates. The middle horizontal line represents 0 Hz. Action A and B activity are represented in opposite directions for clarity. Action A neuronal activity increases in yellow zones while action B neuronal activity in cyan intervals.

F. Firing rate of the dopaminergic neuron (black line). Dotted horizontal lines indicate the range of DA activity considered: black is the baseline, green is the maximum reward, and red represents the maximum punishment. Dots indicate rewards (green) and punishment (red) events delivered to the agent.

G. Evolution of the learning accuracy of the agent, see section 2.3 for further details. The dotted line marks the accuracy level by chance.
inter-channel (inhibitory connections between D1 MSNs of different channels, and between D2 MSNs of different channels). We obtained four possible subsets of connectivity patterns by keeping or removing each connection type (Fig. 2). We used three difficulty levels for the task: easy, normal and hard. The easy task uses only one stimulus associated with each action (stimulus 1 to action A, stimulus 2 to action B, stimulus 3 to no action). The normal task uses two stimuli per action, and one no-go stimulus. The hard task uses four stimuli per action, and two no-go stimuli.

The results of the easy version of the experiments are shown in the Fig. 4. The models without inter-channel inhibition work worse, as they stabilize with lower values of accuracy. The models with inter-channel inhibition seem to reach a similar level of accuracy but the intra-channel inhibition seems to reduce the learning rate.

In the normal version of the task, we again obtained the best learning performance when using the inter-channel lateral inhibition with asymmetrical structured connection pattern, and the difference increased. In this case, there is no apparent effect in of the intra-channel lateral inhibition in this task (Fig. 5). According to our simulations, lateral inter-channel inhibition facilitates the emergence of one action-related channel over the other one in a winner-take-all manner, as expected.

We saw in previous experiments that the inter-channel lateral inhibition is always increases accuracy, so we will use it always in the following tests. In the hard task we obtained small but significant differences: The accuracy of the network improves faster with the intra-channel lateral inhibition (see Fig. 6). Also, apparently the network with the intra-channel inhibition settled in a more stable regime as it maintains its performance, compared with the network without this intra-channel inhibition which slowly degrades (Supplementary...
Figure 6: Effect of the intra-channel lateral inhibitory connectivity on the performance during a harder version of the RL task. The horizontal dotted line represents the accuracy obtained by a random agent. The curves and the filling color represent the mean, the standard error of the mean, respectively, of the evolution of each agent during the task (n=150), simulated for 500 seconds.

The results so far suggest that both connectivity patterns contribute to a reliable action-selection paradigm.

Taking these results together, it seems that when we use several stimuli associated with each action, intra-channel inhibition improves the RL action selection task. However, when only one stimulus is associated with each action, this intra-channel inhibition does not impact learning performance. When compared with the results in Fig. 5 and 6, it seems that the intra-channel lateral inhibition improves the learning capabilities only with a harder task, but when the task is too simple then the intra-channel connection increases the learning time.

We also explored the effect of connectivity patterns of lateral inhibition different from the proposed by Burke et al. (2017), by adding or removing lateral connections within a channel, within each subpopulation, and between subpopulations of the same channel. All variations from the original resulted in reduced learning performance (Supplementary Fig. 15). In this Figure, the curve #5 represents the network with both lateral inhibition in D1 layer and D2 layer, as well as intra- and inter-channel lateral inhibition. This structure (similar to the one proposed by Burke et al. (2017)) obtains the best accuracy.

3.3. Effect of intra-channel lateral inhibition on neuronal specialization

Intra-channel inhibition seems to facilitate learning in more complex tasks, possibly because it enhances neuron specialization. We saw a strong reduction of correlation at time difference $\delta t = 0$ between action A and B D1 sub-populations caused by intra-channel inhibition (data not shown), but this does not seem to justify the improved accuracy for more complex tasks.

Then, we hypothesized that intra-channel inhibition could encourage neuron specialization to specific cortical patterns. We tested this idea by analyzing the preferred stimuli for each neuron after the learning process (Fig. 7), and obtained the opposite result: the intra-channel lateral inhibition affects D1 neurons by forcing them to share more evenly their activity over several stimuli, in addition to reducing their average activity. This is in contrast with the network without intra-channel lateral inhibition, where the activity is more focused on the favorite stimuli and has higher mean activity.

According to these results, although individual neurons of the network with intra-channel inhibition have less precise representation of individual sensorial stimuli, these models have higher precision to associate rewarding actions. This can be explained assuming some sparse representation of the stimuli, where the simultaneous firing of several (but not many) neurons are needed to indicate the presence of an input stimuli. This more sparse representation emerges due to the combination of stronger inhibition and the homeostatic mechanisms: a neuron avoids firing when it is inhibited, so
the homeostatic mechanisms tend to compensate for this activity reduction by increasing its chances to fire in response to several stimuli. This sparse representation has been suggested to facilitate sensory pattern recognition in other brain areas, such as the cerebellar cortex, the mushroom body, and the dentate gyrus of the hippocampus (Cayco-Gajic and Silver, 2019).

In the context of our model, the sparse representation due to intra-channel inhibition plays a role in the action selection process, which can be seen as a form of classification. Here, the goal is not to classify stimuli per se, but to assign stimuli to appropriate actions. The sparse coding helps to achieve more efficient and robust action selection by reducing the overlapping between representations of different sensorial states, minimizing interference, and enabling more reliable decision-making.

3.4. Comparison with previous models of reinforcement learning and basal ganglia

We presented a point-neuron model of the BG that can solve complex action-selection tasks using a RL paradigm. We do so by using multiple mechanisms proposed in the literature: the STDE learning rule that implements synaptic modification in cortex-MSN connections (Gurney et al., 2015), combined with homeostatic mechanisms (Galindo et al., 2020) and an oscillatory input signal (Masquelier et al., 2009; Garrido et al., 2016) in a network with asymmetrical structured lateral inhibition (Burke et al., 2017) can rapidly and consistently learn to detect the presence of rewarded input patterns. These processes have been described in biological systems and here proved to be robust.

Simpler STDP-like rules have been used for RL tasks (Izhikevich, 2007; Legenstein et al., 2008), but they were employed in simpler networks, single neurons, and simple tasks. Beyond the state-action mapping role proposed in this article for the striatum, other theories exist about the action decision process. However, computational models of BG in the literature have considerably evolved during the last two decades (Rubin et al., 2021), and there is still no consensus about how to achieve goal-oriented learning in a BG model. Previous models ranged from those with action-selection features but no learning (Beiser et al., 1997; Gillies and Arbuthnot, 2000; Humphries et al., 2006; Lo and Wang, 2006a; Berns and Sejnowski, 1998; Gurney et al., 2001; Sen-Bhattacharya et al., 2018; Frank, 2006; Ratcliff and Frank, 2012; Bogacz, 2007) (but see Frank, 2005) to simple forms of learning, with RL (Bogacz and Larsen, 2011), rate-based learning rules (Hong and Hikosaka, 2011), or based on modulated STDP with eligibility traces (Humphries et al., 2009; Gurney et al., 2015; Baladron et al., 2019). These models considered direct and indirect pathways (as "selection" and "control" routes, respectively), composed of MSN D1 and D2 striatal neurons controlling GPe and SNr. Many models assume that the BG work as an actor-critic model (Bogacz and Larsen, 2011; O’Doherty et al., 2004), and actor-critic frameworks have
been used for RL tasks like maze navigation (Frémaux et al., 2013; Potjans et al., 2009; Vasilaki et al., 2009) and cartpole (Frémaux et al., 2013). More biologically-constrained models of the BG have been proposed to explain the origin of diseases like Parkinson’s disease (Lindahl and Kotaleski, 2016) and the role of specific interneurons (Goenner et al., 2021) or pathways (Girard et al., 2021) during action-selection. Recent accumulation-to-bound models describe the decision process as an accumulation of evidence for each alternative action until a decision threshold is exceeded in one of these actions (Mulder, 2014). It would be interesting to explore how these models could be incorporated with the proposed model, potentially requiring additional brain areas. In this regard, some models incorporate recurrent activity loops with the cortex through the thalamus (Lo and Wang, 2006b).

Moreover, we acknowledge that similar models can already deal with more complex action-selection tasks than the one used in this work, such as cart-pole, inverted pendulum, or simple mazes (Frémaux et al., 2013). However, there exist some important differences between their model and the one proposed in this article. First, our network does not include a critic. Second, their learning rule requires a temporal difference (TD) signal that would need additional circuitry. Third, their model requires an additional place-cell layer with unsupervised learning to represent complex input patterns. However, it remains as a future work to embed the network model into a closed-loop experimental setup requiring continuously graded output (instead of selecting an action in a discrete set of possibilities). This way, the model could deal with a larger set of RL tasks. In our case, we have integrated a spiking neural network with spike-time pattern representation that scales well with different patterns complexity at the pattern classification layer. Future work will explore how our model could be extended for such complex action control frameworks.

4. Conclusion

In this article we tested the respective roles in learning of the different mechanisms used during our simulations: homeostatic mechanisms make the neurons change their response to compensate for long-lasting changes in the input level, making learning faster and more robust to the configuration. The asymmetrical lateral inhibition consistently outperformed other connectivity configurations. By adding intra-channel lateral inhibition to the network model, we induced the channels to generate a sparse representation of each stimulus relevant for the task. This made the network less prone to errors as the model had to recruit more neurons to take an action. Lastly, by segregating striatal and action neurons in independent channels for each action and incorporating MSN D1 (Go neurons) and MSN D2 (No-Go) sub-populations with different learning kernels, the model effectively learned arbitrary mappings from sensorial input states to action output in a two-choice action-selection task. MSN D1 neurons and MSN D2 neurons cooperatively facilitated action selection with contrary effects; MSN D1 neurons learned to potentiate preferred actions while MSN D2 neurons learned to inhibit non-preferred actions.

Acknowledgements

This research is supported by the Spanish Grant INTSENSO (MICINN-FEDER-PID2019-109991GB-I00), Regional grants Junta Andalucía-FEDER (CEREBIO P18-FR-2378 and A-TIC-276-UGR18). This research has also received funding from the EU Horizon 2020 Framework Program under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3) and the EU Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 891774 (NEUSEQBOT). Additionally,
the main author has been funded with a national research training grant (FPU17/04432). Finally, this research was also supported by the Vetenskapsrådet (VR-M-2017-02806, VR-M-2020-01652); Swedish e-science Research Center (SeRC); KTH Digital Futures.

References


Gurney K, Prescott TJ, Redgrave P. A computational model of action selection in the basal ganglia. i. a new functional anatomy. Biological cybernetics


5. Supplementary materials

5.1. Supplementary methods

5.1.1. Single-striatal-neuron model and experiments

Figure 8: Pattern detection experiments with reinforcement learning and a single striatal neuron. A. Single-striatal-neuron model setting, with serial and oscillatory input currents feeding to a cortical layer. In this simulation, two different input patterns are used and colored in green and red. The cortical layer feeds the striatal neuron with plastic synapses with STDE, where learning occurs. A reward or punishment signal is delivered to a global dopaminergic neuron that modulates the plastic synapses. B. Raster plot of the cortical neurons (blue dots), with input patterns containing only half of the cortical neurons, oscillatory driving current (solid red line), and the striatal neuron (bottom, red dots). C. Evolution of the striatal neuron’s response to each input pattern through time measured using a uncertainty coefficient (see details in the methods section). Insets show the distribution of synaptic weights at the beginning and the end of the learning procedure.

In order to assess the learning capabilities of the proposed model, we define two types of experiments: pattern detection and action-selection. The latter one is already explained in the main text. During pattern detection experiments (Supplementary Fig. 8B), we train a simple model to detect one specific pattern within a noisy input stream. Two (the so-called selected and non-selected) non-temporally-overlapping repeating patterns are presented 20% of the time each (40% in total). We test the STDE learning rule in a RL setting, where a reward (excitation to the dopaminergic neuron) is given if the striatal neuron spikes sometime after the selected pattern is presented. Otherwise, if the striatal neuron fires in response to the non-selected pattern, punishment (inhibition to the dopaminergic neuron) is given to the striatal neuron. Finally, as a stress test, we added a policy swapping procedure for switching the rewarded pattern every 200 seconds (Supplementary Fig. 9). This way, we can test how robust is our combination of synaptic and homeostatic rules during learning.

For this first set of experiments, we used a model with only one striatal neuron that learns to solve a simple RL task. This model allows the validation of the proposed learning mechanisms. It uses the input protocol explained in the oscillatory drive section 2.1.2. A dopaminergic signal modulates the synapses that connect from the cortical neurons to the striatal neuron (Supplementary Fig. 8A), implementing the STDE learning rule as well as the homeostatic mechanisms. Rewards (punishments) delivered by the environment alter the dopaminergic modulatory signal by exciting (inhibiting) the dopaminergic neuron every time the striatal neuron spikes when the input pattern is correct (incorrect). The environment delivers rewards and punishments with some delay (fixed to 300 ms by default). If the striatal neuron does not fire, the environment delivers no reward nor punishment to the DA neuron.
5.1.2. Mutual information

In order to measure how good the detection is in the pattern detection experiments, we calculated the mutual information (MI) between the presentation of each input pattern and the striatal neuron activity, as previously done in Garrido et al. (2016). We consider that the striatal neuron responded to the pattern if it fires at least once during the stimulus presentation, lasting from 100 to 500 ms following a uniformly distributed random distribution. For each stimulus used in the pattern detection experiments, we consider the possible states S of the pattern (present or absent) and the possible response R of the striatal (neuron fired or not). The MI is then defined in equation (6).

\[
MI = H(S) + H(R) - H(S, R)
\]

where \( H(S) \) is the entropy of the stimuli patterns, \( H(R) \) is the entropy of the responses, and \( H(S, R) \) the joint entropy of the stimuli patterns and the responses. These values are defined as in Garrido et al. (2016). The upper bound of the MI for a perfect detector would be \( MI_{max} = H(S) \), so we can obtain a normalized measurement of performance called uncertainty coefficient (UC) defined in equation (7). The UC is calculated independently for both the rewarded and the non-rewarded patterns during pattern detection experiments.

\[
UC = \frac{MI}{MI_{max}} = \frac{H(S) + H(R) - H(S, R)}{H(S)}
\]

5.1.3. Parameters used

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Cortical</th>
<th>Striatal</th>
<th>Action</th>
<th>Dopaminergic</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \epsilon_{exc} ) (mV)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>( \epsilon_{inh} ) (mV)</td>
<td>-85.0</td>
<td>-85.0</td>
<td>-85.0</td>
<td>-85.0</td>
</tr>
<tr>
<td>( \tau_{AMPA} ) (ms)</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>( \tau_{GABA} ) (ms)</td>
<td>10.0</td>
<td>30.0</td>
<td>60.0</td>
<td>10.0</td>
</tr>
<tr>
<td>( \tau_{ref} ) (ms)</td>
<td>1.0</td>
<td>15.0</td>
<td>15.0</td>
<td>1.0</td>
</tr>
<tr>
<td>( C_m ) (pF)</td>
<td>250.0</td>
<td>50.0</td>
<td>100.0</td>
<td>250.0</td>
</tr>
<tr>
<td>( g_{leak} ) (nS)</td>
<td>25.0</td>
<td>10.0</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td>( V_{thr} ) (mV)</td>
<td>-40.0</td>
<td>-50.0</td>
<td>-40.0</td>
<td>-65.0</td>
</tr>
<tr>
<td>( \epsilon_{leak} ) (mV)</td>
<td>-65.0</td>
<td>-65.0</td>
<td>-65.0</td>
<td>-40.0</td>
</tr>
</tbody>
</table>

Table 1: Neuron parameters used in the model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_{lo}^{-} )</td>
<td>0.0</td>
<td>( k_{lo}^{-} )</td>
<td>-1.0</td>
</tr>
<tr>
<td>( k_{lo}^{+} )</td>
<td>-1.0</td>
<td>( k_{lo}^{+} )</td>
<td>1.0</td>
</tr>
<tr>
<td>( k_{hi}^{-} )</td>
<td>-1.0</td>
<td>( k_{hi}^{-} )</td>
<td>0.0</td>
</tr>
<tr>
<td>( k_{hi}^{+} )</td>
<td>1.0</td>
<td>( k_{hi}^{+} )</td>
<td>-1.0</td>
</tr>
</tbody>
</table>

Table 2: STDE parameters used in the model.
5.2. Supplementary results

5.2.1. Single-striatal-neuron experiments

In a previous article by Masquelier et al. (2009), an oscillatory driving signal greatly facilitates the recognition of complex patterns over noise with STDP-like rules. We have extended this learning rule to account for a rewarding signal in a RL paradigm. During a whole learning task (lasting 200 seconds), two different repeating and non-overlapping input patterns are presented. Only one of them produces a rewarding signal if, and only if, the striatal neuron fires simultaneously to the pattern presentation, providing reward modulation to the learning rule. Using this RL framework, the striatal neuron becomes selective to the presentation of the rewarded pattern only (Supplementary Fig. 8B). It usually takes less than 100 seconds of simulated time to consistently generate spikes with the presentation of the rewarded pattern (Fig. 8B). The detection capabilities of this network are also evidenced by the evolution of the uncertainty coefficient (green line in Supplementary Fig. 8C), which remains stable between 0.6 and 0.8 after 80 seconds of discontinuous pattern presentation (Supplementary Fig. 8C), while the punished pattern receives no considerable response (red line in Supplementary Fig. 8C). It can also be observed how the initial uniform weight distribution (insets in Supplementary Fig. 8C) turns into a binomial distribution with a small number of synapses with near-maximum weights and most of the synapses near the minimum weight.

Once demonstrated the effectiveness of the STDE learning rule, we aim to assess if it allows detection of rewarded patterns with policy swapping (i.e., the pattern that offers rewarding signals is swapped every 400 seconds of simulation). Every time that the rewarding policy swaps, the neuron temporarily reduces its average firing rate (cyan line in Supplementary Fig. 9D), and consequently, the adaptive firing threshold approaches the resting potential (pink line in Supplementary Fig. 9D). Once the threshold is low enough, the neuron starts learning the new rewarded pattern, increasing the activity of the dopaminergic neuron as a consequence (Supplementary Fig. 9Ei). This is an important feature because neurons can recover from silent states caused by sudden changes in the reward policy.

Inspired by the different types of neurons existing in the striatum, we adapted the synaptic model parameters to reproduce the differential operation of the learning rule for the MSN D1 and the MSN D2 neurons (MSN D1 and D2 parameters for STDE in Supplementary Table 2). Thus, we adjusted different kernel shapes for low and high DA (Supplementary Fig. 9Bi and Bii, left and right, respectively). According to our simulations, the neuron equipped with a D1 kernel learns to detect only the rewarded pattern (Supplementary Fig. 9Ci). In contrast, the striatal neuron equipped with MSN D2 kernel parameters (a reversed version of MSN D1) learns to detect the non-rewarded pattern (Supplementary Figs. 9Cii, 9Dii and 9Eii). These results point out that, in a network of MSNs with D1 and D2 subpopulations, the D1 subpopulation learns to respond to rewarded patterns while the D2 neurons learn to fire in response to the punished (or non-rewarded) patterns. In this way, the output layer makes simple decisions by just weighting the activity of these subpopulations.

5.2.2. Homeostatic mechanisms: non-Hebbian strengthening and adaptive threshold

Aiming to check the influence of the homeostatic mechanisms, we have replicated the same policy-swapping learning framework with a more complex task (five different input patterns) and different configurations of the homeostatic rules. In the absence of non-Hebbian strengthening, successful learning requires fine-tuning of the learning rule parameters and maximum weight for
Figure 9: Pattern detection experiments with two different STDE sets of parameters. **Xi** column shows the training resulting from using a learning kernel adapted to learn rewarded patterns, as used in MSN D1 synapses. **Xii** column shows the same training results for the MSN D2 kernel used. Note that this kernel is learning the opposite (punished) pattern. **Ai** row shows the kernel functions used with different levels of DA. **Bi** row shows the response of the striatal neuron. The background color indicates which pattern is being rewarded at that specific time frame, and the vertical dotted lines indicate when the rewarding policy swaps. **Ci** row shows the evolution of the adaptive threshold and the firing rate of the striatal neuron. **Di** row shows the firing rate of the dopaminergic neuron, which represents the amount of reward obtained by the striatal neuron through the task. The horizontal green, black and red dotted lines indicate the maximum, baseline, and minimum dopaminergic activity.

Each simulation seed (data not shown). Thus, we barely managed to find a set of parameters suitable for multiple seeds without this homeostatic mechanism. For this reason, in all the simulations shown in this article we employ the non-Hebbian strengthening mechanism.

On the other hand, the adaptive threshold is not strictly necessary for successful learning. However, the learning performance (in terms of UC) with adaptive threshold increases faster and more reliably than without adaptive threshold (Supplementary Fig. 10). It is important to highlight that lack of homeostatic mechanisms often resulted in the more frequent inability of detecting cortical patterns, as demonstrated by lower MI values for the 25-percentile of the simulations (lower boundary of blue shadow in Supplementary Fig. 10, right). In the absence of these mechanisms, the striatal neuron activity extinguishes when the reinforcement policy swaps and, in many cases, remain silent for the rest of the simulation. We tested different learning rates and time constants of dopamine and, in every case, learning was faster with adaptive threshold, as shown in Supplementary Fig. 14. Thus, these homeostatic rules provide the STDE rule with the ability to re-learn different patterns reliably. Moreover, using both of these mechanisms also makes learning robust within a broader parameter space and makes it unnecessary to fine-tune the parameters for each experiment. Although only one of these homeostatic mechanisms would be enough to avoid silent
neurons, we saw in our tests that the system recovered faster and more reliably by using both.

![Graph showing the effect of adaptive threshold](image)

Figure 10: Effect of the adaptive threshold in the learning performance of the single-striatal-neuron model. In this experiment we used a more complex version of the policy-swap task with 5 (one rewarded, the rest punished) different patterns instead of 2. The left curves and the filling represent the evolution of the mean uncertainty coefficient and standard error during a repeated 400-s learning protocol (n=300). The asterisks marked intervals indicate where the means are statistically different with 95% confidence level. The right plot shows the percentiles 5-25-50-75-95, with dashed lines (5 and 95 percentiles), fillings (25 and 75 percentiles) and solid lines (50 percentiles).

5.2.3. Effect of reward delay and input pattern

We wondered how the delay between the action decision (in response to cortical stimulus) and the rewarding signal affects the learning capabilities of our system. In order to evaluate the impact of this parameter, we carried out network simulations with different reward delays (we did not have to adjust any other parameter due to the robustness of the model). We found the best performance when the rewarding signal was provided 300 ms after the sensorial presentation (blue line in Supplementary Fig. 11). Longer or shorter delays resulted in decaying learning accuracy. This result is similar to what can be found in biology (Yagishita et al., 2014).

Since our implementation of the DA-modulated learning rule is based on eligibility traces, we wondered if this optimal delay was somehow related to the duration of the stimulation patterns. Then, we evaluated the reward delay effect on learning when sensorial patterns were longer (300-700 ms and 500-900 ms) than in the control case (100-500 ms). However, our simulations show similar learning accuracy with longer cortical patterns (orange and green lines in Supplementary Fig. 11) as in control conditions (blue line in Supplementary Fig. 11). So that it seems unlikely that the pattern generation algorithm influenced the preferred delay.

Finally, we also studied how the frequency of pattern presentation influences the accuracy achieved at the end of the simulation. We compared the results obtained presenting the patterns 80 percent of the time (as in the rest of the experiments made) with the results obtained by presenting the patterns 40 percent. In order to compensate for the lower exposure of the striatal neurons to input patterns (since in the latter, the network will only see the patterns half the time), we simulated twice as long (up to 1000 seconds). According to our simulations, the proposed network similarly managed to successfully associate cortical inputs to associated actions independently.
of how often the patterns are presented, as long as it experiences enough trials (Supplementary Fig. 12).
Figure 11: Effect of the delay of the rewarding feedback in the learning accuracy.  

A. Simulations with different pattern lengths: within 100 to 500 ms (blue), within 300 to 700 ms (in orange), and within 500 to 900 ms (in green). Every point represents the mean accuracy level obtained in the last 100 seconds of simulation with different delay values, and the shaded area shows the standard error of the mean (n=10).  

B. Notched box plot of all the values. Notice that “if the notches about two median do not overlap, the medians are, roughly, significantly different at about a 95% confidence level” (see McGill et al. (1978) for details).
Figure 12: Effect of the delay of the reward in the learning performance with input pattern proportion of 0.8 (in blue), and with input pattern proportion of 0.4 (in orange). Every notched box [McGill et al., 1978] represents the median (n=10) performance level obtained in the last 100 seconds of simulation for different delay values.
5.2.4. Effect of lateral inhibition in harder experimental setting

Figure 13: Same as in Fig. 6 but simulated for 1000 seconds.
5.2.5. Effect of DA time constant, learning rate and adaptive threshold

Figure 14: Learning performance for different values of DA time constant, learning rate and adaptive threshold. At the top, mean and standard error are shown for each condition. At the bottom, boxplots of the last 200 seconds of simulation (n=80).
5.2.6. Lateral connectivity patterns effect

Figure 15: Learning performance for different connectivity patterns of lateral inhibition. Left: Connectivity topologies tested in these experiments. Note that all these tests assume inter-channel inhibition, as they clearly outperformed other models. Right: evolution of the learning accuracy during 500s of simulation with the medium-complexity task. Every line is marked with the same color of the topology under test. Each line represents the average value with $n = 10$ seeds.