# Dual Competition between the Basal Ganglia and the Cortex: from Action-Outcome to Stimulus-Response

**Meropi Topalidou,**[1,2,3,4,†] **Daisuke Kase,**[2,3,5,†] **Thomas Boraud,**[2,3,5,6,‡] **and Nicolas P. Rougier**[1,2,3,4,‡,*]
[1]INRIA Bordeaux Sud-Ouest 33405 Talence, France [2]Institut des Maladies Neurodégénératives, Université de Bordeaux, 33000, Bordeaux, France [3]Institut des Maladies Neurodégénératives, CNRS, UMR 5293, 33000, Bordeaux, France [4]LaBRI, Université de Bordeaux, Institut Polytechnique de Bordeaux, CNRS, UMR 5800, 33405 Talence, France [5]CNRS, French-Israeli Neuroscience Lab, 33000 Bordeaux, France [6]CHU de Bordeaux, Institut MN Clinique, 33000 Bordeaux, France [†]These authors have contributed equally to this work. [‡]These authors have also contributed equally to this work.
[*]Corresponding author: Nicolas.Rougier@inria.fr

**Action-outcome (A-O) and stimulus-response (S-R) processes that are two forms of instrumental conditioning that are important components of decision making and action selection. The former adapts its response according to the outcome while the latter is insensitive to the outcome. An unsolved question is how these two processes emerge, cooperate and interact inside the brain in order to issue a unique behavioral answer. Here we propose a model of the interaction between the cortex, the basal ganglia and the thalamus based on a dual competition. We hypothesize that the striatum, the subthalamic nucleus, the internal pallidum (GPi), the thalamus, and the cortex are involved in closed feedback loops through the hyperdirect and direct pathways. These loops support a competition process that results in the ability for the basal ganglia to make a cognitive decision followed by a motor decision. Considering lateral cortical interactions (short range excitation, long range inhibition), another competition takes place inside the cortex allowing this latter to make a cognitive and a motor decision. We show how this dual competition endows the model with two regimes. One is oriented towards action-outcome and is driven by reinforcement learning, the other is oriented towards stimulus-response and is driven by Hebbian learning. The final decision is made according to a combination of these two mechanisms with a gradual transfer from the former to the latter. We confirmed these theoretical results on primates using a two-armed bandit task and a reversible bilateral inactivation of the internal part of the globus pallidus.**

**Keywords:** Cortex, Basal Ganglia, Competition, Short-range Excitation, Long-range Inhibition, Segregated Loops, Direct Pathway, Hyperdirect Pathway, Reinforcement Learning, Hebbian Learning, Covert Learning, Transfer Learning, Stimulus-Response, Action-Outcome

## Introduction

Action-outcome (A-O) and stimulus-response (S-R) processes that are two forms of instrumental conditioning and important components of behavior. The former evaluates the benefit of an action in order to choose the best action among those available (action selection) while the latter is responsible for automatic behavior (routines), eliciting a response as soon as a known stimulus is present (Mishkin, Malamut, & Bachevalier, 1984; Graybiel, 2008), independently of the hedonic value of the stimulus. Action selection can be easily characterized using a simple operant conditioning setup such as for example, a two-armed bandit task where an animal must choose between two options of different value, the value being probability, magnitude or quality of reward (Pasquereau et al., 2007; Guthrie, Leblois, Garenne, & Boraud, 2013). After some trials and errors, a wide variety of vertebrates are able to select the best option (Herrnstein, 1974; Graft, Lea, & Whitworth, 1977; Matthews & Temple, 1979; Bradshaw, Szabadi, Bevan, & Ruddle, 1979; Dougan, McSweeney, & Farmer, 1985; Herrnstein, Vaughan, Mumford, & Kosslyn, 1989; Lau & Glimcher, 2005, 2008; Gilbert-Norton, Shahan, & Shivik, 2009). This selection is believed to result from the behavioral expression of the action-selection system. If the associated values are to be changed after only a few trials, the animal can still adapt its behavior and select rapidly the new best option. However, after intensive training (that depends on the species and the task) and if the same values are used all along, the animal will tend to become insensitive to change and persist in selecting the formerly best option (Lau & Glimcher, 2005; Yin & Knowlton, 2006).

Most of the studies on action selection and habits/routines agree on a slow and incremental transfer from the action-outcome to the stimulus-response system such that after extensive training, the S-R system takes control of behavior and the animal becomes insensitive to reward devaluation (Packard & Knowlton, 2002; Seger & Spiering, 2011). But very little is known on the exact mechanism underlying such transfer and one difficult question that immediately arises is when and how the brain switches from a flexible action selection system to a more static one. Our working hypothesis is that there is no need for such an explicit switch. We propose instead that an action expressed in the motor area results from both the continuous cooperation (acquisition) and competition (expression) of the two systems.

To do so, we consider the now classical actor-critic model of decision making elaborated in the 1980s that posits there are two separate components in order to explicitly represent the policy independently of the value function. The actor is in charge of choosing an action in a given state (policy) while the critic is in charge of evaluating (criticizing) the current state (value function). This classical view has been used extensively for modelling the basal ganglia (Suri, R E & Schultz, W, 1999; Suri, 2002; Frank, 2004; Doya, 2007; Glimcher, 2011; Doll, Bradley B, Simon, Dylan A, & Daw, Nathaniel D, 2012) even though the precise anatomical mapping of these two components is still subject to debate and may diverge from one model to the other (Redgrave, Peter, Gurney, Kevin, & Reynolds, John, 2008; Niv, Yael & Langdon, Angela, 2016). However, all these models share the implicit assumption that the actor and the critic are acting in concert, i.e. the actor determines the policy exclusively from the values estimated by the critic, as in Q-Learning or SARSA. Interestingly enough, (Sutton, R S & Barto, A G, 1998) noted in their seminal work that one could imagine intermediate architectures in which both an action-value function and an independent policy would be learned. We support this latter hypothesis based

on a decision-making model that is grounded on anatomical and physiological data and that identify the cortex-basal ganglia (CBG) loop as the actor. The critic — of which the Substantia Nigra pars compacta (SNc) and the Ventral Tegmental Area (VTA) are essential components — interacts through dopamine projections to the striatum (Leblois, Boraud, Meissner, Bergman, & Hansel, 2006). Decision is generated by symmetry breaking mechanism that emerges from competitions processes between positives and negatives feedback loop encompassing the full CBG network (Guthrie et al., 2013). This model captured faithfully behavioural, electrophysiological and pharmacological data we obtained in primates using implicit variant of two-armed bandit tasks — that assessed both learning and decision making — but was less consistent with the explicit version (i.e. when values are known from the beginning of the task) that focus on the decision process only.

We therefore modified this early model by adding a cortical module that has been granted with a competition mechanism and Hebbian learning (Doya, 2000). This improved version of the model predicts that the whole CBG loop is actually necessary for the implicit version of the task, however, when the basal ganglia feedback to cortex is disconnected, the system is still able to choose in the explicit version of the task. Our experimental data fully confirmed this prediction (Piron et al., 2016) and allowed to solve an old conundrum concerning the pathophysiology of the BG which was that lesion or jamming of the output of the BG improve Parkinson patient motor symptoms while it affects marginally their cognitive and psycho-motor performances. An interesting prediction of this generalized actor-critic architecture is that the valuation of options and the behavioural outcome are segregated. In the computational model, it implies that if we block the output of the basal ganglia in a two-armed bandit task before learning, and because reinforcement learning occurred at the striatal level under dopaminergic control, this should induce covert learning when the model chooses randomly. The goal of this study is thus twofold: i) to present a comprehensive description of the model in order to provide the framework for an experimental paradigm that allow to unravel covert learning and ii) to test this prediction in monkeys.

# Materials and Methods

## The task

We consider a variant of a n-armed bandit task (Katehakis & Veinott, 1987; Auer, Cesa-Bianchi, Freund, & Schapire, 2002) where a player must decide which arm of $n$ slot machines to play in a finite sequence of trials such as to maximize his accumulated reward. This task has received much attention in the literature (e.g. machine learning, psychology, biology, game theory, economics, neuroscience, etc.) because it provides a simple model to explore the trade-off between exploration (trying out a new arm to collect information about its payoff) and exploitation (playing the arm with the highest expected payoff) (Robbins, 1952; Gittins, 1979). This task has been shown to be solvable for a large number of different living beings, with (Plowright & Shettleworth, 1990; Keasar, 2002; Steyvers, Lee, & Wagenmakers, 2009) or without a brain (Reid et al., 2016), and even a clever physical apparatus can solve the task (Naruse et al., 2015).

### The computational task

In the present study, we restrict the n-armed bandit task to $n = 2$ with an explicit dissociation between the choice of the option (*cognitive* choice) and the actual triggering of the option (*motor* choice). This introduces a supplementary difficulty because only the motor choice – the physical (and visible) expression of the choice – will be taken into account when computing the reward. If cognitive and motor choices are incongruent, only the motor choices matters. Unless specified otherwise, we consider a set of

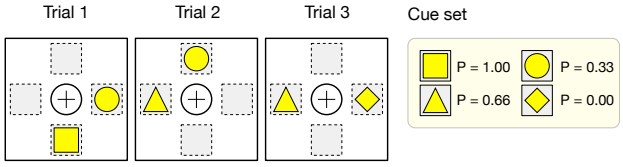


**Figure 1.** Three task trials from a 4-items cue set (□, △, ○, ◇) with respective reward probabilities (1, 2/3, 1/3, 0).

cues $\{C_i\}_{i\in[1,n]}$ associated with reward probabilities $\{P_i\}_{i\in[1,n]}$ and a set of four different locations ($\{L_i\}_{i\in[1,4]}$) corresponding to the *up, down, left, right* positions on the screen. A trial is made of the presentation of two random cues $C_i$ and $C_j$ ($i \neq j$) at two random locations ($L_i$ and $L_j$) such that we have $L_i \neq L_j$ (see Fig. 1). A session is made of $n$ successive trials and can use one to several different cue sets depending on the condition studied (e.g. reversal, devaluation). Unless specified otherwise, in the present study, exactly one cue set is used throughout a whole session.

Once a legal motor decision has been made, reward is computed by drawing a random uniform number between 0 and 1. If the number is less or equal to the reward probability of the chosen cue, a reward of 1 is given, else, a reward of 0 is given. If no motor choice has been made or if the motor choice leads to an empty location (illegal choice), the trial is considered to be failed and no reward is given, which is different from giving a reward of 0. Best choice for a trial is defined as the choice of the cue associated with the highest reward probability among the two presented cues. Performance is defined as the ratio of best choices over the total number of trials. A perfect player with full-knowledge can achieve a performance of 1 while the mean expectation of reward is directly dependent on the cue sampling policy[1].

### The behavioral task

*With kind permission from the authors (Piron et al., 2016), we reproduce here the details of the experimental task which is similar.*

The primates were trained daily in the experimental room and familiarized with the setup, which consisted of 4 buttons placed on a board at different locations (0°, 90°, 180°, and 270°) and a further button in a central position, which detects contact with a monkey's hand. These buttons correspond to the 4 possible display positions of a cursor on a vertical screen. The monkeys were seated in chairs in front of this screen at a distance of 50cm (Fig. 2). The monkeys initiated a trial by keeping their hands on the

[1] For example on Fig. 1, if we consider a uniform cue sampling policy for 6*n trials, the mean expected reward for a perfect player with full knowledge is $3/6 \times 1 + 2/6 \times 2/3 + 1/6 \times 1/3 = 14/18 \approx 0.777\ldots$

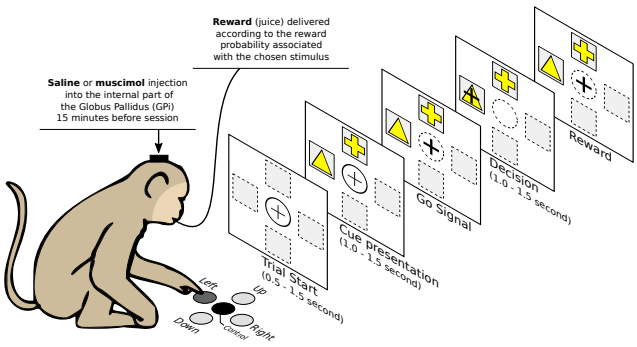


**Figure 2.** The behavioral task. The monkeys initiate a trial by keeping their hands on the central button, which induced the appearance of the cursor in the central position of the screen. After a random delay, two cues appears in 2 different positions. The monkey has a random duration time window (0.5s to 1.5s) to press the button associated with one cue. It moves the cursor over the chosen cue and it has to maintain the position for some duration. After this delay, the monkey is rewarded (0.3 ml of water) or not according to the reward probability of the chosen cue.

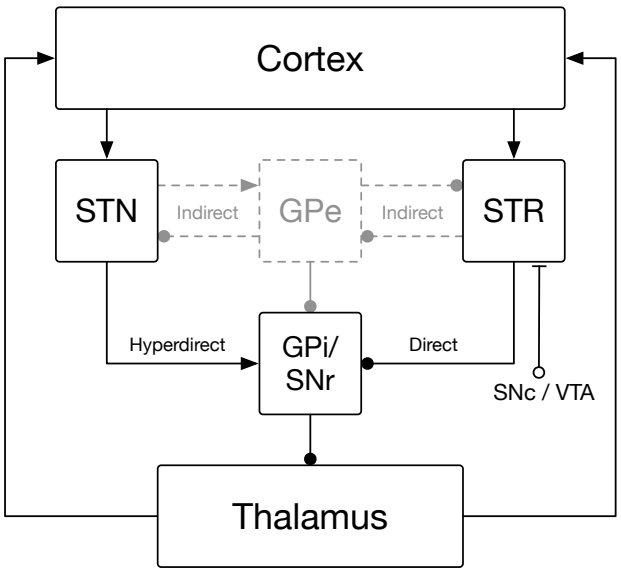


**Figure 3.** Architecture of the model. The architecture of the model is centered around the hyperdirect pathway (cortex → subthalamic nucleus → GPi/SNr → thalamus → cortex), the direct pathway (cortex → striatum → GPi/SNr → thalamus → cortex) and the cortex where lateral interactions take place (not represented on the figure). The external part of the globus pallidus, while not present in the model, is represented on the figure as a reminder of the actual connectivity in the BG. Similarly, the substantia nigra pars compacta is not explicitly represented in the model.

central button, which induced the appearance of the cursor in the central position of the screen. After a random delay (0.5s to 1.5s), 2 cues appeared in 2 (of 4) different positions determined randomly for each trial. Each cue had a fixed probability of reward ($P_1$ =0.75 and $P_2$ = 0.25) and remains the same same during a session. Once the cues were shown, the monkeys had a random duration time window (0.5s to 1.5s) to press the button associated with one cue. It moves the cursor over the chosen cue and they have to maintain the position for 0.5 s to 1.5 s. After this delay, the monkeys were rewarded (0.3 ml of water) or not according to the reward probability of the chosen target. An end-of-trial signal corresponding to the disappearance of the cursor was given, indicating to the monkeys that the trial was finished and they could start a new trial after an inter-trial interval between 0.5 s and 1.5s.

## The model

The model is designed to study the implications of a dual competition between the cortex and the basal ganglia (BG). The competition inside the cortex is conveyed through direct lateral interactions (short-range excitation and long range inhibition, (H. R. Wilson & Cowan, 1972, 1973; Coultrip, Granger, & Lynch, 1992; Muir & Cook, 2014; Deco et al., 2014)) while the competition within the BG is conveyed through the direct and hyperdirect pathways (Leblois et al., 2006; Guthrie et al., 2013). Therefore, the indirect pathway and the external segment of the globus pallidus (GPe) are not included.

### Architecture

Our model contains five main groups (see Fig. 3). Three of these groups are excitatory. These are the cortex (CTX), the thalamus (THL), and the subthalamic nucleus (STN). Two populations are inhibitory. They correspond to the sensory–motor territories of the striatum (STR) and the GPi. The model has been further tailored into three segregated loops (Alexander, DeLong, & Strick, 1986; Alexander & Crutcher, 1990; Alexander, Crutcher, & DeLong, 1991; Mink, 1996; Haber, 2003), namely the motor loop, the associative loop and the cognitive (or limbic) loop. The motor loop comprises the motor cortex (supplementary motor area (SMA), primary cortex (M1), premotor cortex (PMC), cingulate motor area (CMA)), the motor striatum (putamen), the motor STN, the motor GPi (motor territory of the pallidum and the substantia nigra) and the motor thalamus (ventrolateral thalamus (VLm and VLo)). The associative loop comprises the cognitive cortex (dorsolateral prefrontal cortex (DLPFC), the lateral orbitofrontal cortex (LOFC)) and the associative striatum (associative territory of the caudate). The cognitive loop comprises the cognitive cortex (anterior cingulate area (ACA), medial orbitofrontal cortex (MOFC)), the cognitive striatum (ventral caudate), the cognitive STN, the cognitive GPi (limbic territory of the pallidum and the substantia nigra and) the cognitive thalamus (ventral anterior thalamus (VApc, VAmc)).

### Populations

The model comprises 12 populations: 5 motor populations, 4 cognitive populations and 2 associative populations (Fig. 4). These populations comprises from 4 to 16 neural assemblies and possess each a specific geometry whose goal is to facilitate connectivity description. Each assembly is modeled using a neuronal rate model (Hopfield, 1984; Shriki, Hansel, & Sompolinsky, 2003) that give account of the spatial mean firing rate of the neurons composing the assembly. Each assembly is governed by the following equations:

$$\tau \frac{dV}{dt} = -V + I_{syn} + I_{ext} + h \quad (1)$$

$$U = f(V + V.n)) \quad (2)$$

where $\tau$ is the assembly time constant (decay of the synaptic input), $V$ is the firing rate of the assembly, $I_{syn}$ is the synaptic input to the assembly, $I_{ext}$ is the external input representing the sensory visual salience of the cue, $h$ is the threshold of the assembly, $f$ is the transfer function and $n$ is the (correlated, white)

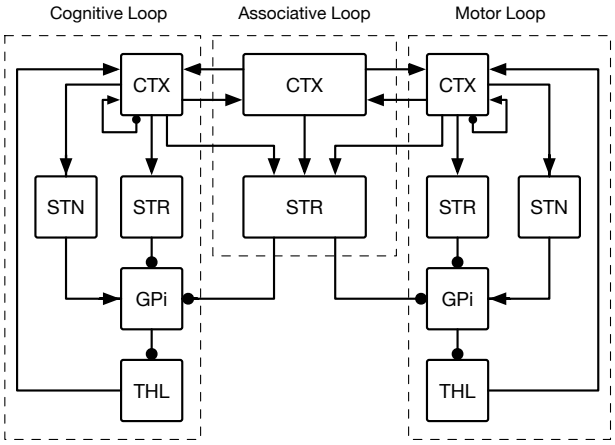


**Figure 4.** Segregated loops. The model is further detailed into three segregated circuits (cognitive, associative, motor). The cognitive and motor circuit each comprises a cortical, a striatal, a thalamic, a subthalamic, and a pallidal population while the associative loop only comprises a cortical and a striatal population. This latter interacts with the two other circuits via diffused connections to the pallidal regions and from all cortical populations. Arrows, excitatory connections. Dots, inhibitory connections.

| Population | | Geometry | $\tau$ | Threshold | Noise |
|---|---|---|---|---|---|
| Cortex | associative | (4,4) | 10ms | -3 | 1.0% |
| | cognitive | (4,1) | 10ms | -3 | 1.0% |
| | motor | (1,4) | 10ms | -3 | 1.0% |
| Striatum | associative | (4,4) | 10ms | 0 | 0.1% |
| | cognitive | (4,1) | 10ms | 0 | 0.1% |
| | motor | (4,1) | 10ms | 0 | 0.1% |
| GPi | cognitive | (4,1) | 10ms | -10 | 3.0% |
| | motor | (1,4) | 10ms | -10 | 3.0% |
| STN | cognitive | (4,1) | 10ms | -10 | 0.1% |
| | motor | (1,4) | 10ms | -10 | 0.1% |
| Thalamus | cognitive | (4,1) | 10ms | -40 | 0.1% |
| | motor | (1,4) | 10ms | -40 | 0.1% |

**Table 1.** Population parameters

| Name | Value |
|---|---|
| $V_{min}$ | 1 |
| $V_{max}$ | 20 |
| $V_h$ | 16 |
| $V_c$ | 3 |

**Table 2.** Parameters for striatal sigmoid transfer function

noise term. Each population possess its own set of parameters according to the group it belongs to (see Table 1). Transfer function for all population but the striatal population is a ramp function ($f(x) = max(x, 0)$). The striatal population that is silent at rest (Sandstrom & Rebec, 2002), requires concerted coordinated input to cause firing (C. J. Wilson & Groves, 1981), and has a sigmoidal transfer function (nonlinear relationship between input current and membrane potential) due to both inward and outward potassium current rectification (Nisenbaum & Wilson, 1995). This is modeled by applying a sigmoidal transfer function to the activation of cortico-costriatal inputs in the form of the Boltzmann equation:

$$f(x) = V_{min} + \frac{V_{max} - V_{min}}{1 + e^{\frac{V_h - x}{V_c}}}$$

where $V_{min}$ is the minimum activation, $V_{max}$ the maximum activation, $V_h$ the half- activation, and $V_c$ the slope. This is similar to the use of the output threshold in the (Gurney, Prescott, & Redgrave, 2001) model and results in small or no activation to weak inputs with a rapid rise in activation to a plateau level for stronger inputs. The parameters used for this transfer function are shown in Table 2 and were selected to give a low striatal output with no cortical activation (1 spike/s), starting to rise with a cortical input of 10 sp/s and a striatal output of 20 spike/s at a cortical activation of 30 spike/s.

## Connectivity

Even though the model takes advantage of segregated loops, they cannot be entirely separated if we want the cognitive and the motor channel to interact. This is the reason why we incorporated a divergence in the corticostriatal connection followed by a re-convergence within the GPi (Graybiel, Aosaki, Flaherty, & Kimura, 1994; Parent et al., 2000) (see Fig. 5). Furthermore, we considered the somatotopic projection of the pyramidal cortical neurons to the striatum (Webster, 1961) as well as their arborization(Cowan & Wilson, 1994; C. J. Wilson, 1987; Parent et al., 2000; Parthasarathy, Schall, & Graybiel, 1992) resulting in specific localized areas of button formation (Kincaid, Zheng, & Wilson, 1998) and small cortical areas innervating the striatum in a discontinuous pattern with areas of denser innervation separated by areas of sparse innervation (Brown, Smith, & Goldbloom, 1998; Flaherty & Graybiel, 1991). We also cinsidered the large reduction in the number of neurons from cortex to striatum to GPi (Bar-Gad & Bergman, 2001; Oorschot, 1996). These findings combined lead to striatal areas that are mostly specific for input from one cortical area alongside areas where there is overlap between inputs from two or more cortical areas (Takada et al., 2001) and which are here referred to as the associative striatum.

The gain of the synaptic connection from population $A$ (presynaptic) to population $B$ (postsynaptic) is denoted as $G_{A \to B}$, and the total synaptic input to population B is:

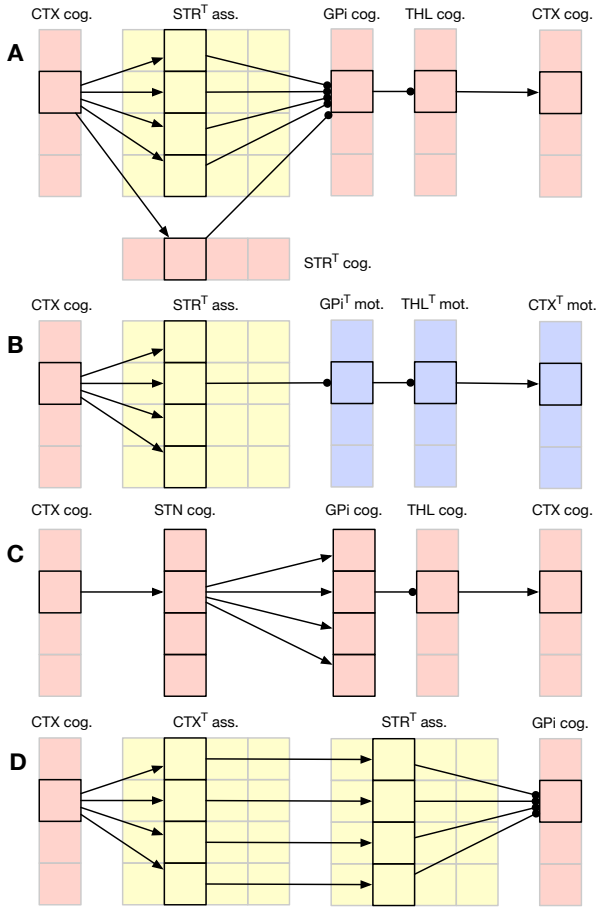


**Figure 5.** Partial connectivity in the cognitive and associative loops. For clarity, only one assembly has been considered. The motor loop is symmetric to the cognitive one. The "T" symbol on some name means the geometry of the group has been transposed (for readability). **A** The direct pathway from cognitive cortical assemblies diverge from cortex to associative and cognitive striatum. the pathway converges into cognitive GPi, send parallel projection to the thalamus and forms a closed loop with the original cognitive cortical assembly. **B** Thanks to the convergence of motor and cognitive pathways in association striatum, there is a cross-talking between the motor and cognitive loops. This allow a decision made in the cognitive loop to influence the decision in motor loops and vice-versa. **C** The hyperdirect pathway from cognitive cortical assembly diverges from STN to GPi, innervating all cognitive, but not motor, GPi regions and feeds back to all cognitive cortical assemblies. **D** The pathway from associative cortex and associative striatum is made of parallel localized projections.

| Pop. A | Pop. B | Pathway | Pattern | Gain |
|---|---|---|---|---|
| Cortex | Striatum | cog. → cog. • | (i,1) → (i,1) | 1.0 |
| | | mot. → mot. | (i,1) → (i,1) | 1.0 |
| | | ass. → ass. | (i,j) → (i,j) | 1.0 |
| | | cog. → ass. | (i,1) → (i,*) | 0.2 |
| | | mot. → ass. | (1,i) → (*,i) | 0.2 |
| | STN | cog. → cog. | (i,1) → (i,1) | 1.0 |
| | | mot. → mot. | (1,i) → (1,i) | 1.0 |
| | Thalamus | cog. → cog. | (i,1) → (i,1) | 0.1 |
| | | mot. → mot. | (1,i) → (1,i) | 0.1 |
| | Cortex | cog. → cog. | (i,1) → (*,1) | ±0.5 |
| | | mot. → mot. | (1,i) → (1,*) | ±0.5 |
| | | ass. → ass. | (i,j) → (*,*) | ±0.5 |
| | | ass. → mot. | (*,i) → (1,i) | 0.025 |
| | | ass. → cog. | (i,*) → (i,1) | 0.01 |
| | | cog. → ass. • | (i,1) → (i,*) | 0.025 |
| | | mot. → ass. | (1,i) → (*,i) | 0.01 |
| Striatum | GPi | cog. → cog. | (i,1) → (i,1) | -2.0 |
| | | mot. → mot. | (1,i) → (1,i) | -2.0 |
| | | ass. → cog. | (i,*) → (i,1) | -2.0 |
| | | ass. → mot. | (*,i) → (1,i) | -2.0 |
| STN | GPi | cog. → cog. | (i,1) → (i,1) | 1.0 |
| | | mot. → mot. | (1,i) → (1,i) | 1.0 |
| GPi | Thalamus | cog. → cog. | (i,1) → (i,1) | -1.0 |
| | | mot. → mot. | (1,i) → (1,i) | -1.0 |
| Thalamus | Cortex | cog. → cog. | (i,1) → (i,1) | 1.0 |
| | | mot. → mot. | (1,i) → (1,i) | 1.0 |

**Table 3.** Connectivity gains and pattern between the different populations. For connectivity patterns, "*" means all. For example, (1,i) → (1,*) means one-to-all connectivity while (1,i) → (1,i) means one-to-one connectivity. Plastic pathways are indicated by a "•" symbol.

$$I^B_{syn} = G_{A\to B} \sum_A U_A$$

where $A$ is the presynaptic assembly, $B$ is the postsynaptic assembly, and $U_A$ is the output of presynaptic assembly $A$. The gains for each pathway are shown in Table 3. Gains to the corresponding cognitive (motor) assembly are initially five times higher than to each receiving associative area. Re-convergence from cognitive (motor) and association areas of striatum to cognitive (motor) areas of GPi are evenly weighted.

## Task encoding

At the trial start, assemblies in the cognitive cortex encoding the two cues $C_1$ and $C_2$ receive an external current (7Hz) and assemblies in the motor cortex encoding the two positions $M_1$ and $M_2$ receive similarly an external current (7Hz). These activities

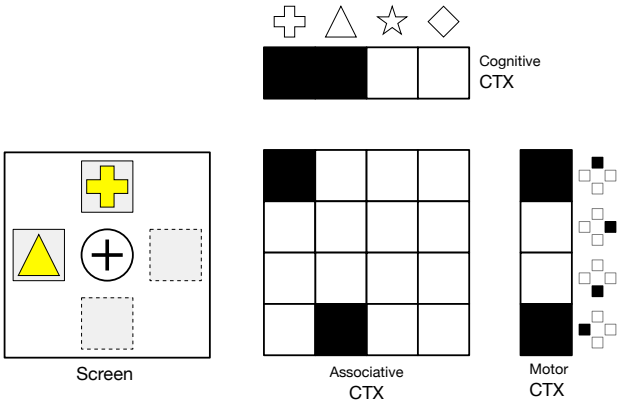


**Figure 6.** Task encoding. Assemblies in the cognitive cortex encoding the two cues $C_1$ and $C_2$ receive an external current and assemblies in the motor cortex encoding the two positions $M_1$ and $M_2$ receive similarly an external current. These activities are not sufficient to disambiguate between the situation ($C_1/M_1$, $C_2/M_2$) and the situation ($C_1/M_2$, $C_2/M_1$). This is the reason why the associative cortex encoding one of these two situations receives an external current, ($C_1/M_1$, $C_2/M_2$) in the present case.

are not sufficient to disambiguate between the situation ($C_1/M_1$, $C_2/M_2$) and the situation ($C_1/M_2$, $C_2/M_1$). This is the reason why the associative cortex encoding one of these two situations receives an external current (7Hz), ($C_1/M_1$, $C_2/M_2$) in the present case (see Fig. 6. The decision of the model is decoded from the activity in the motor cortex *only*, i.e. independently of the activity in the cognitive cortex. If the model chooses a given cue but produces the wrong motor command, the cognitive choice will not be taken into account and the final choice will be decoded from the motor command that may lead to an irrelevant choice.

## Dynamic

There exist two different competition mechanisms inside the model. One is conveyed through the direct and hyperdirect pathways, the other is conveyed inside the cortex through short-range excitation and long range inhibition. The former has been fully described and analyzed in Leblois et al., 2006 while the latter been extensively studied in a number of experimental and theoretical papers (von der Malsburg, 1973; H. R. Wilson & Cowan, 1972, 1973; Amari, 1977; Callaway, 1998; Taylor, 1999). Each of these two competition mechanisms can lead to a decision as illustrated on Fig. 7 that shows the dynamic of the motor loop for all the population in three conditions. In the absence of the cortical interactions (gain of cortical lateral connections has been set to 0), the direct and hyperdirect pathway are able to promote a competition that result in the selection of one of the two assemblies in each group. In the absence of GPi output (connection has been cut), the cortical lateral connections are able to support a competition that result in the selection of o the two assemblies, even though such decision is generally slower than the basal one. The result of the dual competition is a faster selection of one of the two assemblies prior to learning, when there is no possibility for the two competition to be non congruent (one competition tends to select move A while the others tend to select move B). We'll see in the results section that if the result of the two competitions is non-congruent, the decision is slower.

## Learning

Learning has been restricted to the cognitive channel on the cortico-striatal synapse (between the cortex cognitive and the striatum cognitive) and the cortico-cortical synapse (between the cortex cognitive and the cortex associative). There is most probably learning in other structures and pathways, but the aim here is to show that the proposed restriction is sufficient to produce the behavior under consideration. All synaptic weights are initialized to 0.5 (SD 0.005) that are used as as a multiplier to the pathway gain to keep the factors of gain and weight separately observable. All weights are bound between $W_{min}$ and $W_{max}$ (see Table 4) such that for any change $\Delta W(t)$, weight $W(t)$ is updated according to the equation:

$$W(t) \leftarrow W(t) + \Delta W(t)(W_{max} - W(t))(W(t) - W_{min})$$

**Reinforcement learning** At the level of corticostriatal synapses, phasic changes in dopamine concentration have been shown to be necessary for the production of long-term potentiation (LTP) (Kerr & Wickens, 2001; Reynolds, Hyland, & Wickens, 2001; Surmeier, Ding, Day, Wang, & Shen, 2007; Pawlak & Kerr,

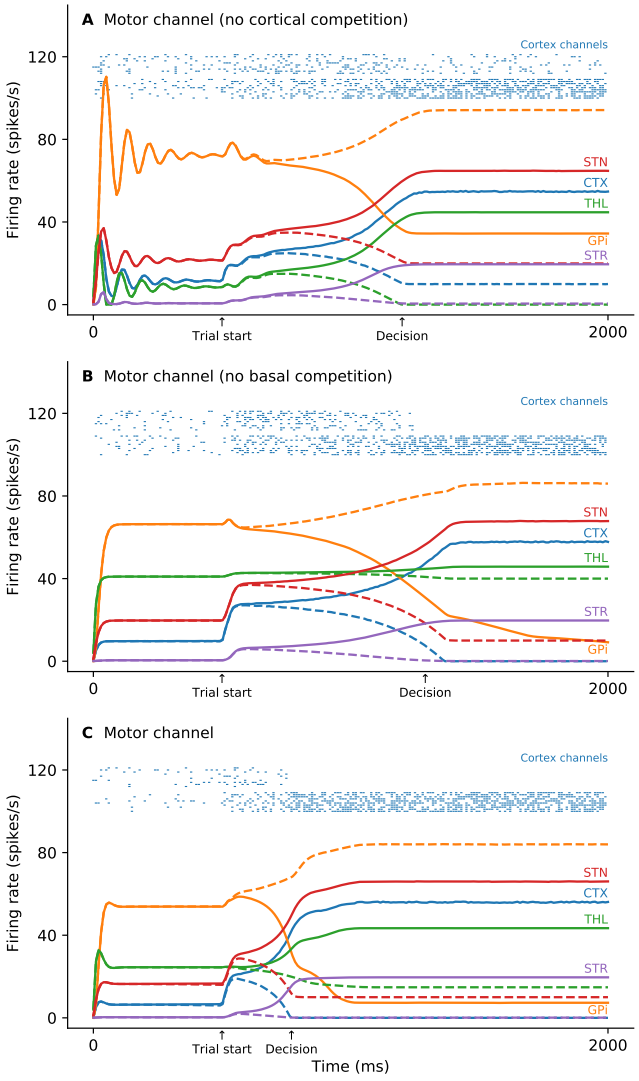


**Figure 7.** Activity in the different populations during a single trial of action selection before learning. The model is started at time t=0ms and allowed to settle to a steady state before the presentation of the cues at t=500ms. Solid lines represents activity related to the selected population, dashed lines represent activity related to the non selected population. Decision threshold has been set to 40 spikes/s between the two cortical populations and is indicated on the x axis. Raster plots are related to the cortical populations and has been generated from the firing rate of 10 neurons. **A** Activity in the motor populations in the absence of lateral competition in the cortical populations. The damped oscillations during the settling phase are characteristic of the delayed feedback from the subthalamic nucleus (excitation) and the striatum (inhibitory) through the globus pallidus and the thalamus. **B** Activity in the motor populations in the absence of the feedback from the basal ganglia (GPi) to the cortical populations via the thalamus. Decision threshold is reached thanks to the direct lateral competition in both cognitive and motor cortical channels. There is no damped oscillation since there is no delay between the cortical populations and the decision times are slower than in the previous case. **C** Activity in the motor populations in the full model with a dual competition, one cortical, one basal. When congruent (cortical and basal decision are the same), decision time for both the motor and cortical channels are faster than in the absence of one of the competition loop.

| Name | Value |
|---|---|
| $W_{min}$ | 0.25 |
| $W_{max}$ | 0.75 |
| $LTP_{RL}$ | 0.050 |
| $LTD_{RL}$ | 0.030 |
| $LTP_{HL}$ | 0.005 |
| $\alpha$ | 0.025 |

**Table 4.** Learning parameters

2008). After each trial, once reward has been received (0 or 1), the corticostriatal weights are updated according to:

$$\Delta W_B^A = LTP_{RL} \times RPE \times U_B \text{ if RPE} > 0 \quad (3)$$
$$= LTD_{RL} \times RPE \times U_B \text{ if RPE} < 0 \quad (4)$$

where $\Delta W_B^A$ is the change in the weight of the corticostriatal synapse from cortical assembly A to striatal assembly B, $RPE$ is the reward prediction error, the amount by which the actual reward delivered differs from the expected reward, $U_B$ is the activation of the striatal assembly, and $\mu$ is the actor learning rate. Generation of LTP and long-term depression (LTD) in striatal MSNs has been found to be asymmetric (Pawlak & Kerr, 2008). Therefore, in the model, the actor learning rate is different for LTP and LTD. The RPE is calculated using a simple critic learning algorithm:

$$RPE = R - V_i$$

where R, the reward, is 0 or 1, depending on whether a reward was given or not on that trial. Whether a reward was given was based on the reward probability of the selected cue (which is most of the time the one associated with the direction chosen). i is the number of the cue chosen, and $V_i$ is the value of cue i. The value of the chosen cue is then updated using the PE:

$$V_i \leftarrow V_i + \alpha PE$$

**Hebbian learning** At the level of cortico-cortical synapse, only the co-activation of two assemblies is necessary for the production of long-term potentiation (Bear & Malenka, 1994; Caporale & Dan, 2008; Feldman, 2009; Hiratani & Fukai, 2016). After each trial, once a move has been initiated, the cortico-cortical weights are updated according to:

$$\Delta W_B^A = LTP_{HL} \times U_A \times U_B$$

where $\Delta W_B^A$ is the change in the weight of the cortico-cortical synapse from cognitive cortical assembly A to associative cortical assembly B. This learning rule is thus independent of reward.

## Experimental setup

*With kind permission from the authors (Piron et al., 2016), we reproduce here the details of the experimental setup as well as the surgical procedure since the two same monkeys were used for these new experiments.*

Experimental data were obtained from 2 female macaque monkeys (Macaca mulata). Experiments were performed during the daytime. Monkeys were living under a 12h/12h diurnal rhythm. Although food access was available ad libitum, the primates were kept under water restriction to increase their motivation to work. A veterinary skilled in healthcare and maintenance in nonhuman primates supervised all aspects of animal care. Experimental procedures were performed in accordance with the Council Directive of 20 October 2010 (2010/63/ UE) of the European Community. This project was approved by the French Ethic Comity for Animal Experimentation (50120111-A).

### Surgical Procedure

Cannula guides were implanted into the left and right GPi in both animals under general anesthesia. Implantation was performed inside a stereotaxic frame guided by ventriculography and single-unit electrophysiological recordings. A ventriculographic cannula was introduced into the anterior horn of the lateral ventricle and a contrast medium was injected. Corrections in the position of the GPi were performed according to the line between the anterior commissure (AC) and the posterior commissure (PC) line. The theoretical target was AP: 23.0mm, L: 7.0 mm, P: 21.2 mm.27 A linear 16-channel multielectrode array was lowered vertically into the brain. Extracellular single-unit activity was recorded from 0mm to 24 mm relative to the AC–PC line with a wireless recording system. Penetration of the electrode array into the GPi was characterized by an increase in the background activity with the appearance of active neurons with a tonic firing rate (around the AC–PC line). The exit of the electrode tips from the GPi was characterized by the absence of spike (around 3-4 mm below the AC–PC line). When a clear GPi signal from at least 3 contacts had been obtained, control radiography of the position of the recording electrode was performed and compared to the expected position of the target according to the ventriculography. If the deviation from the expected target was less than 1mm, the electrode was removed and a cannula guide was inserted with a spare cannula inside so that the tip of the cannula was superimposed on the location of the electrode array in the control radiography. Once the cannula guide was satisfactorily placed, it was fixed to the skull with dental cement.

### Bilateral Inactivation of the GPi

Micro-injections were delivered bilaterally 15 minutes before a session. For both animals injections of the $GABA_A$ agonist muscimol hydrobromide (Sigma) or saline (NaCl 9‰) were randomly assigned each day. Muscimol was delivered at a concentration of 1$\mu$g/$\mu$l (dissolved in a NaCl vehicle). Injections (1$\mu$l in each side) were performed at a constant flow rate of 0.2 $\mu$l/min using a micro-injection system. Injections were made through a 30-gauge cannulae inserted into the 2 guide cannulae targeting left and right GPi. Cannulas were connected to a 25 $\mu$l Hamilton syringe by polyethylene cannula tubing.

## Data Analysis

Theoretical and experimental data were analyzed using Kruskal-Wallis rank sum test between the three conditions (saline (C0), muscimol (C1) or saline following muscimol (C2)) for the 6 samples (12×10 first trials of C0 (control), 12×10 last trials of C0 (control), 12×10 first trials of C1 (GPi Off/muscimol); 12×10 last trials of C1(GPi OFF/muscimol); 12×10 first trails of C2(GPi On/saline); 12×10 last trials of C2(GPi On/saline)) with

posthoc pairwise comparisons using Dunn's-test for multiple comparisons of independent samples. P-values have been adjusted according to the false discovery rate (FDR) procedure of Benjamini-Hochberg. Results were obtained from raw data using the PMCMR R package (Pohlert, 2014). Significance level was set at $P < 0.01$.

Experimental raw data is available from (Kase & Boraud, 2017) under a CC0 license, Theoretical raw data and code are available from (Rougier & Topalidou, 2017) under a CC0 license (data) and BSD license (code).

# Results

Our model predicts that the valuation of options and the behavioural outcome are two separate (but entangled) processes. This means that if we block the output of the basal ganglia before learning, reinforcement learning still occurs at the striatal level under dopaminergic control and this should induce covert learning of stimuli value even though the behavioral choice would appear as random.

## Protocol

The protocol has been consequently split over two consecutive conditions (C1 & C2) using the same set of stimuli and a dissociated control (C0) using a different set of stimuli (using same probabilities as for C1 & C2).

C0 60 trials, GPi On (model), saline injection (primates), stimulus set 1 ($A_1$, $B_1$) with $P_{A_1} = 0.75$, $P_{B_1} = 0.25$

C1 60 trials, GPi Off (model), muscimol injection (primates), stimulus set 2 ($A_2$, $B_2$) with $P_{A_2} = 0.75$, $P_{B_2} = 0.25$

C2 60 trials, GPi On (model), saline injection (primates), stimulus set 2 ($A_2$, $B_2$) with $P_{A_2} = 0.75$, $P_{B_2} = 0.25$

## Computational results

| H0 | statistic (H) | p value |
|---|---|---|
| C0 start = C2 start | 2.965 | 0.0051 |
| C1 start = C2 start | 4.986 | 1.8e-6 |
| C1 end = C2 start | 3.099 | 0.0036 |

**Table 5.** Theoretical results statistical analysis. Kruskal-Wallis rank sum test between the three conditions (saline (C0), muscimol (C1) or saline following muscimol (C2)) with posthoc pairwise comparisons using Dunn's-test for multiple comparisons of independent samples.

We tested our hypothesis on the model using 12 different sessions (corresponding to 12 different initializations of the model). On day 1, we suppressed the GPi output by cutting the connections between the GPi and the thalamus. When the GPi output has been suppressed on day 1, the performance is random at the beginning as shown by the average probability of choosing the best option (expressed in mean±SD) in the first 10 trials (0.408 ±0.161) and remain so until the end of the session (0.525 ±0.164). Statistical analysis revealed no significant difference between the 10 first and the 10 last trials. On day 2, we re-established connections between the GPi and the thalamus and the model has to perform the exact same task as for day 1 using the same set of stimuli. Results shows a significant change in behavior: the model starts with an above-chance performance on the first 10 trials (0.717 ±0.241) and this change is significant (see Table 5 and Fig. 8) as compared to the start of C1, as compared to the end of C1 and as compared to the start of C0, confirming our hypothesis that the BG have previously learned the value of stimuli even though they were unable to alter behavior.

## Experimental results

| H0 | statistic (H) | p value |
|---|---|---|
| C0 start = C2 start | 3.181 | 0.0024 |
| C1 start = C2 start | 3.738 | 0.0004 |
| C1 end = C2 start | 2.803 | 0.0069 |

**Table 6.** Experimental results statistical analysis. Kruskal-Wallis rank sum test between the three conditions (saline (C0), muscimol (C1) or saline following muscimol (C2)) with posthoc pairwise comparisons using Dunn's-test for multiple comparisons of independent samples.

We tested the prediction on two female macaque monkeys which have been implanted with two cannula guides into the left and right GPi (see Materials and Methods section for details). In order to inhibit the GPi, we injected bilaterally a GABA agonist (muscimol, 1$\mu$g) 15 minutes before working session on day 1 (C1). The two monkeys were trained for 7 and 5 sessions respectively, each session using the same set of stimuli. Results on day 1 shows that animals were unable to choose the best stimulus in such condition from the start (0.433 ±0.236) to the end (0.492 ±0.250) of the session. Statistical analysis revealed no significant difference between the 10 first and the 10 last trials on day 1. On day 2 (C2), we injected bilaterally a saline solution 15 minutes before working session and animals had to perform the exact same protocol as for day 1. Results shows a significant change in behavior (see Table 6 and Fig. 8): animals start with an above-chance performance on the first 10 trials (P=0.667 ±0.213), as compared to the start of C1, as compared to the end of C1 and as compared to the start of C0, confirming our hypothesis that the BG has previously learned the value of stimuli.

# Discussion

## Covert learning in the BG

These results reinforce the classical idea that the basal ganglia architecture is based on an actor critic architecture where the dopamine serves as a reinforcement signal. However, the proposed model goes beyond this classical hypothesis and proposes a more general view on the role of the BG in behaviour and the entanglement with the cortex. Our results, both theoretical and experimental, suggest that the critic part of the BG extends its role beyond the basal ganglia and makes it *de facto* a central component in behavior that can evaluate any action, independently of their origin. This hypothesis is very congruent with the results introduced in Charlesworth, Warren, and Brainard (2012) where authors show that the anterior forebrain pathway in Bengalese finches contributes to skill learning even when it is blocked and

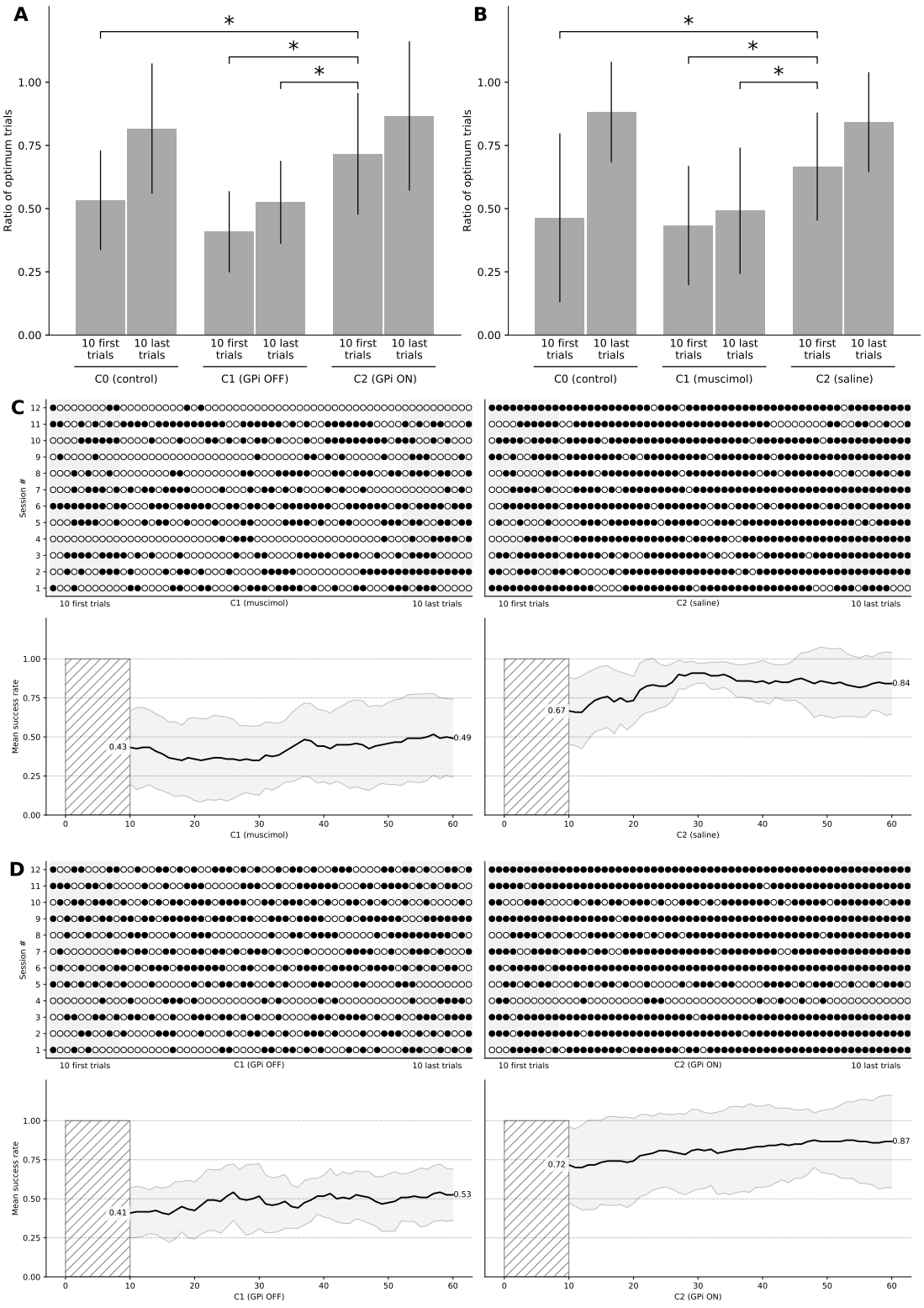


**Figure 8.** Theoretical and experimental results. Histograms show the mean performance at the start and the end of a session in day 1 and day 2 conditions for both the model (**A**) and the monkeys (**B**). At the start of day 2, the performance for both the model and the monkeys is significantly higher compared to the start and end of day 1, suggesting some covert learning occurred during day 1 even though performances are random during day 1. **C** Individual trials (n=2x60) for all the sessions (n=12) for the primates. **D** Individual trials (n=2x60) for all the sessions (n=12) for the model. A black dot means a successful trial (the best stimulus has been chosen), an outlined white dot means a failed trial (the best stimulus has not been chosen). Measure of success is independent of the actual reward received after having chosen one of the two stimuli. The bottom part of each panel shows the mean success rate over a sliding window of ten consecutive trials and averaged across all the sessions. The thick black line is the actual mean and the gray-shaded area represents the standard deviation (STD) over sessions.

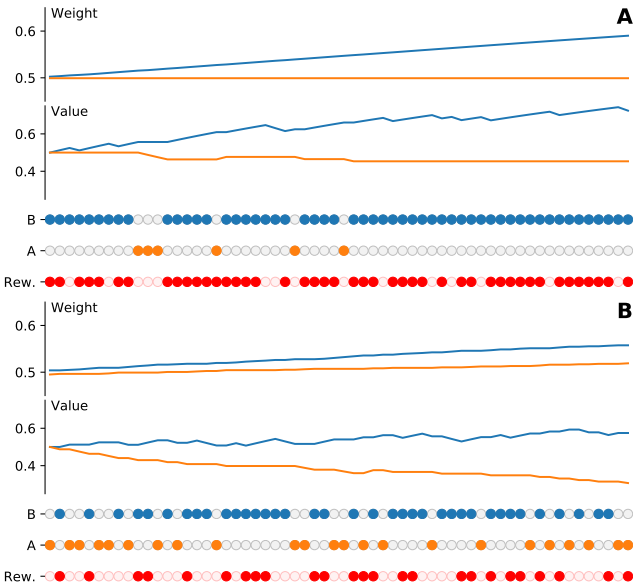


**Figure 9.** Model performance during a single session. Filled dots indicate the chosen cue between A and B. Filled red dots indicate if a reward has been received following the choice. Reward probability is 0.75 for cue A and 0.25 for cue B. **A** Intact model (C0). The BG output drives the decision and evaluates the value of cue A and cue B with a strong bias in favor of A because this cue is chosen more frequently. In the meantime, the Hebbian weight relative to this cue is strongly increased while the weight relative to the other cue doesn't change significantly. **B** Lesioned model (C1). The BG output has been suppressed and decisions are random. Hebbian weights for cue A and cue B are both increased up to similar values at the end of the session. In the meantime, the value of cue A and cue B are evaluated within the BG and the random sampling of cue A and cue B leads to an actual better sampling of value A and B. This is clearly indicated by the estimated value of B that is very close the theoretical value (0.25).

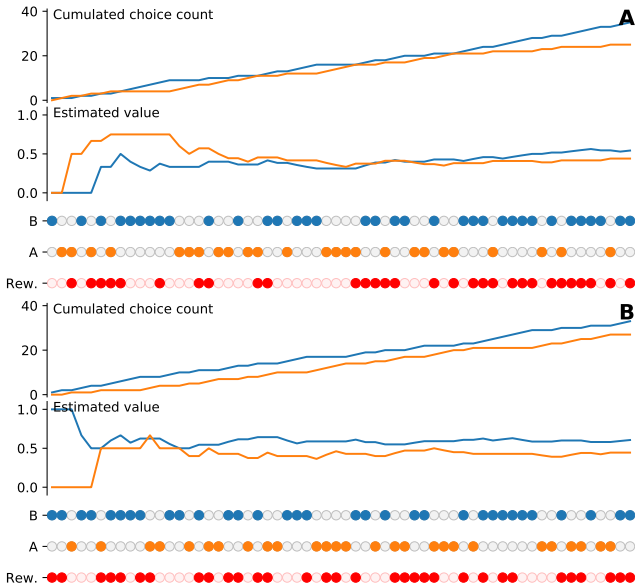


**Figure 10.** Monkey performance during a single session. Filled dots indicate the chosen cue between A and B. Filled red dots indicate if a reward has been received following the choice. Reward probability is 0.75 for cue A and 0.25 for cue B. **A** In saline condition (C0), the monkey is able to slowly choose for the best cue with a slight preferences for A at the end of the 60 trials. Estimation of the perceived value of the two cues shows the actual value of A is greater than the value of B at the end of the session **B** In muscimol condition (C1), the monkey choose cues randomly as indicated by the overall count of choices A and B. Estimation of the perceived value of the two cues (dashed lines) reveals a greater estimation of the value of A compared to the value of B.

does not participate in the behavioural performance. This is also quite compatible with (Ashby, Turner, & Horvitz, 2010; Hélie, Ell, & Ashby, 2015) who propose that the BG is a general purpose trainer for cortico-cortical connections. Here, we introduced a precise computational model using both reinforcement and Hebbian learning, supported by experimental data, that explains precisely how this general purpose trainer can be biologically implemented.

This can be simply understood by scrutinizing a session in control and lesion condition (see Fig. 9). In control condition, the model learns to select the best cue thanks to the BG. Because it learns what is the best stimulus, this induces a preferential selection of the best stimulus in order to obtain a higher probability of reward. If the process is repeated over many trials, this leads implicitly to an over-representation of the more valuable stimuli at the cortical level and since cortex learns with Hebbian learning, it is implicitly learned. Said differently, the value of the best stimulus has been *converted* to the temporal domain. In lesion condition, the selection is random and each stimulus is roughly selected with equal probability and this allows the BG to evaluate the value of the two stimuli even more precisely. We believe this is the same for the monkeys even though we do not have access to internal value and weights. However, we can see on Fig. 10 that the estimated value of stimuli (computed as the probability of reward) reflects the highest value for the best stimulus. Similarly, the number of time a given stimulus has been selected is correlated with its actual value even if it is not significant.

## From action-outcome to stimulus-response

These new results, together with our previous results (Piron et al., 2016) shed a new light on a plausible neural mechanism responsible for the gradual mix between an action-outcome behavior and a stimulus-response one. The novelty in our hypothesis is that there is no transfer *per se*. There is instead a joint combination of the two systems that act and learn together and we tend to disagree with the hypothesis of a hierarchical system (Dezfouli & Balleine, 2013). In our case, the final behavioral decision results from a subtle balance between the two decisions. When a new task needs to be solved, the basal ganglia initially drives the decision because it has initially a faster dynamic. In the meantime, the cortex takes advantage of this driving and gradually learns the decision independently of the reward. We've shown how this could be the case for monkeys, even though we lack experimental evidence that the decision in muscimol condition is actually driven by the cortex. The actual combination of the two systems might be more complex than a simple weighted linear combination and this make the study even more difficult to carry on. What we see at the experimental level might the projection of a more complex phenomenon. Persisting in a devaluated task does not mean the system is *frozen* but the time to come back from a stimulus-response oriented behavior might be simply much longer than the time to initially acquire the behavior.

Finally, our results suggest a behavioral decision results from both the cooperation (acquisition) and competition (expression) of two distinct but entangled systems.

# References


Alexander, G. E. & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in neurosciences*. doi:10.1016/0166-2236(90)90107-L

Alexander, G. E., Crutcher, M. D., & DeLong, M. R. (1991). Chapter 6 basal ganglia-thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. In *Progress in brain research* (pp. 119–146). Elsevier. doi:10.1016/s0079-6123(08)62678-3

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*(1), 357–381. doi:10.1146/annurev.ne.09.030186.002041

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87. doi:10.1007/bf00337259

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. doi:10.1016/j.tics.2010.02.001

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, *32*(1), 48–77. doi:10.1137/S0097539701398375

Bar-Gad, I. & Bergman, H. (2001). Stepping out of the box: Information processing in the neural networks of the basal ganglia. *Current Opinion in Neurobiology*, *11*(6), 689–695. doi:10.1016/s0959-4388(01)00270-7

Bear, M. F. & Malenka, R. C. (1994). Synaptic plasticity: LTP and LTD. *Current Opinion in Neurobiology*, *4*(3), 389–399. doi:10.1016/0959-4388(94)90101-5

Bradshaw, C. M., Szabadi, E., Bevan, P., & Ruddle, H. V. (1979). The effect of signaled reinforcement availability on concurrent performances in humans. *Journal of the Experimental Analysis of Behavior*, *32*(1), 65–74. doi:10.1901/jeab.1979.32-65

Brown, L. L., Smith, D. M., & Goldbloom, L. M. (1998). Organizing principles of cortical integration in the rat neostriatum: Corticostriate map of the body surface is an ordered lattice of curved laminae and radial points. *The Journal of Comparative Neurology*, *392*(4), 468–488. doi:10.1002/(SICI)1096-9861(19980323)392:4<468::AID-CNE5>3.0.CO;2-Z

Callaway, E. M. (1998). Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience*, *21*(1), 47–74. doi:10.1146/annurev.neuro.21.1.47

Caporale, N. & Dan, Y. (2008). Spike timing–dependent plasticity: A hebbian learning rule. *Annual Review of Neuroscience*, *31*(1), 25–46. doi:10.1146/annurev.neuro.31.060407.125639

Charlesworth, J. D., Warren, T. L., & Brainard, M. S. (2012). Covert skill learning in a cortical-basal ganglia circuit. *Nature*, *486*(7402), 251–255.

Coultrip, R., Granger, R., & Lynch, G. (1992). A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks*, *5*(1), 47–54. doi:10.1016/s0893-6080(05)80006-1

Cowan, R. L. & Wilson, C. J. (1994). Spontaneous firing patterns and axonal projections of single corticostriatal neurons in the rat medial agranular cortex. *Journal of Neurophysiology*, *71*(1), 17–32.

Deco, G., Ponce-Alvarez, A., Hagmann, P., Romani, G. L., Mantini, D., & Corbetta, M. (2014). How local excitation-inhibition ratio impacts the whole brain dynamics. *Journal of Neuroscience*, *34*(23), 7886–7898. doi:10.1523/jneurosci.5068-13.2014

Dezfouli, A. & Balleine, B. W. (2013). Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computational Biology*, *9*(12), e1003364. doi:10.1371/journal.pcbi.1003364

Doll, Bradley B, Simon, Dylan A, & Daw, Nathaniel D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081.

Dougan, J. D., McSweeney, F. K., & Farmer, V. A. (1985). Some parameters of behavioral contrast and allocation of interim behavior in rats. *Journal of the Experimental Analysis of Behavior*, *44*(3), 325–335. doi:10.1901/jeab.1985.44-325

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*(6), 732–739.

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *HFSP Journal*, *1*(1), 30–11.

Feldman, D. E. (2009). Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience*, *32*(1), 33–55. doi:10.1146/annurev.neuro.051508.135516

Flaherty, A. W. & Graybiel, A. M. (1991). Corticostriatal transformations in the primate somatosensory system. projections from physiologically mapped body-part representations. *Journal of Neurophysiology*, *66*(4), 1249–1263.

Frank, M. J. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943. doi:10.1126/science.1102941

Gilbert-Norton, L. B., Shahan, T. A., & Shivik, J. A. (2009). Coyotes (canis latrans) and the matching law. *Behavioural Processes*, *82*(2), 178–183. doi:10.1016/j.beproc.2009.06.005

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 148–177.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108 Suppl 3*(Supplement 3), 15647–15654.

Graft, D. A., Lea, S. E. G., & Whitworth, T. L. (1977). The matching law in and within groups of rats1. *Journal of the Experimental Analysis of Behavior*, *27*(1), 183–194. doi:10.1901/jeab.1977.27-183

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, *31*(1), 359–387. doi:10.1146/annurev.neuro.29.051605.112851

Graybiel, A. M., Aosaki, T., Flaherty, A. W., & Kimura, M. (1994). The basal ganglia and adaptive motor control. *Science*, *265*(5180), 1826–1831. doi:10.1126/science.8091209

Gurney, K., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. II. analysis and simulation of behaviour. *Biological Cybernetics*, *84*(6), 411–423. doi:10.1007/pl00007985

Guthrie, M., Leblois, A., Garenne, A., & Boraud, T. (2013). Interaction between cognitive and motor cortico-basal ganglia loops during decision making: A computational study. *Journal of Neurophysiology*, *109*, 3025–3040. doi:10.1152/jn.00026.2013

Haber, S. N. (2003). The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy*, *26*(4), 317–330. doi:10.1016/j.jchemneu.2003.10.003

Hélie, S., Ell, S. W., & Ashby, F. G. (2015). Learning robust cortico-cortical associations with the basal ganglia: An integrative review. *Cortex*, *64*, 123–135. doi:10.1016/j.cortex.2014.10.011

Herrnstein, R. J. (1974). Formal properties of the matching law1. *Journal of the Experimental Analysis of Behavior*, *21*(1), 159–164. doi:10.1901/jeab.1974.21-159

Herrnstein, R. J., Vaughan, W., Mumford, D. B., & Kosslyn, S. M. (1989). Teaching pigeons an abstract relational rule: Insideness. *Perception & Psychophysics*, *46*(1), 56–64. doi:10.3758/bf03208074

Hiratani, N. & Fukai, T. (2016). Hebbian wiring plasticity generates efficient network structures for robust inference with synaptic weight plasticity. *Frontiers in Neural Circuits*, *10*. doi:10.3389/fncir.2016.00041

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*(10), 3088–3092.

Kase, D. & Boraud, T. (2017). Covert learning in the basal ganglia: Raw data. FigShare. doi:10.6084/m9.figshare.4753507.v1

Katehakis, M. N. & Veinott, A. F. (1987). The multi-armed bandit problem: Decomposition and computation. *Mathematics of Operations Research*, *12*(2), 262–268. doi:10.1287/moor.12.2.262

Keasar, T. (2002). Bees in two-armed bandit situations: Foraging choices and possible decision mechanisms. *Behavioral Ecology*, *13*(6), 757–765. doi:10.1093/beheco/13.6.757

Kerr, J. N. D. & Wickens, J. R. (2001). Dopamine d-1/d-5 receptor activation is required for long-term potentiation in the rat neostriatum in vitro. *Journal of Neurophysiology*, *85*(1), 117–124.

Kincaid, A. E., Zheng, T., & Wilson, C. J. (1998). Connectivity and convergence of single corticostriatal axons. *Journal of Neuroscience*, *18*(12), 4722–4731.

Lau, B. & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579. doi:10.1901/jeab.2005.110-04

Lau, B. & Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451–463. doi:10.1016/j.neuron.2008.02.021

Leblois, A., Boraud, T., Meissner, W., Bergman, H., & Hansel, D. (2006). Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *Journal of Neurosciences*, *26*, 3567–3583. doi:10.1523/JNEUROSCI.5050-05.2006

Matthews, L. R. & Temple, W. (1979). Concurrent schedule assessment of food preference in cows. *Journal of the Experimental Analysis of Behavior*, *32*(2), 245–254. doi:10.1901/jeab.1979.32-245

Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*(4), 381–425. doi:10.1016/S0301-0082(96)00042-1

Mishkin, M., Malamut, B., & Bachevalier, J. (1984). Memories and habits: Two neural systems. In G. Lynch, J. L. McGaugh, & N. M. Weinberger (Eds.), *Neurobiology of human learning and memory*.

Muir, D. R. & Cook, M. (2014). Anatomical constraints on lateral competition in columnar cortical architectures. *Neural Computation*, *26*(8), 1624–1666. doi:10.1162/neco_a_00613

Naruse, M., Berthel, M., Drezet, A., Huant, S., Aono, M., Hori, H., & Kim, S.-J. (2015). Single-photon decision maker. *Scientific Reports*, *5*(1). doi:10.1038/srep13253

Nisenbaum, E. S. & Wilson, C. J. (1995). Potassium currents responsible for inward and outward rectification in rat neostriatal spiny projection neurons. *Journal of Neuroscience*, *15*(6), 4449–4463.

Niv, Yael & Langdon, Angela. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, *11*, 67–73.

Oorschot, D. E. (1996). Total number of neurons in the neostriatal, pallidal, subthalamic, and substantia nigral nuclei of the rat basal ganglia: A stereological study using the cavalieri and optical disector methods. *The Journal of Comparative Neurology*, *366*(4), 580–599. doi:10.1002/(SICI)1096-9861(19960318)366:4<580::AID-CNE3>3.0.CO;2-0

Packard, M. G. & Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, *25*(1), 563–593. doi:10.1146/annurev.neuro.25.112701.142937

Parent, A., Sato, F., Wu, Y., Gauthier, J., Lévesque, M., & Parent, M. (2000). Organization of the basal ganglia: The importance of axonal collateralization. *Trends in Neurosciences*, *23*, S20–S27. doi:10.1016/s1471-1931(00)00022-7

Parthasarathy, H., Schall, J., & Graybiel, A. M. (1992). Distributed but convergent ordering of corticostriatal projections: Analysis of the frontal eye field and the supplementary eye field in the macaque monkey. *Journal of Neuroscience*, *12*(11), 4468–4488.

Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., Gross, C. E., & Boraud, T. (2007). Shaping of motor responses by incentive values through the basal ganglia. *Journal of Neuroscience*, *27*(5), 1176–1183. doi:10.1523/jneurosci.3745-06.2007

Pawlak, V. & Kerr, J. N. D. (2008). Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *Journal of Neuroscience*, *28*(10), 2435–2446. doi:10.1523/jneurosci.4402-07.2008

Piron, C., Kase, D., Topalidou, M., Goillandeau, M., Orignac, H., N'Guyen, T.-H., Rougier, N. P., & Boraud, T. (2016). The globus pallidus pars interna in goal-oriented and routine behaviors: Resolving a long-standing paradox. *Movement Disorders*, *31*(8), 1146–1154. doi:10.1002/mds.26542

Plowright, C. & Shettleworth, S. J. (1990). The role of shifting in choice behavior of pigeons on a two-armed bandit. *Behavioural Processes*, *21*(2-3), 157–178. doi:10.1016/0376-6357(90)90022-8

Pohlert, T. (2014). *The pairwise multiple comparison of mean ranks package (pmcmr)*. R Package. Retrieved from http://CRAN.R-project.org/package=PMCMR

Redgrave, Peter, Gurney, Kevin, & Reynolds, John. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews*, *58*(2), 322–339.

Reid, C. R., MacDonald, H., Mann, R. P., Marshall, J. A. R., Latty, T., & Garnier, S. (2016). Decision-making without a brain: How an amoeboid organism solves the two-armed bandit. *Journal of The Royal Society Interface*, *13*(119), 20160030. doi:10.1098/rsif.2016.0030

Reynolds, J. N. J., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(6851), 67–70. doi:10.1038/35092560

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–536. doi:10.1090/s0002-9904-1952-09620-8

Rougier, N. P. & Topalidou, M. (2017). Covert learning in the basal ganglia: Code. doi:10.5281/zenodo.598112

Sandstrom, M. I. & Rebec, G. V. (2002). Characterization of striatal activity in conscious rats: Contribution of NMDA and AMPA/kainate receptors to both spontaneous and glutamate-driven firing. *Synapse*, *47*(2), 91–100. doi:10.1002/syn.10142

Seger, C. A. & Spiering, B. J. (2011). A critical review of habit learning and the basal ganglia. *Frontiers in Systems Neuroscience*, *5*. doi:10.3389/fnsys.2011.00066

Shriki, O., Hansel, D., & Sompolinsky, H. (2003). Rate models for conductance-based cortical neuronal networks. *Neural Computation*, *15*(8), 1809–1841. doi:10.1162/08997660360675053

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179. doi:10.1016/j.jmp.2008.11.002

Suri, R E & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91*(3), 871–890.

Suri, R. E. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Networks*, *15*(4-6), 523–533.

Surmeier, D. J., Ding, J., Day, M., Wang, Z., & Shen, W. (2007). D1 and d2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences*, *30*(5), 228–235. doi:10.1016/j.tins.2007.03.008

Sutton, R S & Barto, A G. (1998). *Reinforcement learning: An introduction*.

Takada, M., Tokuno, H., Hamada, I., Inase, M., Ito, Y., Imanishi, M., Hasegawa, N., Akazawa, T., Hatanaka, N., & Nambu, A. (2001). Organization of inputs from cingulate motor areas to basal ganglia in macaque monkey. *European Journal of Neuroscience*, *14*(10), 1633–1650. doi:10.1046/j.0953-816x.2001.01789.x

Taylor, J. G. (1999). Neural 'bubble' dynamics in two dimensions: Foundations. *Biological Cybernetics*, *80*(6), 393–409. doi:10.1007/s004220050534

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, *14*(2), 85–100. doi:10.1007/bf00288907

Webster, K. (1961). Cortico-striate interrelations in the albino rat. *Journal of anatomy*, *95*, 532–544.

Wilson, C. J. (1987). Morphology and synaptic connections of crossed corticostriatal neurons in the rat. *The Journal of Comparative Neurology*, *263*(4), 567–580. doi:10.1002/cne.902630408

Wilson, C. J. & Groves, P. M. (1981). Spontaneous firing patterns of identified spiny neurons in the rat neostriatum. *Brain Research*, *220*(1), 67–80. doi:10.1016/0006-8993(81)90211-0

Wilson, H. R. & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, *12*(1), 1–24. doi:10.1016/s0006-3495(72)86068-5

Wilson, H. R. & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, *13*(2), 55–80. doi:10.1007/bf00288786

Yin, H. H. & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464–476. doi:10.1038/nrn1919

# Abbreviations

| | |
|---|---|
| **A-O** | Action – Outcome |
| **AC** | Anterior Commissure |
| **CMA** | Cingulate motor area |
| **CTX** | Cortex |
| **DLPFC** | Dorsolateral prefrontal cortex |
| **DLS** | Dorsolateral striatum |
| **DMS** | Dorsomedial striatum |
| **FEF** | Frontal eye fields |
| **GPi** | Internal part of the globus pallidus |
| **GPe** | External part of the globus pallidus |
| **LTP** | Long-term potentiation |
| **LTD** | Long-term depression |
| **LOFC** | Lateral orbitofrontal cortex |
| **M1** | Primary motor cortex |
| **MOFC** | Medial orbitofrontal cortex |
| **OFC** | Orbitofrontal cortex |
| **PC** | Posterior commissure |
| **PFC** | Prefrontal cortex |
| **PMC** | Premotor cortex |
| **RPE** | Reward prediction error |
| **SMA** | Supplementary motor area |
| **SNc** | Substantia nigra pars compacta |
| **SNr** | Substantia nigra pars reticulata |
| **STN** | Subthalamic nucleus |
| **STR** | Striatum |
| **S-R** | Stimulus – Response |
| **THL** | Thalamus |
| **VLm** | Ventrolateral thalamus, pars medialis |
| **VLo** | Ventrolateral thalamus, pars oralis |
| **VApc** | Ventral anterior thalamus, pars parvocellularis |
| **VAmc** | Ventral anterior thalamus, pars magnocellularis |