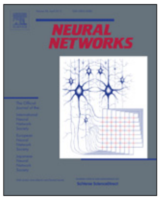
Neural Networks 41 (2013) 212–224

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



2013 Special Issue

A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning

Fabian Chersi [a,*], Marco Mirolli [a], Giovanni Pezzulo [b,a], Gianluca Baldassarre [a]

[a] *Institute of Cognitive Sciences and Technologies, National Research Council. Via San Martino della Battaglia 44, 00185 Roma, Italy*

[b] *Institute of Computational Linguistics "Antonio Zampolli", National Research Council. Via Giuseppe Moruzzi 1, 56124 Pisa, Italy*

ARTICLE INFO

Keywords:
Autonomous learning
Goal-directed and habitual actions
Motor sequences
Basal ganglia
Spiking neurons

ABSTRACT

Dual-system theories postulate that actions are supported either by a goal-directed or by a habit-driven response system. Neuroimaging and anatomo-functional studies have provided evidence that the prefrontal cortex plays a fundamental role in the first type of action control, while internal brain areas such as the basal ganglia are more active during habitual and overtrained responses. Additionally, it has been shown that areas of the cortex and the basal ganglia are connected through multiple parallel "channels", which are thought to function as an action selection mechanism resolving competitions between alternative options available in a given context.

In this paper we propose a multi-layer network of spiking neurons that implements in detail the thalamo-cortical circuits that are believed to be involved in action learning and execution. A key feature of this model is that neurons are organized in small pools in the motor cortex and form independent loops with specific pools of the basal ganglia where inhibitory circuits implement a multistep selection mechanism.

The described model has been validated utilizing it to control the actions of a virtual monkey that has to learn to turn on briefly flashing lights by pressing corresponding buttons on a board. When the animal is able to fluently execute the task the button–light associations are remapped so that it has to suppress its habitual behavior in order to execute goal-directed actions.

The model nicely shows how sensory-motor associations for action sequences are formed at the cortico-basal ganglia level and how goal-directed decisions may override automatic motor responses.

1. Introduction

A prominent feature of animal intelligence is the ability to process different types of stimuli at the same time and to effortlessly solve real-world tasks of varying complexity simultaneously or in rapid sequences. In order to do this the brain has developed two distinct mechanisms for action selection, with separate neurological bases. One process, termed "goal directed" (Dickinson & Balleine, 2000; Hommel, 2003), motivates action by the integration of an expectation that a given action will have a specific outcome and a desire for that outcome. On the other hand, stimulus-driven "habitual" actions occur as an automatic response to sensory inputs with which the action has become associated, for example through reinforcement learning (Balleine & Dickinson, 1998). Although apparently simple, the latter mechanism can produce very complex behavioral patterns combining basic learned responses (Donahoe, Burgos, & Palmer, 1993).

In the first mechanism, the tendency to select a particular action depends on the currently predicted value of the outcome. On the contrary, once a habitual action is learned it is elicited in response to a stimulus irrespective of what the value of the outcome may be.

In healthy individuals, the acknowledgment that the current situation suddenly cannot be solved in the habitual way evokes the generation of an alternative plan. On the contrary, adults with brain damage in areas such as the frontal lobes are usually prone to continue to unsuccessfully perform habitual actions elicited by known stimuli. Such patients provide a clear demonstration of how processes governing automatic action execution can operate independently of decision-making processes.

Recently, experimental studies, in particular on rodents, have begun to elucidate the neural substrates underlying these types of behavioral control processes. More specifically, a series of works have shown that the dorsomedial striatum and the prelimbic cortex subserve goal-directed actions (Corbit & Balleine, 2003; Killcross & Coutureau, 2003; Miyachi, Hikosaka, & Lu, 2002; Yin, Knowlton, & Balleine, 2005), whereas habit formation is reflected in a shift in control toward the dorsolateral striatum (Yin & Knowlton, 2006; Yin, Knowlton, & Balleine, 2004). Importantly,

[*] Corresponding author. Tel.: +39 06 44595206.
E-mail address: fabian.chersi@istc.cnr.it (F. Chersi).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.11.009

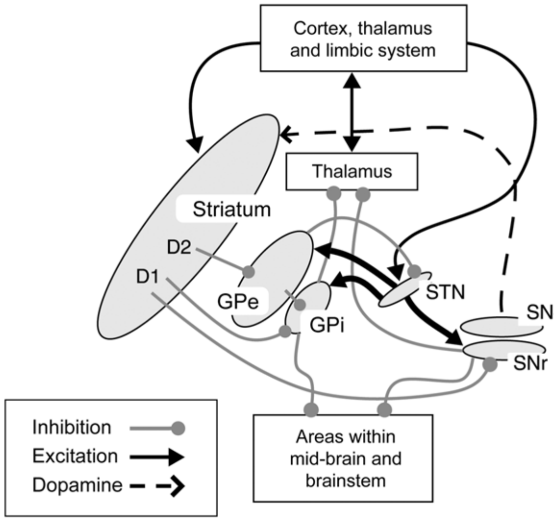

Cortex, thalamus and limbic system
Thalamus
Striatum
D2
D1
GPe
GPi
STN
SNc
SNr
Inhibition
Excitation
Dopamine
Areas within mid-brain and brainstem

Fig. 1. Principal connections of the basal ganglia. Abbreviations: STN, subthalamic nucleus; GPe/GPi: external/internal globus pallidus; SNc/SNr, substantia nigra pars compacta and pars reticulata; D1, D2, striatal neurons preferentially expressing dopamine receptors subtypes D1 and D2.
Source: Modified from Gurney et al. (2001).

dopamine seems to be strongly involved in the reinforcement of habits (Faure, Haberland, Conde, & El Massioui, 2005; Wise, 2004).

Some of the mentioned areas are part of a larger ensemble of subcortical nuclei, the basal ganglia (BG), supposedly involved in vertebrate action selection, in both motor and cognitive domains (Mink, 1996; Alexander, DeLong, & Strick, 1986). The traditional components of the BG are as follows (see Fig. 1): the striatum (STR, also called caudate/putamen in primates), the subthalamic nucleus (STN), the external globus pallidus (GPe), the internal globus pallidus (GPi), the substantia nigra pars reticulata (SNr), and the pars compacta (SNc).

The BG output nuclei are the GPi and the SNr, which send projections to the subcortical areas responsible for posture and locomotion (McHaffie, Stanford, Stein, Coizet, & Redgrave, 2005) and to parts of the motor thalamus, which in turn project to the motor cortex (Gerfen, Staines, Arbuthnott, & Fibiger, 1982).

The cortical input to the STN originates in the primary and supplementary motor areas (M1 and SMA) and in the frontal eye field (FEF). The STN sends its excitatory (glutamatergic) outputs to the GPe and the GPi, and in turn receives inhibitory (GABAergic) projections from the GPe. The GPe projects not only to the STN but also directly to the GPi, the SNr, and to the striatum, which in turn projects mainly to the GPe and the GPi. Striatal neurons are traditionally classified in two categories: those having dominant dopamine receptor type 1 (D1R) and those having dominant dopamine receptor type 2 (D2R).

The purpose of the present work was to develop a spiking neuron model of the circuit comprising the basal ganglia and parts of the sensory, the motor and the prefrontal cortices, with the aim of testing its functioning during stimulus-driven responses, its action selection mechanisms, its switching capabilities from habitual to goal-directed behavior, and finally its unsupervised learning abilities of motor sequences.

One of the important aspects of this study is the fact that the circuit is used to control in closed-loop modality an agent that has to act autonomously in a realistic experimental setup. This implies that the agent not only has to learn to select and combine actions in the correct order, but also has to fine-tune the execution timing of each element in order to comply with real-world constraints.

In the following sections, we will first describe the task utilized in this experiment and then explain the details of the model (from the network configuration to the single neuron model). In Section 4, we will present the results of our experiments. Finally, Section 5 will conclude our paper with the discussion of the results and predictions. Parts of this study have been described in Chersi, Mirolli, Gurney, Redgrave, and Baldassarre (2010).

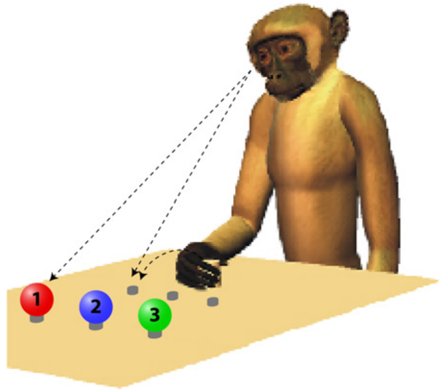


Fig. 2. Experimental setup. A monkey is seated at a table on top of which there are three lights that can be turned on by pressing the corresponding buttons. Initially, the correspondence between buttons and lights is not known to the animal. The task is to learn the correct movement sequence that turns on the indicated light.

2. Task description

In the task simulated in the present work, a monkey sits at a table on which there are three buttons and three lights (see Fig. 2). Each light can be turned on by pressing the corresponding button. However, the correct association between buttons and lights is initially not known to the animal and can be changed at any time.

Each trial begins with a brief flash of one of the three lights, indicating which one has to be turned on. At this point, starting with the hand in the home position, the monkey has to reach and press a button and then look at the light to check if it has been turned on. To render the task biologically more realistic (and interesting) we supposed that the animal is able to reach a button only if it has previously acquired the corresponding coordinates by means of visual inspection. Instead, it is not necessary for the animal to fixate the button in order to press it. Given these constraints, the correct motor sequence for this task consists of the following actions: looking at the button, reaching it, pressing it, and looking at the light.

Finally, we assumed that, if the animal observes the cued light turning on (as a result of its own actions), it receives a reward that causes a temporary dopamine increase in the brain, the effect of which is to modify the connectivity between neurons that were active during the trial (for a more detailed description see below).

3. Model details

3.1. Network architecture

The network presented in this paper aims at modeling the circuit formed by the prefrontal cortex (PFC), the motor cortex (MC), the sensory cortex, the thalamus (TH), and the basal ganglia (BG). The latter has been further subdivided into striatum (STR_{D1} and STR_{D2}), subthalamic nucleus (STN), substantia nigra pars reticulata (SNr), and the external globus pallidus (GPe) (see Fig. 3).

Each of these regions has been modeled as a layer within the network. Neurons in each layer are grouped into small pools (100 units) that share the same properties (i.e. fire in a coherent way). In general, these neurons possess strong local connections and a small number of long-range connections reaching neurons in other pools and other layers, in a configuration known as "small-world" topology (Watts & Strogatz, 1998). This type of configuration

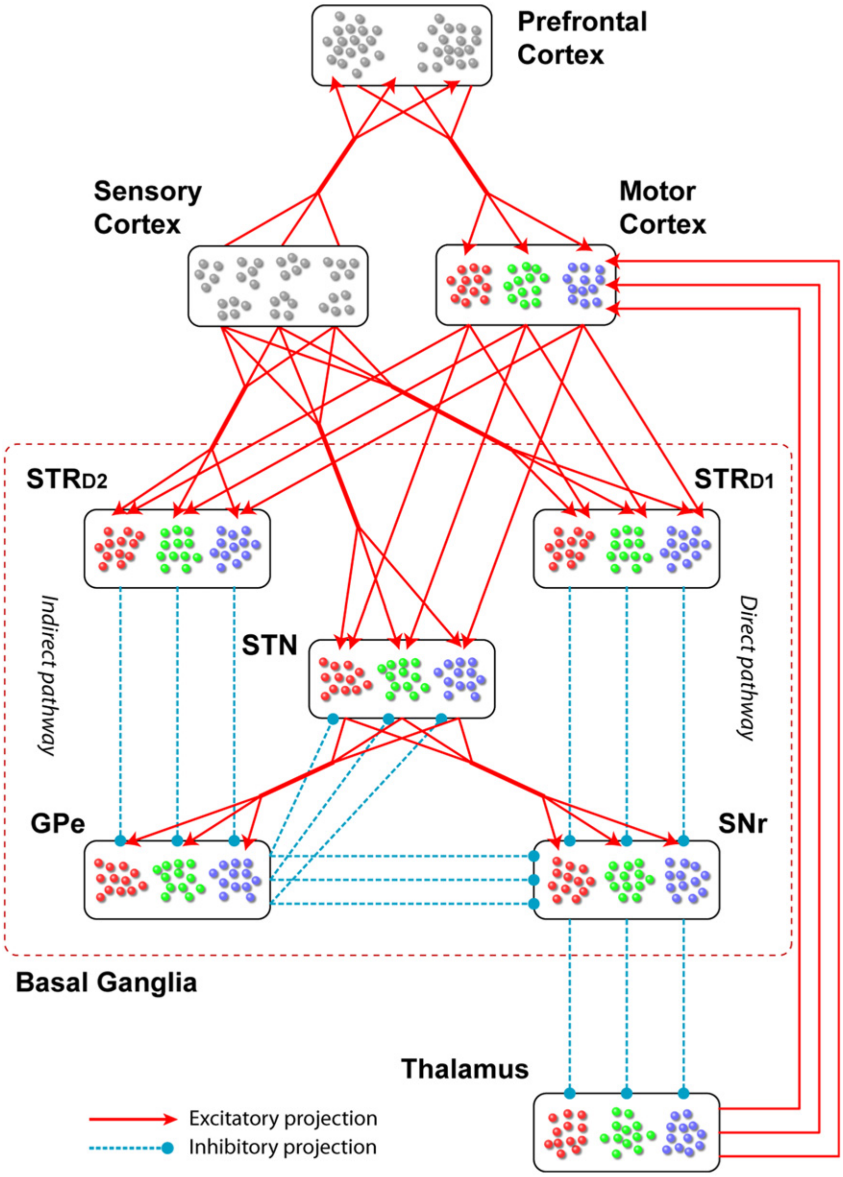


Fig. 3. Scheme of the complete network. Rectangles represent different brain areas and contain segregated groups of neurons. Their different colors represent separate channels within the cortico-basal ganglia network. For the sake of simplicity, only three channels have been shown. Within the BG circuit, competition takes place only between actions involving the same effector (hand or eye). Bundled connections in the figure stand for all-to-all connections. Weights between the sensory areas and the BG, as well as between the prefrontal cortex and the motor cortex, are learned during the task. The sensory cortex receives input from the eye and from proprioceptive sensors, while the motor cortex controls arm, hand and eye movements. Abbreviations: STR, striatum; STN, subthalamic nucleus; GPe, globus pallidus pars externa; SNr, substantia nigra pars reticulata.

is considered to match the connectivity in real neural systems better than either local or random connectivity (Luebke & von der Malsburg, 2004; Mountcastle, 1997), and to optimize the ratio between neurons and connections.

In the present model, the sensory layer comprises multiple external input signals such as proprioception and vision. In particular, pools in this layer convey information about the current fixation target of the eye (home position, button 1, 2 or 3, or light 1, 2 or 3), the perception of a cue signal (i.e. a flash of light 1, 2 or 3), the position of the hand (home position, on button 1, 2 or 3, or having pressed button 1, 2 or 3), for a total of 17 different sensory signals. Neurons in this layer project only to neurons of the same pool, or to neurons in the prefrontal and the basal ganglia layers (more precisely to the STR and the STN) in an all-to-all way. We assumed that only the connections to the basal ganglia neurons are plastic. External inputs, such as visual, proprioceptive and other feedback stimuli, are simulated by means of incoming spike trains directed to specific pools.

The PFC layer encodes the three action goals: turn on light 1, 2 or 3. These pools do not interact with each other but they are connected to neurons in the motor layer in an all-to-all fashion. For the sake of simplicity, we have not replicated the true path followed by PFC signals to the motor cortex (which would have implied having to model also parietal and premotor cortices) but have inserted a propagation delay of 50 ms for compensation. PFC–MC connections are plastic, and they are modified during the learning phase.

Pools in the MC layer encode elementary oculomotor and hand motor acts such as reaching, pressing, and foveating at the different buttons and lights. Neurons that encode hand movements are not specific for the various buttons, i.e. there are no pools encoding “Reach Button 1” or “Reach Button 2”, but instead there are six generic “Reaching” and six generic “Pressing” pools (indicated as “Reaching (01)”, “Reaching (02)”, etc.). For these neurons we have adopted a deictic representation where the target of the action is provided by the visual system. The main reason for choosing this multiplicity was to study the spontaneous formation of goal specificity among motor neurons, a phenomenon that has been experimentally clearly shown to exist (Bonini et al., 2011; Fogassi et al., 2005).

Similarly to their sensory counterpart, motor pools are also connected in a one-to-one fashion to pools in the striatum layer and in the STN layer of the basal ganglia. In this implementation we have isolated the channels coding the six eye movements from those coding the twelve hand actions in order to better match neuroanatomical evidence which indicates that motor, oculomotor, "prefrontal", and "limbic" signals belong to segregated cortico-striatal circuits (Alexander et al., 1986). The weights of these connections, as well as those from sensory areas, are initially set to low random values, and are learned during the experiment.

Pools in the BG layers of our network encode the same elementary actions as in the motor layer and belong to channels that traverse all layers forming a double inhibition architecture (Gurney, Prescott, & Redgrave, 2001). More precisely, pools in the striatum ($\mathrm{STR_{D1}}$ and $\mathrm{STR_{D2}}$) that are activated at various levels by the input from the cortex representing competing saliences inhibit the corresponding pools in the SNr and the GPe layers, while pools in the STN receive excitatory input from the cortex and in turn excite all the pools in the SNr and the GPe. The active pools in the SNr (which correspond to the "non-winning" channels) inhibit the corresponding pools in the thalamus. Finally, the single output from the thalamus returns to the cortex selectively enhancing the activity of the corresponding pool, which in turn projects back to the basal ganglia creating a self-sustaining loop that converges toward a stable decision.

There are some important points to take into account. The first one is the fact that striatal neurons have been observed to be mostly in a hyperpolarized "down state" or in a depolarized "up state" (Wilson & Kawaguchi, 1996), possibly produced by a variation of the tonic level of dopamine (see Eq. (2)) or by a decrease/enhancement of L-type Ca^{2+} currents (Hernández-López, Bargas, Surmeier, Reyes, & Galarraga, 1997). In our model we exploited this characteristic, hypothesizing that during the exploration phases of the experiment striatal neurons are brought into the "up state" producing low random firing activity (of the order of 10–15 spk/s), while in default condition the BG does not autonomously evoke motor activity.

The second is that since the thalamus receives only inhibitory signals from the SNr, it must possess an additional source (internal or external) of tonic activity which, when disinhibited, propagates to the cortex to activate the neurons therein. In the default state this output is inhibited by the activity of the SNr, which thus must be tonically active. In order to be able to function, this circuit has been endowed with an external source of activity (Bias) that provides a constant input signal to the TH, the STN, and the SNr.

3.2. The neuron model

Neurons are described by a leaky integrate-and-fire model (Dayan & Abbott, 2001) with dynamic channels activated by AMPA, NMDA, and GABA receptors. Below threshold, the membrane potential $V(t)$ of a cell is described by

$$\tau_m \frac{dV(t)}{dt} = -(V(t) - V_L) + R_m \cdot I_{syn}(t), \qquad (1)$$

where $V_L = -70$ mV is the resting (leak) potential, R_m the membrane resistance, and τ_m the membrane time constant.

When the membrane potential $V(t)$ reaches the threshold $V_{thr} = -55$ mV, a spike is generated, the membrane potential is reset to its default value $V_{reset} = -59$ mV, and for a period $t_{ref} = 3.0$ ms the neuron cannot emit another spike.

Following mostly (Humphries, Stewart, & Gurney, 2006), we have chosen, as average values, $R_{cortex} = 50\ \mathrm{M\Omega}$ and $\tau_{cortex} = 30$ ms, $R_{striatum} = 42\ \mathrm{M\Omega}$ and $\tau_{striatum} = 25$ ms, $R_{STN} = 18\ \mathrm{M\Omega}$ and $\tau_{STN} = 6$ ms, $R_{SNr} = 112\ \mathrm{M\Omega}$ and $\tau_{SNr} = 8$ ms, and $R_{GP} = 88\ \mathrm{M\Omega}$ and $\tau_{GP} = 14$ ms. In the current implementation, for each neuron of a given type, these parameters have been chosen randomly from a range between -10% and $+10\%$ of the average value reported above.

Finally, $I_{syn}(t)$ represents the total synaptic current flowing into the cell at time t. This current can be expressed as

$$I_{syn}(t) = [I_{AMPA}(t) + I_{NMDA}(t) + I_{GABA}(t)](1 + \alpha \cdot \Delta_{DA}), \qquad (2)$$

i.e. the sum of glutamatergic excitatory components (NMDA and AMPA) and inhibitory components (GABA), modulated by a factor α (the synaptic efficacy) times the variation of dopamine concentration with respect to the default value. We have chosen $\alpha = +0.3$ for $\mathrm{STR_{D1}}$, $\alpha = -0.3$ for $\mathrm{STR_{D2}}$, $\alpha = -0.11$ in STN, $\alpha = -0.15$ for GP (for additional details see Humphries et al., 2006). Increasing the value of Δ_{DA} produces a hypersensitivity of the network to a point that, for Δ_{DA} of the order of 0.5, there is no competition between channels anymore, and all input signals are maximally activated. On the other hand, decreasing the level of tonic dopamine ($\Delta_{DA} = -0.5$) produces a system that is less reactive and that filters out all input signals that are not sufficiently high.

We consider that external excitatory contributions are produced through AMPA receptors, while the excitatory recurrent synaptic currents are produced through AMPA and NMDA receptors (see Fig. 4).

The current generated by each receptor of type X follows the general form (Destexhe, Mainen, & Sejnowski, 1996):

$$I_X(t) = \bar{g}_X \cdot (V(t) - E_X) \cdot \sum_i W_i \cdot r_i, \qquad (3)$$

where $\bar{g}_X$ is the maximal conductance, $V(t)$ is the post-synaptic voltage, E_X is the reversal potential, and r_i and W_i are the fraction of receptors in the open state and the connection strength with pre-synaptic neuron i, respectively. For AMPA and NMDA synapses $E = 0$ mV, and for GABA synapses $E = -70$ mV.

In case of NMDA an additional multiplicative term $B(V)$ has to be added, representing the magnesium block of the receptor channel. This block takes place extremely quickly compared to the other kinetics of the receptor. The block can therefore be accurately modeled as an instantaneous function of voltage (Jahr & Stevens, 1990):

$$B(V) = \left[1 + \frac{\left[\mathrm{Mg}^{2+}\right] \exp(-0.062 \cdot \mathrm{V})}{3.57\ \mathrm{mM}}\right]^{-1}, \qquad (4)$$

where $[\mathrm{Mg}^{2+}]$ is the external magnesium concentration (in our case 1 mM).

The fraction r of the receptors in the open state is well described by the following first-order kinetic equation:

$$\frac{dr}{dt} = \alpha \cdot [T] \cdot (1 - r) - \beta \cdot r. \qquad (5)$$

The values of parameters α, β, and $[T]$ (neurotransmitter concentration) used in our simulations were taken from Destexhe et al. (1996).

The total current I_{syn} flowing into neuron i can also be written as the sum of different components:

$$I_{syn,i} = \sum_j W_{ij}^{(exc)} \cdot R_j + \sum_k W_{ik}^{(inh)} \cdot R_k + \sum_l W_{il}^{(far)} \cdot R_l, \qquad (6)$$

where $W_{ij}^{(exc)}$ and $W_{ik}^{(inh)}$ define the strength of local excitatory and inhibitory connections, respectively, while $W_{il}^{(far)}$ represents the connectivity to neurons belonging to other pools and R_x the contribution of receptor type X (see Eq. (3)).

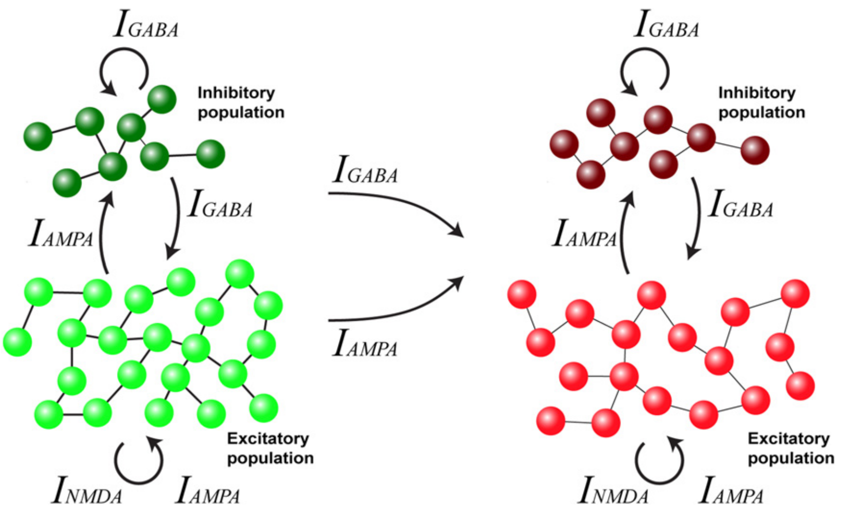


Fig. 4. Schematic representation of two connected neuronal pools. Inhibitory neurons are represented as darker elements while excitatory neurons are represented as lighter elements. Arrows and the corresponding neurotransmitter specify the type of connection between neurons within and outside the pool.

In our implementation, 20% of the neurons in the cortex (i.e. MC, PFC, sensory areas) and TH layers are randomly chosen to be inhibitory. Each neuron is approximately connected to 20% of the neurons in its pool, and to 2% of the neurons belonging to other pools (following the scheme shown in Fig. 3).

In the STR all neurons are inhibitory, in the STN all neurons are excitatory, and in the GPe and in the SNr 80% of the neurons are inhibitory. In these areas the local connectivity is around 5%.

The weights of the local and long-range connections have been obtained through an automatic fitting procedure that had to respect a number of constraints. The first one was that in the default state the cortex, the TH, and the STR (in the "down" state) had to show no activity, while the average firing rate of neurons in the STN was 25 spk/s, in GP 55 spk/s, and in the SNr 70 spk/s (van Albada & Robinson, 2009). On the other hand, in the active state the maximal firing rate of cortical and TH neurons was 100 spk/s, of STR neurons 35 spk/s, of STN neurons 60 spk/s, of GPe neurons 80 spk/s, and of SNr neurons 100 spk/s.

In this way the average local connectivity strength was found to be 10.2 ± 4.1 in the cortex and the TH, 7.0 ± 4.5 in the STR, 61.0 ± 12.0 in the STN, 4.2 ± 4.0 in the GPe and 2.8 ± 2.2 in the SNr. The large number for STR must not surprise as it is due to the low value of R_m of its neurons.

Another constraint for the fitting procedure was that stimulating the basal ganglia (through the STR and the STN) by activating one projecting cortical pool to fire at 80 spk/s and the others at 60 spk/s resulted in only one pool (corresponding to the highest firing channel) to be active in the thalamus. Note that, since the channels form closed loops (cortex → BG → cortex), the activity level of the "wining" channel in the TH is in great part due to recurrent input.

Finally, it was required that stimulating the BG with equal inputs produced no "winning" output.

Given these constraints, the resulting average inter-pool connectivity was $W_{IN-STN} = 0.98$, $W_{IN-STR} = 9.7$, $W_{STR1-SNr} = -40.8$, $W_{STR2-GPe} = -17.3$, $W_{STN-SNr} = 35.86$, $W_{STN-GPe} = 6.7$, $W_{GPe-STN} = -11.8$, $W_{GPe-SNr} = -8.6$, $W_{SNr-TH} = -33.4$, $W_{TH-MC} = 70.2$, $W_{Bias-SNr} = -95.1$, $W_{Bias-STN} = 12.6$.

Note that these values should be taken as indicative, as they depend also on the specific characteristics of the neurons and on the local connectivity.

Fig. 5 shows the spiking activity and the average firing rate as a function of time of 12 neurons in the cortico-basal ganglia circuit belonging to two competing channels (red and blue). The average firing rate, calculated using a sliding window of 40 ms, is superimposed onto the rasterplot of the spikes of the single neurons recoded during ten trials. The spontaneous activity in the striatum has been switched on here only between $t = 400$ and $t = 800$ ms for illustrative purposes.

All together the network comprises 146 pools, each of which is composed of 100 neurons, for a total of 14,600 neurons.

In order to avoid numerical instabilities associated with fast spiking activity, each neuron was simulated with a time step of 0.1 ms using the first-order Euler method.

3.3. The learning rule

In the present model, learning takes place through the update of the connections' strength between the neurons in the sensory, the striatum and the STN layers, as well as between the neurons in the PFC and in the motor layers. The implemented learning rule is the combination of two components: the first is spike-timing-dependent plasticity (STDP) (Abbott & Nelson, 2000; Bi & Poo, 2001; Kepecs, van Rossum, Song, & Tegner, 2002; Song, Miller, & Abbott, 2000) and the second is eligibility traces (Izhekevich, 2007; Singh & Sutton, 1996).

According to the first component, the connection between two neurons is reinforced if the post-synaptic neuron fires shortly after the pre-synaptic neuron, while it is weakened when the pre-synaptic neuron fires after the post-synaptic neuron. The term $F(\Delta t)$ that takes into account the dependency of the synaptic efficacy on the past activity the neuron (Gonzalez-Burgos, Krimer, Urban, Barrionuevo, & Lewis, 2004; Zucker & Regehr, 2002) can be written as

$$F(\Delta t) = \begin{cases} A_+ \cdot \exp(-\Delta t/\tau_+) & \text{if } \Delta t > 0 \\ A_- \cdot \exp(\Delta t/\tau -) & \text{if } \Delta t < 0, \end{cases} \tag{7}$$

with $\Delta t = t_j - t_i$, where t_i and t_j are the time of occurrence of the pre-synaptic and post-synaptic spikes, respectively. A_+ and A_- determine the maximum amount of synaptic modification (in our case 1), and the parameters τ_+ and τ_- determine the ranges of pre-to-post-synaptic interspike intervals over which synaptic change occurs (in our case $\tau_+ = 8$ ms and $\tau_- = 10$ ms).

Finally, the complete equations that include also the second component, i.e. the eligibility traces, are the following:

$$\begin{cases} \tau_E \dfrac{dE_{pre,post}}{dt} = -E_{pre,post} + F(\Delta t) \\ \tau_W \dfrac{dW_{pre,post}}{dt} = -W_{pre,post} + \alpha \cdot R \cdot E_{pre,post}, \end{cases} \tag{8}$$

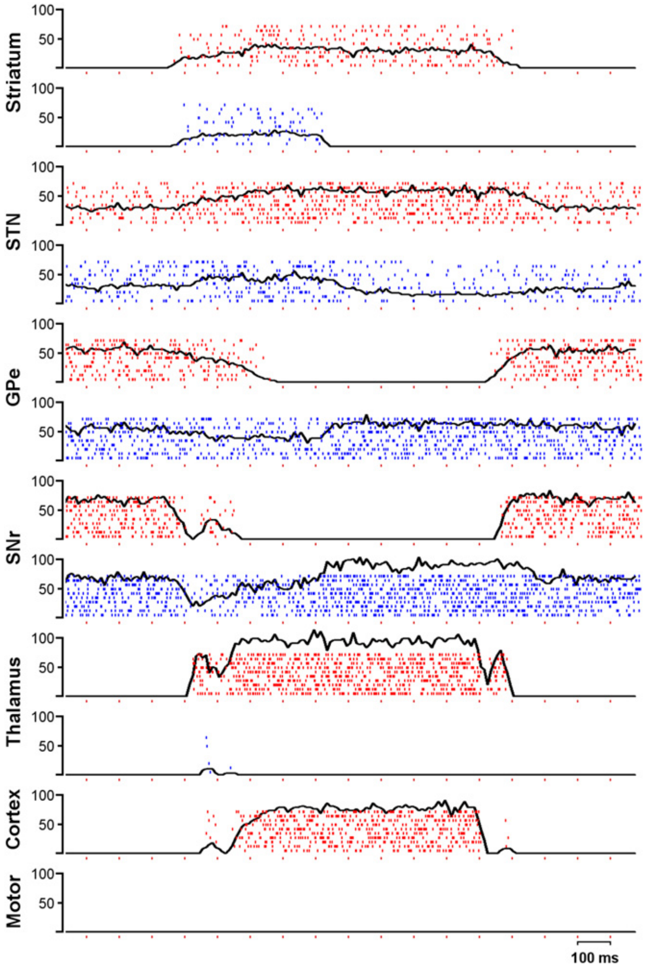


Fig. 5. Spike rasterplots and mean firing rates of the activity of 12 neurons (belonging to two channels: red and blue) of the cortico-basal ganglia circuit. Spontaneous activity in the striatum (active between 400 and 800 ms) propagates along each channel in the other areas of the BG. Here a double-inhibition circuit enhances the most active channel, eliciting only for this a high firing rate in the thalamus and in the motor cortex, which then return to the striatum amplifying the effect (at $t = 1400$ the network is reset). For each channel and for each nucleus one representative neuron has been chosen. The average firing rate is calculated for ten repetitions and a sliding window of 40 ms every 10 ms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where $W_{pre,post}$ is the connection strength between the pre and the post neuron, α determines the learning rate, R is the reward, $E_{pre,post}$ is the eligibility trace of the co-activation of the pre and the post neuron, and τ_E and τ_W are the decay time constants of the eligibility trace (2 s) and of the weights (10 s), respectively. An important effect of the last term is to prevent an unbounded increase in the value of the connection weights.

4. Results

4.1. Learning phase

The experiment simulated in this paper is divided into two phases. The first phase, during which the agent has to learn to solve the task described above, consists of 3000 repetitions of the basic task in its three variants (the three lights).

Each trial lasts either for 5 seconds or until the agent reaches the goal. In the latter case the procedure is stopped and the network receives a reward (simulated here as a phasic increase of the dopamine level), which causes a modification of plastic synaptic weights proportionally to the intensity of the trace of the neurons' past co-activation according to Eq. (8). On the other hand, if the final configuration is not reached within the time limit, a penalty is given.

In this implementation, we have assumed that the agent is already capable of executing elementary motor acts such as reaching, pressing, and foveating, but does not know in which order to execute them to most successfully reach the goal. Additionally, in order to increase the level of realism of the simulations, we assumed that the execution of any motor command is not instantaneous but takes 200 ms to complete (during which the network dynamics continue to evolve). This has the effect of adding the temporal dimension to the constraints of the task: not only has the agent to learn the correct sequence, but also the correct timing between motor acts. Executing the right actions too early or too late produces unsuccessful motor sequences. Here, for example, as a result of overtraining, the synaptic weights become so strong that motor acts are activated too early when not all the necessary conditions are met. To overcome this problem, the reward policy employed in this model entails that in the case of failure a small punishment is administrated (in form of a drop of the dopamine concentration), which causes the connection strengths to be decreased, producing a slowdown in the motor activation and thus a return to the correct timing.

After the execution is completed, the state of the agent (i.e. its eyes and hand position) and that of the environment are updated, and a confirmation signal is sent to the basal ganglia in order to reset the corresponding neural races.

The activity of multiple components of the network during the execution of two typical trials, one at early stages and one after 1500 trials, are represented in Fig. 6 in the left and right panel, respectively. The upper group (SI∗) represents the various sensory inputs (visual and proprioceptive), and the middle group (STR∗) the activity of the striatum, while the lower group (MC∗) represents the activity of the motor cortex. As can be seen, at the beginning of the experiment (left panel) the agent chooses the motor acts in a random way, executing actions that are not necessary. As learning proceeds (right panel), the agent becomes more skilled and able to complete the correct sequence in the minimal number of acts.

One important point in this experimental setup is the order in which the learning trials are administered and how much reward is given to the agent.

We recall that, in the present model, the reward takes the form of an increase of the dopamine level in the network, which modifies the rate of change of the connection strength of plastic synapses (see Eq. (8)).

From an operative point of view, the strategy utilized to solve the task is the classic "trial and error" method: driven initially only by noise (due to internal stochastic processes or to external inputs), the basal ganglia randomly generate sequences of motor acts to execute. Those sequences that lead to the requested goal states elicit a reward that produces a reinforcement of the associations between states and responses. As the training proceeds, the contribution of the learned association in the selection process becomes greater than that due to noise (which remains constant when present).

One first important result is that this system cannot successfully learn if the tasks are administered in batches with always the same light to turn on. The reason for this is that, due to the random nature of the choices, the system finds many paths in the action space that lead to the desired goal. Several of these intersect or overlap with trajectories that would lead to other goals, but the learning algorithm "rewires" them all to the current goal. If one goal is trained too often (batch), this modification becomes so

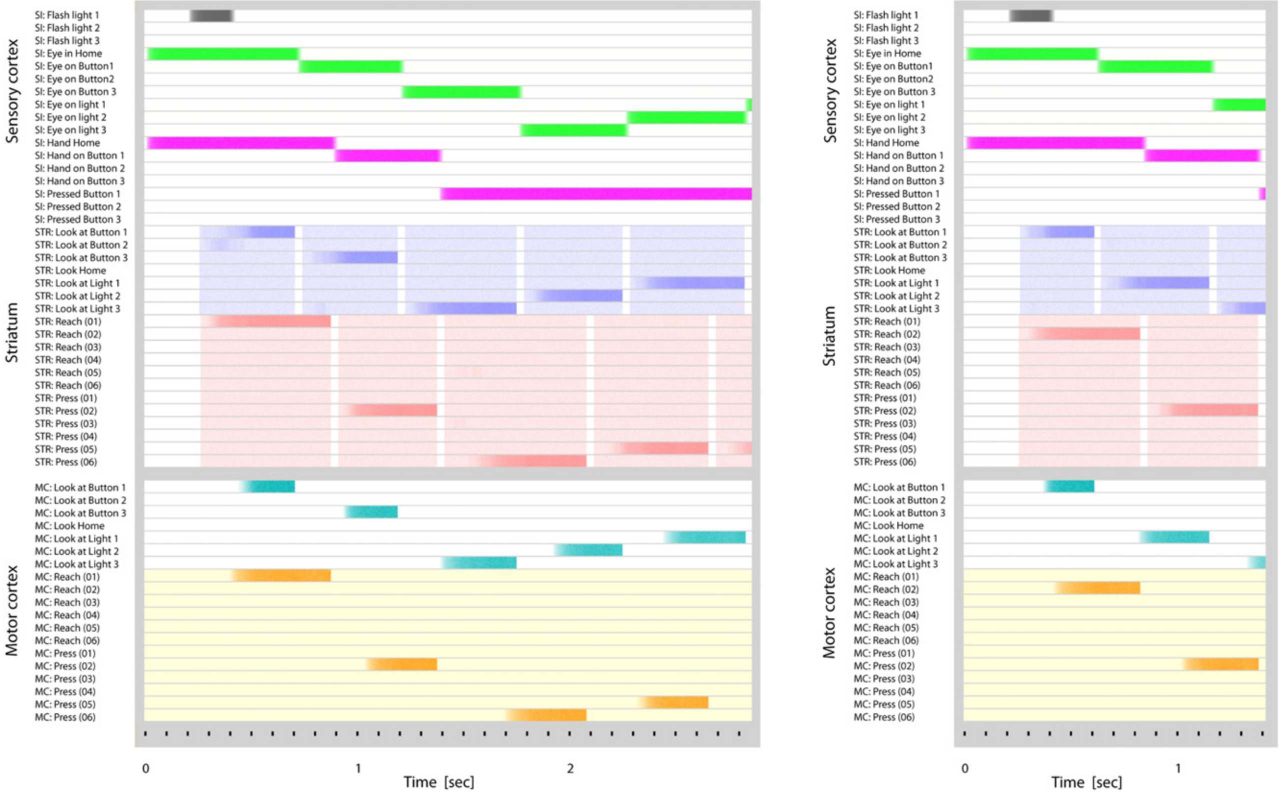


Fig. 6. Two examples of neural activity of different pools as a function of time within various areas of the striato-cortical loops (for clarity, not all areas have been shown). Darker tones indicate higher firing rates. Left panel: successful trial (for switching on light 1) at the early stages of the experiment; right panel: same task but after 1500 trials. Abbreviations: SI, sensory input; STR: striatum; MC, motor cortex.

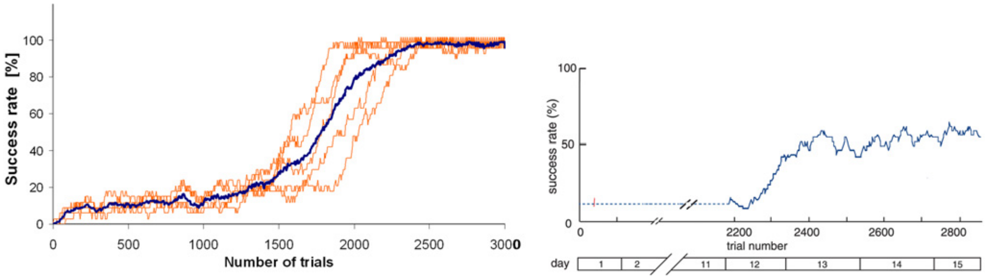


Fig. 7. Left panel: success rate of the agent in five consecutive experiments (the thick blue line is the average performance) as a function of the number of trials in a task in which all three button-light associations have to be learned. The success rate is calculated as the moving average of 100 trials. For comparison, the right panel shows the performance of a monkey learning by trial and error to use a rake to retrieve food.
Source: Modified from Ishibashi et al. (2000).

strong that the system can subsequently reach another goal only if the previous associations are forgotten.

To obviate this problem we have utilized two training methods.

In the first one, we force the agent to repeat the same trial (i.e. turning on the same light) until it is able to reach the goal. This ensures that all goals receive the same training and reach the same performance. The corresponding results are shown in Figs. 7 and 8.

As can be seen in the left panel of Fig. 7, performance is initially very low and almost constant; then, after 1500 trials, the success rate rapidly gradually increases to the maximum level (33.3% for each goal).

Fig. 8 depicts the time taken by the agent to complete the task as a function of the number of learning trials. As can be noted, not only does the agent increase its performance (deducible by the fact that the orange dots become denser on the x axis), but it improves its average completion time (blue curve). This is due to the fact that with exercise the agent develops a tendency to store and execute shorter sequences. This is easily explained by considering that on the one hand these are statistically more frequent and on the other they allow receiving a higher reward because their eligibility trace is stronger at the moment of dopamine administration.

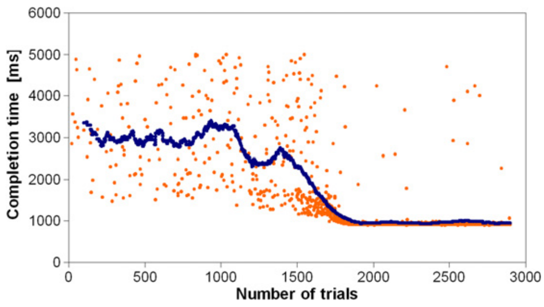


Fig. 8. Performance of the agent measured as the time necessary to complete successful trials as a function of the executed trials. The blue curve represents the moving average of the completion time over a window of 200 trials.

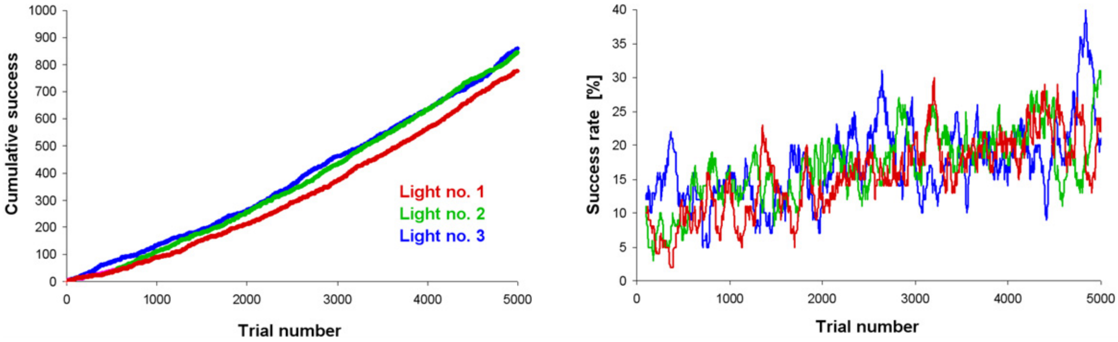


Fig. 9. Performance of the agent as a function of time when the goals are presented in random order and the reward is inversely proportional to the success rate for that goal. Left panel: cumulative successes; right panel: success rate calculated over a sliding window of 100 trials.

The second method consists of running all trial conditions in random order, allowing only one attempt per trial, and administrating an amount of dopamine that is inversely proportional to the recent success rate for each goal. In this way less frequent results are rewarded more, thus avoiding a "repetition bias" that in our case is harmful for a balanced learning. The rationale behind this training method derives from neurophysiological evidence showing that, as animals become more skilled and complete specific tasks more frequently, their successes become less surprising, and this causes the release of a smaller amount dopamine (Ljungberg, Apicella, & Schultz, 1992; Schultz, Dayan, & Montague, 1997). Results are shown in Fig. 9. As can be seen, the performance does not reach the optimal level as in the previous case, probably because noise has too big an influence on the choices and implicitly also conditions the amount of reward. In all conditions, at initialization the connection strength of all plastic synapses is set to small random values. These weights are autonomously learned during the training period according to the rules described above.

We want to point out that, in this model, learning takes place in different areas and thus at different levels of abstractness/concreteness (see Fig. 10). At the lower level, which we here assume to be the basal ganglia, stimulus–response associations are formed, that, although simple, combined allow the production of complex behavior.

At the higher level, i.e. between the prefrontal and the motor cortex, links from intention (goal) pools to elementary actions are learned. The PFC–MC path plays a fundamental role in goal-directed motor control as it allows overriding habitual responses (see the next section). It is important to note that in this model these connections do not specify the temporal order of the execution of the single motor acts but only which motor pools will have to be activated in order to reach a desired goal.

Fig. 11 represents the strength of the connections between pools in different layers of the circuit at the end of the learning period: red squares represent excitatory connections between sensory areas and motor representations in the striatum, while green squares represent the connectivity between prefrontal areas and the pools in the motor cortex. As can be seen, the network is clearly able to learn the correct associations between states and actions that allow it to successfully complete all three variants of the task. For example, analyzing the graph, it is possible to infer that if the agent experiences a "Flashed light 1" situation, it will activate most highly the "Look at button 1" action (in the graph the biggest square in the first column). Note that when light 1 flashes the PFC will understand that Goal 1 has to be pursued so it will send an input to the "Look at button 1" pool in the motor cortex which in turn will propagate the activity to the corresponding basal ganglia pool. There exist also weaker "spurious" connections due to random activations that produce longer and less frequent sequences which are nevertheless correct and are therefore reinforced as well.

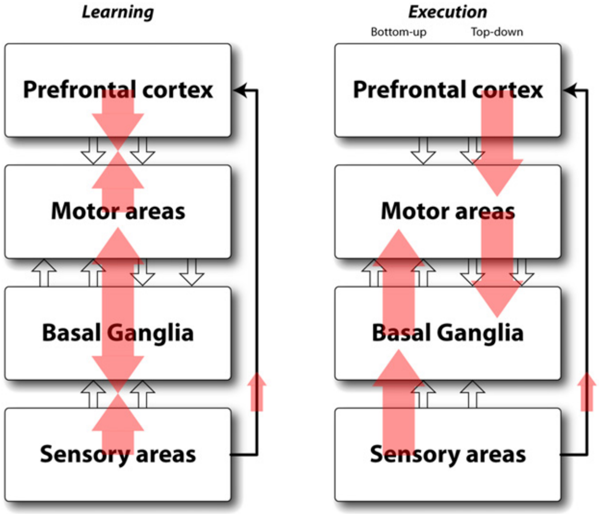


Fig. 10. Left panel: Schematic representation of the flow of neural activity during learning. The (random) activity originating in the BG on the one hand is confronted with incoming sensory signals producing the update of the connection weights, and on the other reaches the motor areas where it is confronted with PFC signals allowing the formation of goal–action associations. Right panel: During habitual action, execution sensory signals elicit the response of the BG which sends motor commands to the MC. During goal-directed control, the prefrontal cortex activates specific motor representations in MC which override the bottom-up signals from BG.

Interestingly, simulations reveal that usually for each goal only specific hand-related channels are recruited. More precisely, when the agent has to turn on light 1, results show (see Fig. 11) that the channels encoding "Looking at button 1", "Reaching (03)", "Pressing (03)", and "Looking at light 1" are activated, while, when it has to turn on light 2, the network uses "Looking at button 2", "Reaching (03)", "Pressing (05)", and "Looking at light 2". Finally, when it has to turn on light 3 it uses "Looking at button 3", "Reaching (03)", "Pressing (04)", and "Looking at light 3". What should be noted is the fact that, although all three sequences contain "Pressing" motor acts, which are undistinguishable from the information content point of view as they code target-aspecific actions, the structure and the learning rule adopted in this model produce a system that differentiates the motor acts on the base of the final goal of the action sequences (i.e. light 1, 2 or 3). This result is important because the same goal specificity has been recently observed in parietal and premotor cortex neurons (Bonini et al., 2011; Fogassi et al., 2005).

Another interesting result that was obtained from the simulations is that, besides the motor sequence consisting of "Look at button X", "Reach it", "Press it", and then "Look at light X", which was expected to be optimal, the system found another solution which proved to be more rewarding (and thus better): "Look at button X",

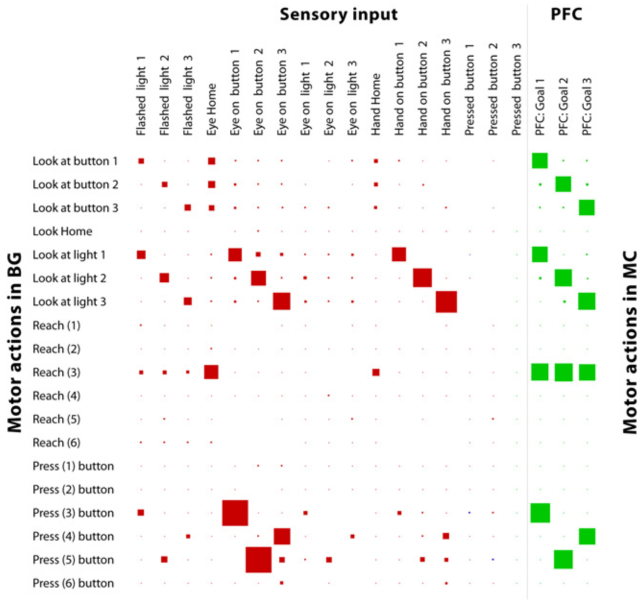


Fig. 11. Hinton graph of the typical connectivity between different pools in the network. Red squares indicate connections between sensory areas and striatum, while green squares (on the right side) indicate connections between prefrontal cortex and motor cortex pools. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

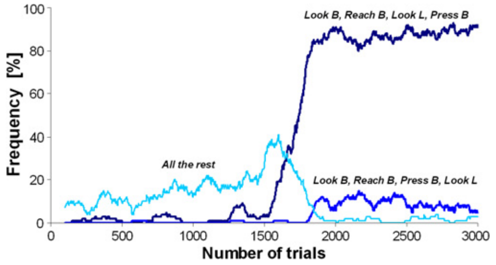


Fig. 12. Frequency of utilization of specific (successful) motor sequences. Initially the agent randomly generates motor sequences of variable composition and length, of which very few are successful (see “All the rest” curve). As learning proceeds, the agent discovers shorter and thus more rewarding sequences which are used much more often than to all the others. Abbreviations: Look B = Look button, Reach B = reach button, Press B = press button, Look L = Look light.

“Reach it”, “Look at light X”, and the “Press button X” (see Fig. 12). The advantage of this sequence is not the number of motor acts, which remains the same, but the order: while the agent is executing an action with one effector, its brain can prepare the action to execute with the other effector, because the cortico-basal ganglia loops involved do not interact blocking each other. The duration of this second solution is shorter, and thus the eligibility of the neural activity is higher.

Note that, besides the two most frequent sequences, that make up almost 50% of all the successfully executed sequences (during the training phase), the other valid solutions are obtained from the main solutions with interspersed random motor acts. For example, the third most common sequence is “Look at button X”, “Reach (it)”, “Look at button Y”, “Press (button X)”, and “Look at light X”.

4.2. Second phase: cue reversal

In this phase of the experiment, we wanted to test how the system behaves when the task suddenly changes (thus requiring goal-directed behavior). To this aim, after the usual initial training phase of 3000 trials, we remapped the light–button correspondence, assuming us to be able, if needed, to instruct the animal about this change (for example, through the appearance of an additional cue signal).

In particular, the change consisted of the fact that a brief flash of light 1 now meant switch on light 2, one of light 2 meant switch on light 3, and one of light 3 meant switch on light 1.

In human subjects, these rules are understood immediately, stored in working memory, and employed to solve the new task. In monkeys, this takes more time, but eventually they are able to learn the switching with relative ease. In our neural network, we assumed that this kind of one-shot remapping takes place somewhere in the prefrontal cortex, and the effect is that the activation of a neural pool in the sensory layer automatically elicits the activation of the corresponding new goal pool in the PFC layer. We recall that in these simulations the PFC does not learn the associations between cues and goals, which are assumed to be already known at the time of the experiment. Instead it only learns the links between goal representations and actions, thus being able to bias the activation of specific neuronal pools directly in the motor cortex depending on the sensory input and context.

To test the contribution of the prefrontal cortex, we devised two variants of this experiment: one in which we imagined changing the cue–goal association without informing the agent, and one in which we did. The result is represented in Fig. 13, which shows the agent's performance as a function of time during both phases of the experiment. The cue reversal takes places at $t = 3000$ ms, and immediately the agent's performance deteriorates. Interestingly though, in one situation (“with PFC control”) the recovery is relatively rapid. This is due to the fact that, when the association rules between cues and action sequences are remapped, the PFC instantly adjusts to the new situation (because we assume that the agent understands and immediately cognitively assimilates the new rules) and takes control of the action, while the basal ganglia needs much more time to learn the new sensory-motor contingencies.

In the other case (“without PFC control”), the rules learned by the PFC suddenly become useless (or even hindering), and the network has to relearn the whole task again.

Fig. 14, similarly to Fig. 8, shows the time taken by the agent to complete a single task as a function of the number of executed trials. In this case, both experimental phases are represented subsequently. The cue reversal point is for $t = 3000$ ms. The right panel illustrates the behavior when both the BG and the PFC have to learn the new configuration from scratch, while the left panel represents the situation in which the PFC is able to take predominant control of the actions. It is interesting to note that, in this case, after the cue reversal the completion times do not reach the previous optimal level. This occurs because, once the agent is able to regularly complete the trials mostly due to the strong contribution from the PFC, there is no punishment for unsuccessful sequences, thus impeding the correction of non-optimal solutions. On the other hand, in the experimental condition without PFC control, the connections are basically all forgotten and completely relearned, thus erasing all traces of past experiences.

We want to underline the fact that the first variant has been devised not to model problem-solving capabilities in monkeys or even less in humans, as the underlying mechanisms differ radically from what described here, but instead to study stimulus–response formation and habitual motor learning.

What is the detailed neuronal dynamics that underlies the execution of the tasks described above?

Fig. 15 shows the typical behavior of 14 neurons belonging to two competing “channels” (red and blue) belonging to the 18-channel cortico-basal ganglia circuit implemented in this work. The left panel depicts the situation in which the random activity within the striatum alone elicits a race between neural channels, the final result of which is the activation of thalamus and subsequently motor cortex neurons corresponding to the “winning”

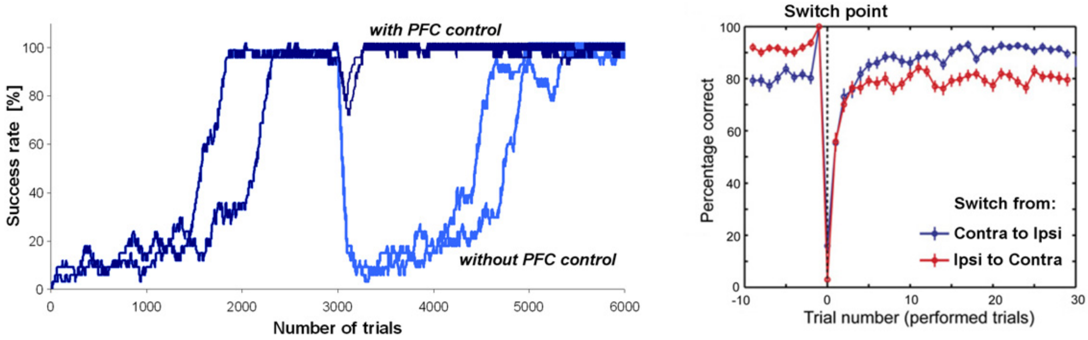


Fig. 13. Left panel: performance of the agent in four trials with a cue switch at $t = 3000$ ms. The dark blue curves represent the performance of the agent that was informed about the switch (thus implying an active participation of the prefrontal cortex), while the dark blue curve shows the performance of the agent that cannot use the control from the PFC. For comparison to the PFC controlled condition, the right panel reports the performance of a monkey in a response-switch task in which it had to change from cue-ipsilateral to cue-contralateral saccades and vice versa, based on the amount of received reward. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Source: Modified from Johnston et al. (2007).

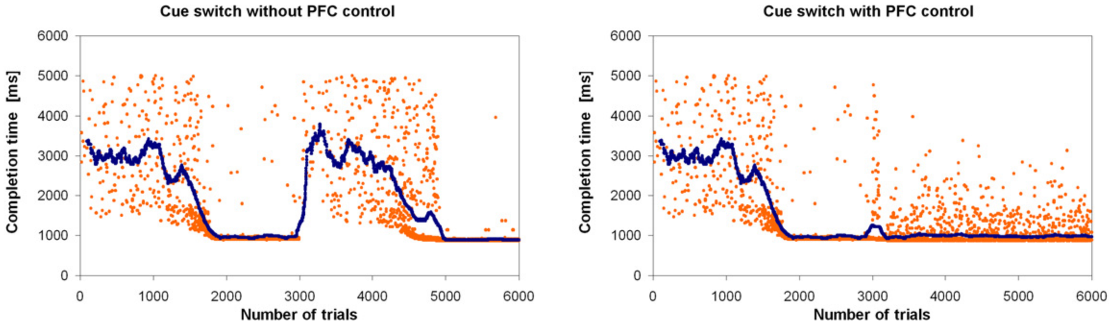


Fig. 14. Performance of the agent measured as the time necessary to complete a successful task as a function of the number executed trials. The cue switch takes place at $t = 3500$ ms. The blue curve represents the average over 100 trials. Left panel: the agent is not informed about the cue switch; right panel: the agent is informed.

channel (in this case the red one). As can be seen, although the activity of the two neurons in the striatum is initially very similar, the "selection" ability of the basal ganglia ensures that at the output stage one channel will clearly dominate over the others.

The middle panel shows the behavior of the network when a sensory input (top inset, blue curve) arrives at a specific pool in the basal ganglia: the corresponding channel rapidly dominates the others and the resulting output activates the motor cortex.

Finally, the right panel illustrates the neural dynamics for the cue reversal task where the action control is taken over by the prefrontal cortex. In this example, in addition to the sensory input (blue channel) a strong signal from the PFC (not shown), simulating a goal-directed decision, directly activates the motor cortex (red curve in the bottom inset) which in turn back-projects to the striatum.

This behavior is in line with experimental results showing that during goal-directed actions the prefrontal cortex overrides automatic stimulus-driven responses (Ashby, Turner, & Horvitz, 2010; de Wit, Barker, Dickinson, & Cools, 2011; Redgrave et al., 2010). In this network, we have realized this capability by assuming that the PFC also sends a strong input to the STN through the so-called "hyperdirect" pathway (Aron & Poldrack, 2006; Eagle et al., 2008), which has the effect of increasing the activity of inhibitory neurons in the basal ganglia, thus lowering the activity of all channels and allowing the goal-directed (cognitive) selection of the motor acts to execute.

This overriding capacity has a profound impact on the performance of the system when combinations of habitual and goal-directed behaviors are required, or when the two types of behavior are conflicting. This effect cannot be easily seen during automatic task execution, as the choice of the PFC layer is either absent or congruent with the outcome of the basal ganglia circuit. On the contrary, during goal-directed tasks such as the one described above, PFC and BG responses are (initially) in contrast; therefore this effect can be detected.

5. Discussion and conclusion

The work presented in this article provides a detailed model of the mechanisms that underlie unsupervised learning of motor sequences in the basal ganglia. In particular, it describes how a spiking neuron network mimicking the hierarchical organization of the cortico-basal ganglia circuit is able to acquire and recall a repertoire of actions in response to specific stimuli, and how habitual and goal-directed motor controllers may interact within the brain.

One important aspect is that this system uses a delayed associative learning rule—in this case spike-timing-dependent plasticity with eligibility traces—rather than temporal difference (TD) learning (Suri & Schultz, 1998; Sutton, 1988). TD learning requires on the one hand an ongoing representation of predicted reward at each time step, and on the other the calculation of the error between the observed and the expected state (which is used to update the synaptic weights). In this respect, the reinforcement learning-based model described here is simpler, as it uses only the reward signal at the end of the trial to modify the connections between pools of neurons representing sensory and motor information weighted by the trace of their past co-activation. Here, the actual sequencing between the motor primitives within the basal ganglia is not performed by learned pairwise associations between each primitive (as may happen in other areas) but rather by learning which motor primitives should be selected in correspondence of a particular configuration of the agent and the environment.

The second interesting aspect is that the analysis of the connectivity matrix has revealed that the activation of motor

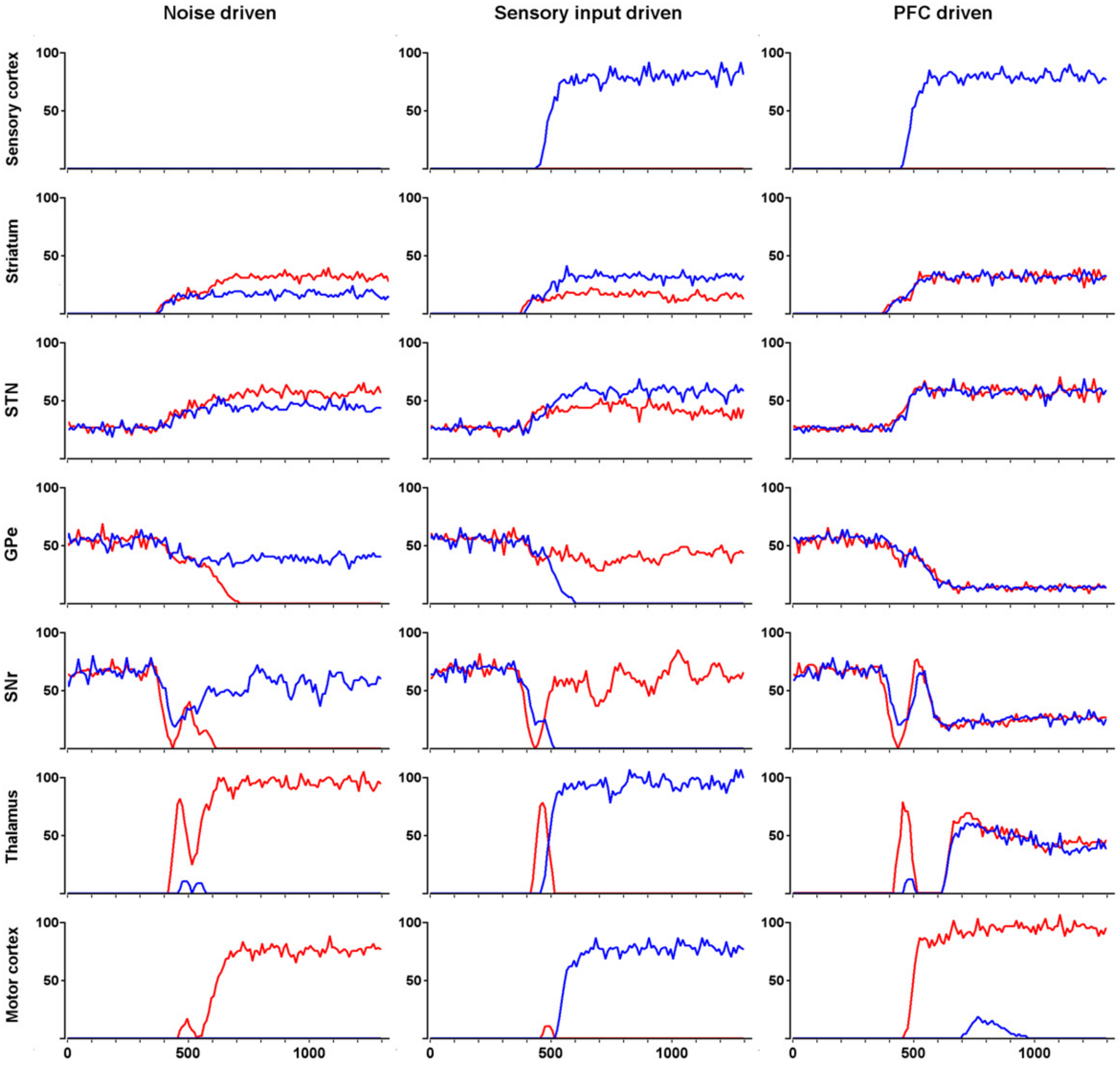


Fig. 15. Mean firing rate of 14 neurons belonging to two competing channels (red and blue) within different brain areas. Left panel: the spontaneous activity of the striatal neurons (starting at $t = 350$ ms) eventually elicits a high firing rate of the thalamus and motor cortex neurons of the red channel (for $t > 550$ ms). Middle panel: in this case, in addition to the spontaneous activity, a strong sensory input leads to the activation of the blue channel. Right panel: in addition to the spontaneous and sensory activity, a strong input from the PFC (not shown) to the motor cortex and to the STN at $t = 450$ ms leads to the activation of the red channel. The input to the STN has the function of deactivating the races within the basal ganglia and let the external activity dominate. Average activity was calculated on 15 repetitions and a sliding window of 40 ms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

primitives often shows a selectivity for the end-goal of the action in which they embedded. This result is in agreement with recent neurophysiological findings (Bonini et al., 2011; Fogassi et al., 2005) and is in keeping with the hypothesis that other parts of the brain, such as the parietal and the premotor cortices, store and recall goal-specific chains of subsequential motor acts (Chersi, Ferrari, & Fogassi, 2011).

The third central point of this work is the distinction between the mechanisms underlying goal-directed actions and stimulus-driven habits, the two main categories of instrumental behavior. The analysis of the network interactions that produce switches between goal-directed actions and habits may have important implications for the study of skill learning, addiction, and various clinical disorders linked to basal ganglia dysfunctions. To this aim, the present model proposes a plausible mechanism of how goal-directed motor commands deriving from the prefrontal cortex interact with stimulus-driven responses originating in the basal ganglia. One of the hypotheses on which the model is based is that motor neurons are organized in independent channels that form loops between the cortex and the basal ganglia (Alexander et al., 1986; Middleton & Strick, 2000). In this view, voluntary commands deriving from the PFC and propagating to the MC together with sensory inputs elicit targeted activity in the cortex which is conveyed to the various components of the basal ganglia. The BG circuits described in this paper work as a selection and "filtering" mechanism between competing signals present in different channels projecting from sensory and motor areas.

In addition to this functionality, in the present model the basal ganglia possess also intrinsic random activity which works as a source of variability for the action selection mechanism in the case of absence of strong external inputs. This ensures autonomous operation even when sensory inputs are ambiguous or there are no salient inputs selective for a specific action.

From the results achieved in this work, we can draw some important conclusions and predictions.

First, simulations have shown that the subdivision of the basal ganglia into sectors related to different effectors is highly beneficial for the performance of the brain, as the BG can simultaneously process input signals of different kinds and activate several motor acts in parallel.

Second, we have observed in our model that, at the beginning of the experiment, striatal neurons tend to fire often and randomly, while, as learning proceeds, neuronal responses become targeted and more seldom. This is due to the fact that neural representations become highly correlated with specific sensory inputs, and neurons are activated only when required. These results lead us to predict that such a behavior may be present in the striatum and possibly other BG nuclei of real brains, where one should observe a progressive firing rate adaptation and specification on long time scales reflecting the acquisition of a low-level representation of the task.

Third, neurophysiological data have shown that a large part of the motor neurons in the parietal and premotor cortex also encode the goal of the action sequence. Given the direct connections between these areas and the basal ganglia and their channeled organization, we predict also finding a certain degree of goal specificity in some nuclei of the basal ganglia.

5.1. Comparison with other models

In the last decade there has been a considerable increase in the number of computational models of the basal ganglia, ranging from a more abstract and system-level description to detailed neuronal simulations.

In particular, in their work, Beiser and Houk (1998) describe a mechanism for the encoding of sequences that emerges as a result of macro-reorganizations of the circuit comprising the prefrontal cortex, the basal ganglia, and the thalamus. They find that neurons in their network show cue-related and sequence-related characteristics similarly to the responses recorded in the prefrontal cortex.

More recently, Frank, Seeberger, and O'reilly (2004) proposed a model that combines action selection and reinforcement learning. They predicted that Parkinson's patients off medication are better at learning to avoid negative outcome choices than they are at learning from positive outcomes, while dopamine medication makes patients more sensitive to positive than negative outcomes.

Furthermore, the role of the basal ganglia in exploration in the framework of reinforcement learning has been studied by the group of Chakravarthy (Sridharan, Prashanth, & Chakravarthy, 2006). They proposed that the STN–GPe circuit generates chaotic dynamics that is used to explore the state-space, and the dopamine-based reward signals from substantia nigra pars compact regulates the striatal–GPe connection weights leading to reinforcement learning.

The main difference between our model and the ones cited above is the use of a biologically more accurate architecture and realistic neuronal dynamics, which allows us to link single neuron behavior to the properties of wider brain systems and to behavioral results.

A work of particular interest that follows this approach, thus being a reference point for us, is that of Humphries et al. (2006) (but see also Schroll, Vitay, & Hamker, 2012). In this study, the authors present a biologically detailed model of the BG that is able to reproduce the firing rates and oscillations observed in animal brains, demonstrated selection and switching between signals representing the salience of actions.

Compared to this and the other works, the major improvement of our model was the addition of thalamo-cortical feedback and cortico-striatal plasticity at the neuronal level. These extensions, together with the validation in a realistic scenario that included timing constraints, unsupervised learning, and switching from habitual to goal-directed behavior, represents in our view a substantial advancement in the modeling of the basal ganglia.

To conclude, we would like to remark on the fact that, given the enormous complexity of the cortico-basal ganglia network, the development of a faithful model presents considerable difficulties. Nevertheless, we feel that this study, although limited in detail and realism, has highlighted important aspects about the relations between cortical and inner brain structures, and about the learning and execution of habitual and goal-directed actions and their mutual interaction.

Future research will aim at building a more realistic model with a richer and more detailed goal-directed control system and apply it to control an embodied agent or a robot.

Acknowledgments

We would like to thank Prof. Kevin Gurney and Prof. Peter Redgrave for the fruitful discussions and their invaluable help in the development of this model. This research has received funding from the European Commission 7th Framework Programme (FP7/2007-2013), "Challenge 2 - Cognitive Systems, Interaction, Robotics", with the projects "IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots", grant agreement No. ICT-IP-231722, and "Goal-Leaders - Goal-directed, Adaptive Builder Robots", grant agreement No. FP7-ICT-270108. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Author contributions: Study design: FC GB MM GP. Performed the experiments: FC. Collected the data: FC. Wrote the paper: FC. Computational modeling: FC.

References

Abbott, L. F., & Nelson, S. (2000). Synaptic plasticity: taming the beast. *Nature Neuroscience*, *3*, 1178–1183.

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

Aron, A. R., & Poldrack, R. A. (2006). Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus. *The Journal of Neuroscience*, *26*(9), 2424–2433. http://dx.doi.org/10.1523/JNEUROSCI.4682-05.2006.

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, *14*(5), 208–215.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4–5), 407–419. http://dx.doi.org/10.1016/S0028-3908(98)00033-1.

Beiser, D. G., & Houk, J. C. (1998). Model of cortical-basal ganglionic processing: encoding the serial order of sensory events. *Journal of Neurophysiology*, *79*(6), 3168–3188.

Bi, G. Q., & Poo, M. M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*, 139–166.

Bonini, L., Ugolotti Serventi, F., Simone, L., Rozzi, S., Ferrari, P. F., & Fogassi, L. (2011). Grasping neurons of monkey parietal and premotor cortices encode action goals at distinct levels of abstraction during complex action sequences. *The Journal of Neuroscience*, *31*(15), 5876–5886.

Chersi, F., Ferrari, P. F., & Fogassi, L. (2011). Neuronal chains for actions in the parietal lobe: a computational model. *PloS One*, *6*(11), e27652. http://dx.doi.org/10.1371/journal.pone.0027652.

Chersi, F., Mirolli, M., Gurney, K., Redgrave, P., & Baldassarre, G. (2010). Goal-directed motor sequence learning based on multiple cortico-basal ganglia loops, Soc. Neurosci. Abs., 380.9.

Corbit, L. H., & Balleine, B. W. (2003). The role of prelimbic cortex in instrumental conditioning. *Behavioural Brain Research*, *146*, 145–157.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience. computational and mathematical modelling of neural systems*. Cambridge, MA: MIT Press.

Destexhe, A., Mainen, Z. F., & Sejnowski, T. J. (1996). Kinetic models of synaptic transmission. In C. Koch, & I. Segev (Eds.), *Methods in neuronal modeling*. Cambridge: MIT Press.

de Wit, S., Barker, R. A., Dickinson, A. D., & Cools, R. (2011). Habitual versus goal-directed action control in Parkinson disease. *Journal of Cognitive Neuroscience*, *23*(5), 1218–1229.

Dickinson, A., & Balleine, B. W. (2000). Causal cognition and goal-directed action. In C. Heyes, & L. Huber (Eds.), *The evolution of cognition* (pp. 185–204). Cambridge: MIT Press.

Donahoe, J. W., Burgos, J. E., & Palmer, D. C. (1993). A selectionist approach to reinforcement. *Journal of the Experimental Analysis of Behavior*, *60*(1), 17–40.

Eagle, D. M., Baunez, C., Hutcheson, D. M., Lehmann, O., Shah, A. P., & Robbins, T. W. (2008). Stop-signal reaction-time task performance: role of prefrontal cortex and subthalamic nucleus. *Cerebral Cortex*, *18*(1), 178–188.

Faure, A., Haberland, U., Conde, F., & El Massioui, N. (2005). Lesion to the nigrostriatal dopamine system disrupts stimulus–response habit formation. *Journal of Neuroscience*, *25*, 2771–2780.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science*, *308*(5722), 662–667. http://dx.doi.org/10.1126/science.1106138.

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943. http://dx.doi.org/10.1126/science.1102941. (New York, N.Y.).

Gerfen, C. R., Staines, W. A., Arbuthnott, G. W., & Fibiger, H. C. (1982). Crossed connections of the substantia nigra in the rat. *The Journal of Comparative Neurology*, *207*, 283–303.

Gonzalez-Burgos, G., Krimer, L. S., Urban, N. N., Barrionuevo, G., & Lewis, D. A. (2004). Synaptic efficacy during repetitive activation of excitatory inputs in primate dorsolateral prefrontal cortex. *Cerebral Cortex*, *14*, 530–542.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. a new functional anatomy. *Biological Cybernetics*, *85*, 401–410.

Hernández-López, S., Bargas, J., Surmeier, D. J., Reyes, A., & Galarraga, E. (1997). D1 receptor activation enhances evoked discharge in neostriatal medium spiny neurons by modulating an L-type Ca^{2+} conductance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(9), 3334–3342.

Hommel, B. (2003). Acquisition and control of voluntary action. In S. Maasen, W. Prinz, & G. Roth (Eds.), *Voluntary action: brains, minds, and sociality* (pp. 34–48). Oxford: Oxford University Press.

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

Ishibashi, H., Hihara, S., & Iriki, A. (2000). Acquisition and development of monkey tool-use: behavioral and kinematic analyses. *Canadian Journal of Physiology and Pharmacology*, *78*(11), 958–966. http://dx.doi.org/10.1139/cjpp-78-11-958.

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

Jahr, C. E., & Stevens, C. F. (1990). Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics. *The Journal of Neuroscience*, *10*, 3178–3182.

Johnston, K., Levin, H. M., Koval, M. J., & Everling, S. (2007). Top-down control-signal dynamics in anterior cingulate and prefrontal cortex neurons following task switching. *Neuron*, *53*(3), 453–462. http://dx.doi.org/10.1016/j.neuron.2006.12.023.

Kepecs, A., van Rossum, M. C. W., Song, S., & Tegner, J. (2002). Spike-timing-dependent plasticity: common themes and divergent vistas. *Biological Cybernetics*, *87*, 446–458.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*, 400–408.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, *67*(1), 145–163.

Luebke, J., & von der Malsburg, C. (2004). Rapid processing and unsupervised learning in a model of the cortical macrocolumn. *Neural Computation*, *16*, 501–533.

McHaffie, J. G., Stanford, T. R., Stein, B. E., Coizet, V., & Redgrave, P. (2005). Subcortical loops through the basal ganglia. *Trends in Neurosciences*, *28*, 401–407.

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Research. Brain Research Reviews*, *31*, 236–250.

Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*(4), 381–425.

Miyachi, S., Hikosaka, O., & Lu, X. F. (2002). Differential activation of monkey striatal neurons in the early and late stages of procedural learning. *Experimental Brain Research*, *146*, 122–126.

Mountcastle, V. (1997). The columnar organization of the neocortex. *Brain*, *120*, 701–722.

Redgrave, p., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., et al. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nature Reviews. Neuroscience*, *11*, 760–772.

Schroll, H., Vitay, J., & Hamker, F. H. (2012). Working memory and response selection: a computational account of interactions among cortico-basalganglio-thalamic loops. *Neural Networks*, *26*, 59–74. http://dx.doi.org/10.1016/j.neunet.2011.10.008.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. (New York, N.Y.).

Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning*, *22*, 123–158.

Song, A., Miller, K., & Abbott, L. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, *3*, 919–926.

Sridharan, D., Prashanth, P. S., & Chakravarthy, V. S. (2006). The role of the basal ganglia in exploration in a neural model based on reinforcement learning. *International Journal of Neural Systems*, *16*(2), 111–124.

Suri, R. E., & Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, *121*, 350–354.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9–44.

van Albada, S. J., & Robinson, P. A. (2009). Mean-field modeling of the basal ganglia-thalamocortical system. I firing rates in healthy and parkinsonian states. *Journal of Theoretical Biology*, *257*(4), 642–663. http://dx.doi.org/10.1016/j.jtbi.2008.12.018.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, *393*, 440–442.

Wilson, C. J., & Kawaguchi, Y. (1996). The origins of two-state spontaneous membrane potential fluctuations of neostriatal spiny neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *16*(7), 2397–2410.

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

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews. Neuroscience*, *7*(6), 464–476.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *The European Journal of Neuroscience*, *19*(1), 181–189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *The European Journal of Neuroscience*, *22*(2), 505–512. http://dx.doi.org/10.1111/j.1460-9568.2005.04219.x.

Zucker, R. S., & Regehr, W. G. (2002). Short-term synaptic plasticity. *Annual Review of Physiology*, *64*, 355–405.