Dissociation of reinforcement and Hebbian learning induces covert acquisition of value in the basal ganglia

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, Talence, France
[2]University of Bordeaux, UMR 5293, IMN, Bordeaux, France
[3]CNRS, UMR 5293, IMN, Bordeaux, France
[4]LaBRI, University of Bordeaux, IPB, CNRS, UMR 5800, Talence, France
[5]CNRS, French-Israeli Neuroscience Lab, Bordeaux, France
[6]CHU de Bordeaux, IMN Clinique, Bordeaux, France
[†]These authors contributed equally to this work
[‡]These authors also contributed equally to this work
[*]Corresponding author: Nicolas.Rougier@inria.fr

This paper introduces a new hypothesis concerning the formation of habits in the cortex of primates under the implicit supervision of the basal ganglia. This hypothesis has been formulated using a theoretical model and confirmed experimentally in monkeys. To do so, and prior to learning, we inactivated the internal part of the globus pallidus (GPi, the main output structure of the BG) with injections of muscimol and we tested monkeys on a variant of a two-armed bandit task where two stimuli are associated with two distinct reward probabilities (0.25 and 0.75 respectively). Unsurprisingly, their performance in such conditions are at the chance level. However, the theoretical model predicts that even if the performance is random, the value of the stimuli are implicitly evaluated and learned. This has been tested and confirmed on the next day, when inhibition has been removed: monkeys instantly showed quasi-optimal performances, demonstrating they knew the relative value of the two stimuli. Said differently, we managed to explicitly dissociate reinforcement learning from Hebbian learning and demonstrated covert learning inside the basal ganglia. These results suggest that a behavioral decision results from both the cooperation (acquisition) and competition (expression) of two distinct but entangled memory systems, the goal-directed system and the habit system that may represent the two ends of the same graded phenomenon.

Keywords: habits, action selection, cortex, basal ganglia, computational model, value acquisition, habit expression, covert learning, reinforcement learning, Hebbian learning

Introduction

When entering a new building, you're confronted to a very simple binary choice: should you pull or push the door to enter? If there is no external sign indicating what is the rightful action, you might as well select a random action or select the one you're the most used to and, in case of failure, you can immediately switch to the alternate action. If you come to this building on a regular basis, you will rapidly learn what is the right action to enter (push) and to exit (pull) the building without ever noticing it. To enter and to exit the building will be soon part of you daily routines. However, if for some reason the door is reversed (because of works), you will find yourself persisting in pushing the door to enter and pulling the door to exit. Even though you realized on the first day something has changed, this situation can last for several days until you actually change your behavior and adapt it to the new situation. But there are some situations where adapting behavior to new conditions is not so easy and overcoming habits becomes extremely difficult. For example, when you live in a right-side driving country, you're used to look first at the left side to watch for incoming cars before crossing the road. If you now travel to a left-side driving country for a short stay, it is extremely difficult to look on your right before crossing the road. The habit is so strongly imprinted in your brain and your body that it cannot be overcome in just a few days. Only a long-term stay can give you a chance (if any) to adapt to the new conditions.

Such habits have been identified, demonstrated and studied for a long time in many different fields of neuroscience and psychology [1–4] but there is still a large degree of uncertainty around their exact definition. According to the review provided by [5], habits can be characterized using five common, although not always present, features: inflexible, slow or incremental, insensitive to reinforcer devaluation, unconscious and automatic. This characterization may further vary across fields depending on the species, tasks and methodologies such that in the end, it is difficult to assess if a given behavior results from an habit or from another process. However, even if insensitivity to reinforcer devaluation is largely considered to be a hallmark of habits [6–11] (although such overtraining is hardly ever defined), things might be more complex as illustrated by the simple example we introduced previously. Habits might be indeed a more graded phenomenon.

To gain a better insight, we have to consider both 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 (habits), eliciting a response as soon as a known stimulus is present [1, 7], independently of the hedonic value of the stimulus. Habits and action selection can be easily characterized using a simple operant conditioning setup such as for example, a two-armed bandit task (see Fig 1) where an animal must choose between two options of different value, the value being probability, magnitude or quality of reward [12, 13]. After some trials and errors, a wide variety of vertebrates are able to select the best option [14–22]. 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 [20, 23]. This selection is believed to result from the behavioral expression of the habit system. Most of the studies on action selection and habits agree on a slow and incremental transfer from the action-selection system to the habit system such that after extensive training, the habit system takes control of behavior and the animal becomes insensitive to reward devaluation [5, 24]. 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 and habitual 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. We therefore upgraded our previous model of reinforcement learning and decision making through the cortex BG loop [13, 25, 26] by adding a cortical module that is granted a competition mechanism and Hebbian learning capacity.

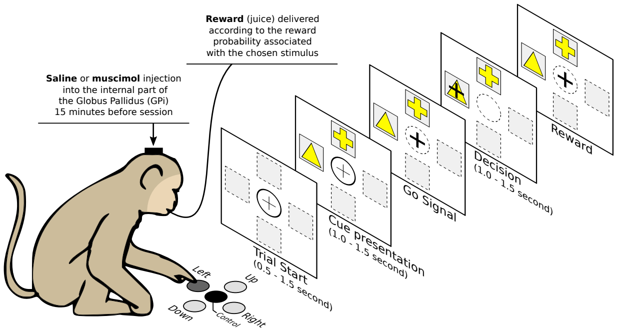


Figure 1: Description of the task. A trial is made of the simultaneous presentation of two cues at two random positions associated with a fixed reward probability. The monkey has to choose a stimulus at the go signal and maintain this choice for one second. Reward is delivered according to the reward probability associated with the chosen stimulus.

Results

We designed a simple two-armed bandit task where two stimuli A and B are associated with different reward probability (respectively 0.25 and 0.75) as shown on Fig 1 (see Material and Methods section for full description of the different protocols). The goal for the subject is to choose the stimulus associated with the highest reward probability, independently of its position. The theoretical model makes a clear distinction between the acquisition of habits — that requires the basal ganglia to imprint the cortex — and the expression of habits — that does not require the basal ganglia *per se*. However, if these two processes are normally entangled and congruent in the course of daily behavior, the model predicts that it is possible to dissociate them experimentally. The idea is to consider the learning of a novel set of stimuli (with respective reward probability 0.25 and 0.75) while the GPi output is suppressed, such that the BG cannot influence the selection through the cortical feedback. Because cortical learning in the model is Hebbian [27] and does not depend on reward, we should observe random choices from the model, leading to an equal amount of learning for any of the two cues. In the meantime, and because the BG still receive the dopaminergic signal following a choice, it is able to learn the actual value of the cue, even if, it cannot influence anymore the actual action.

We first ran a series of control experiments for both the model and the monkeys to ensure the task can be learned without difficulty as shown on Fig 2 (Control). We proceeded with 2 different protocols where the GPi is either inhibited on day 1 (protocol 2) or on day 2 (protocol 1). In the case of protocol 1, and because the control experiment has demonstrated that both the model and the monkeys were able to learn the respective value of A and B, this also induces a preferential selection of stimulus B 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 stimuli B at the cortical level. Said differently, the value of B has been converted into the temporal do-main (i.e. frequency). Considering our hypothesis that Hebbian learning occurs in the same time at the cortical level (LTP/LTD) and depends mostly on the occurrence of coincident events, the cortex can implicitly learn the value of the stimulus through such sequence of repeated trials. This means that on day 2 (D2), we should observe an above chance performance, when the GPi is inhibited. More interestingly, the model also predicts that protocol 1 can be reversed, that is, GPi can be inhibited on day 1 (D1) and later disinhibited on day 2. During the first day and after several trials in such condition, we hypothesize the value of the two stimuli to have been learned within the BG even if the BG cannot influence decision, meaning the actual decision should be random. However, when the inhibition of the GPi is removed on day 2, we should observe a radical change in behavior for subsequent trials. Because the BG has learned the actual value of the cue and can now influence behavior through thalamus and cortex (whose learning resulted in a no differentiation of the two cues), peak performance should be reached quasi instantly.

Theoretical results

Left part of Fig 2 shows the theoretical results for the first 25 trials and the last 25 trials averaged over 25 different simulations. For protocol 2, when the GPi output is suppressed on day 1 (D1), the performance is random from the start (P=0.506 ±0.12) to the end (P=0.555 ±0.15) of the session. However, after GPi inhibition has been removed on day 2 (D2, at trial 120) and ran for another batch of 120 trials, we can observe a significant change in the performance and the model immediately reaches near-optimal performance level (P=0.907 ±0.09) and improves until the end of the session (P=1.000 ±0.00). We further proceeded with the model and tested it on day 3 with a renewed suppression of the GPi output. Interestingly enough, performances do not drop to chance level but start at a high level (P=0.882 ±0.06) and improve until the end of the session (P=0.938 ±0.05) beause habits have been formed at the cortical level. These results are also reported on Fig 3 that display the instantaneous performance using a sliding window of 10 trials. The discontinuity between the end of day 1 and the start of day 2 is particularly clear. More interestingly, the measure of the internal estimation of the reward probability for each of the two stimuli show no such discontinuity as predicted by our hypothesis.

Experimental results

We tested the prediction on two female macaque monkeys which have been implanted two cannula guides into the left and right GPi (see Materials and Methods section for details). We applied the same protocol as in the theoretical model, that is, we injected bilaterally GABA agonist muscimol 15 minutes before working session (see Materials & Methods in supplementary material) on day 1. The two monkeys were trained for 6 and 7 sessions respectively, each session using the same set of stimuli. Results (right part of Fig 2) shows that animals are unable to choose the best stimulus in such condition from start (P=0.36 ±0.05) to end (P=0.41 ±0.05) of the session. On day 2, animals were injected bilaterally a saline solution 15 minutes before working session and they were trained using the exact same protocol as for day 1. Results shows a drastic and significant change in behavior: animals start with near-optimal performance on the first 25 trials (P=0.97 ±0.05), confirming our hypothesis that the BG have previously learned the value of stimuli even though they were unable to alter behavior. Based on these theoretical results and in light of experimental results in the monkey for protocol 1, we can predict that a similar habit formation would occur in the primate frontal cortex.

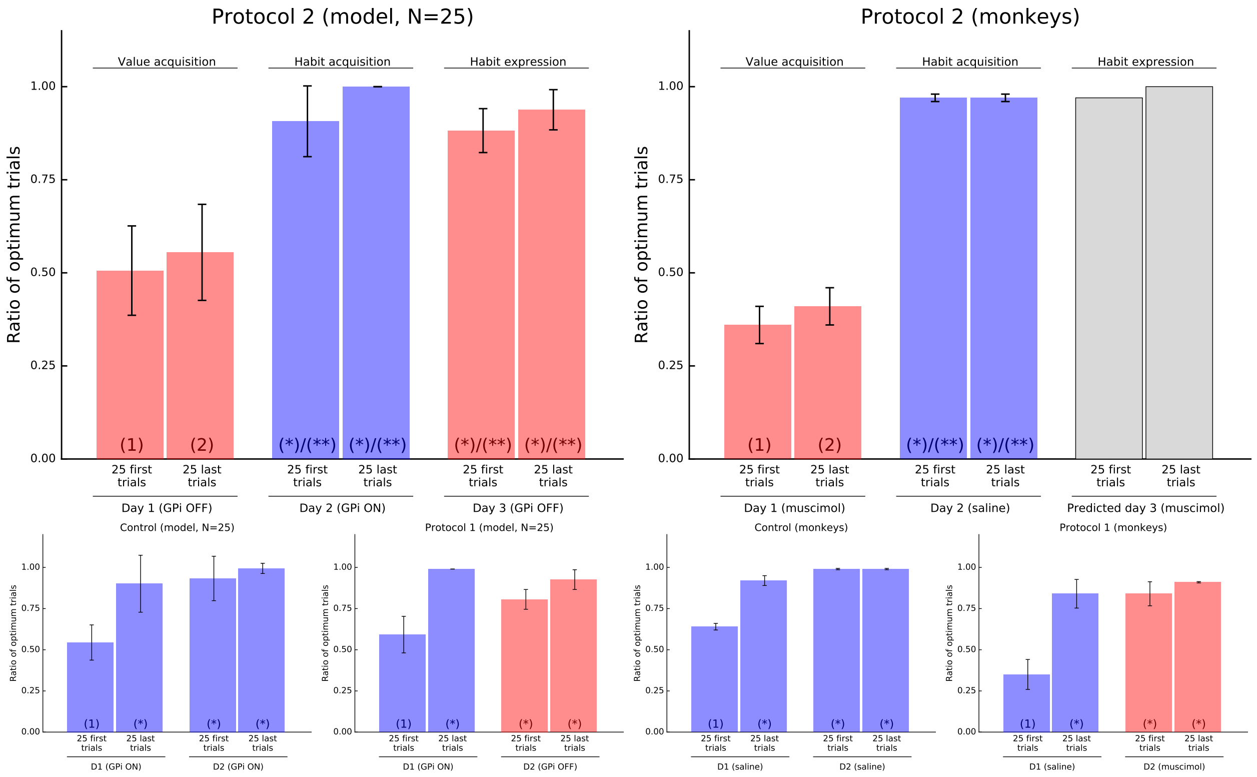


Figure 2: Comparative results. Protocol 2 D1 corresponds to the first day of the experiment where GPi output is suppressed in both the model (removal of GPi-Thalamus connection) and the monkeys (muscimol injection). D2 corresponds to the second day where the suppression of the GPi output is removed. It is remarkable to see that D1 results in random choice while in D2 performances are quasi instantly optimal and improve until the end of the session. This tends to confirm the hypothesis that the BG have learned the value of the stimuli during day 1 even if they were unable to alter behavior. During D3, GPi output is suppressed again and performances remains very high. This is due to the learning of the task at the cortical level in the model. Note that for monkeys, D3 results is only a prediction based on model results and monkey protocol 1 results. They have have not yet been confirmed. **Control condition** reveals that without suppression of the GPi output, both the model and the monkeys are able to quickly reach a very good performance level at the end of day 1 (D1, second column). **Protocol 1** D1 corresponds to the first day of the experiment where the habits are believed to have been acquired by the end of the day. D2 corresponds to the second day where GPi output is suppressed in both the model (removal of GPi-Thalamus connection) and the monkeys (muscimol injection). Even if the performances are a bit lower than at the end of D1, they are significantly higher than at the start of D1. These results in the monkeys allow us to predict performances for protocol 2 and D3 for the monkeys. (*) (resp. (**)) means significant difference with (1) (resp. (2)), all other differences being not significant.

Material and Methods

Behavioral experiments

Experimental procedures were performed in accordance with the Council Directive of 20 October 2010 (2010/63/UE) of the European Community. This project

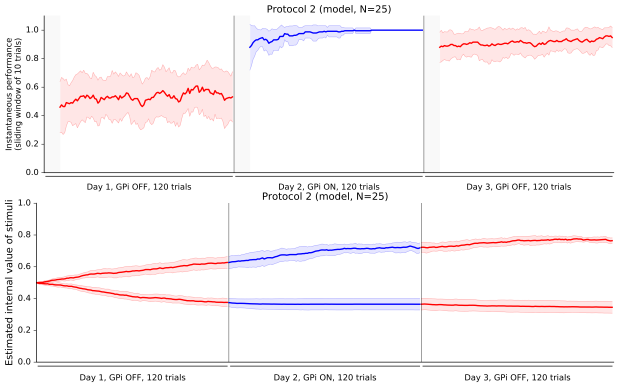


Figure 3: Theoretical results. Top panel. The measure of the instantaneous performance over a sliding window of 10 trials (strong line) clearly shows the discontinuity in performances between end of day 1 and start of day 2, indicating values have been learned during day 1. **Bottom panel.** The measure of the internal estimation of stimuli value shows no discontinuity and is thus independent of the state of the GPi, as predicted by our hypothesis. Final values are 0.76± 0.02 and 0.35±0.04 respectively, which are close to the actual reward probability of the two stimuli (0.75 and 0.25 respectively). Since the lowest probability stimulus is less often chosen (after learning has occurred) it is expected its value to be only roughly approximated.

was approved by the French Ethic Comity for Animal Experimentation (#50120111-A). Data were obtained from two female macaque monkeys that were previously used in a related set of experiments. All the details concerning animal care, experimental setup, surgical procedure, bilateral inactivation of the GPi and histology can be found in [28].

Computational modeling

Architecture

The model is an extension of previously published models [13, 25]. The model by [25] introduced an action selection mechanism which derives from the competition between a positive feedback through the direct pathway and a negative feedback through the hyperdirect pathway in the cortico-basal-thalamic loop. The model has been extended in [13] in order to explore the parallel organization of circuits in the BG. This model includes all the major nuclei of the basal ganglia (but GPe) and is organized along three segregated loops (motor, associative and cognitive) that spread over the cortex, the basal ganglia and the thalamus [29–32]. It incorporates a two-level decision making with a cognitive level selection (lateral prefrontal cortex, LPFC) based on cue shape and a motor level selection (supplementary motor area, SMA, and primary motor cortex, PMC) based on cue position (see Fig 4). In this latter model, the cortex is mostly an input/output structure under the direct influence of both the task input and the thalamic output resulting from the basal ganglia computations. Consequently, this cortex cannot take a decision of it own. In the present work, and to cope with our main hypothesis, we added a lateral competition mechanism in all three cortices (motor, cognitive, associative) based on short range excitation and long range inhibition and connections between the associative cortex and cognitive (resp. motor ones) to allow for the cross talking of these structures. This competition results in the capacity for the cortex to make a decision as shown on Fig 5, although slower than the BG.

Dynamics

The dynamic of a decision in the model is illustrated on Fig 5 before any learning has occurred. The top part shows the dynamics of the unlesionned model where a decision occurs a few milliseconds after stimulus onset. However, in the lesioned model, the suppression of the GPi output slows down considerably the decision process compared to the intact model. This means that the decision is initially driven by the basal ganglia.

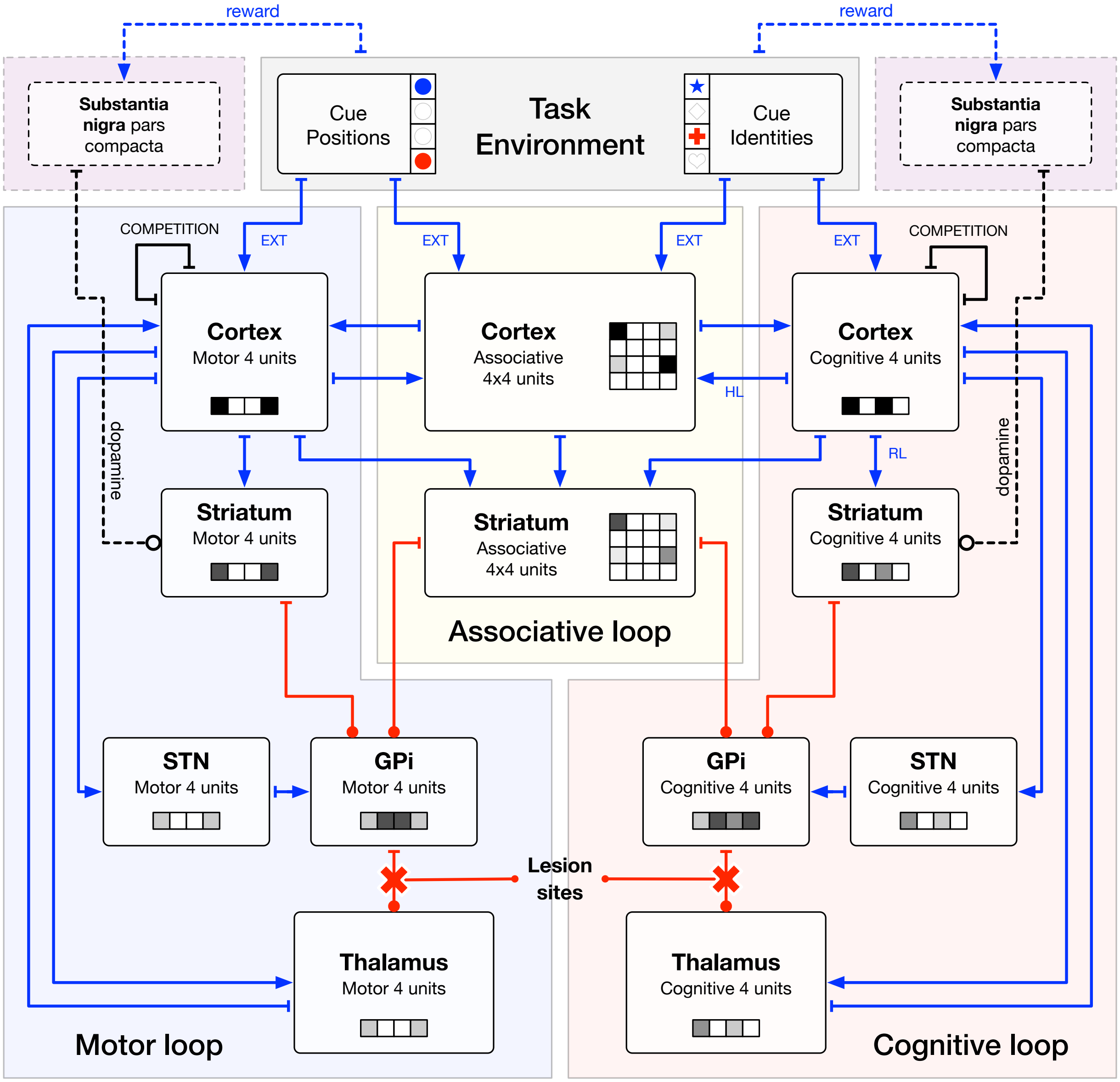


Figure 4: Architecture of the model. The computational model is made of 12 neural groups organized along three segregated loops (motor, associative and cognitive) that spread over the cortex, the basal ganglia and the thalamus. Blue lines represent excitatory pathways, red lines represent inhibitory pathways and dashed lines represent emulated pathways (they are not physically present in the model but their influence is taken into account). Red crosses represent lesion sites emulating the muscimol injection in the GPi of the monkeys. The color of the different units has only an illustrative purpose and does not represent actual activation.

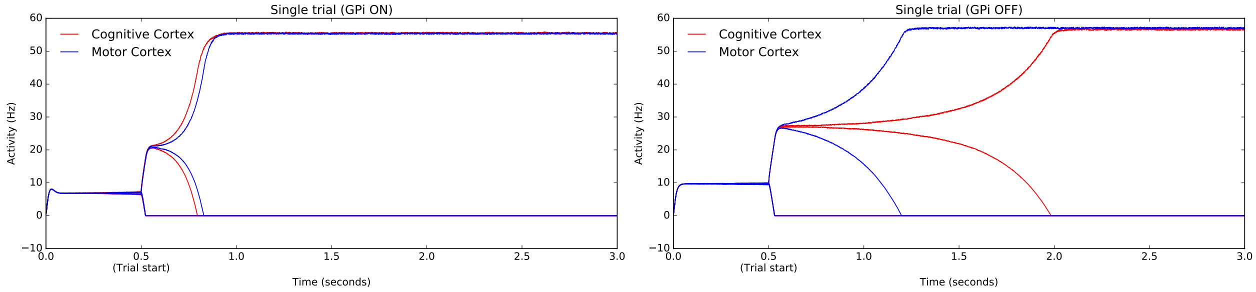


Figure 5: Dynamic of a decision At the beginning of the trial, all the neurons are settled to their steady state. Each red line represents one cognitive cortical neuron and each blue line one motor cortical neuron. At t=500 ms, stimuli are presented to the model. The activity of cortical neurons associated to cues or positions that are not shown are immediately suppressed while the activity of the present cues and positions start to compete such that in the end only one cognitive and one motor neuron stay active (i.e. a decision has been made). The suppression of the GPi output (bottom panel) slows down considerably the decision process compared to the intact model (top panel).

Learning

Dopamine modulates learning using reinforcement learning (RL) between the cognitive cortex and the cognitive striatum such that the decision made at the cognitive level can be used to bias the decision at the motor level. Hebbian learning (HL) occurs just after a motor action has been selected and carried out and modifies the connections (LTP) between the cognitive cortex and the associative cortex. It *does not depend* on reward but only on the actual cognitive and motor choices. It is to be noted that the cortical selection (resulting from lateral competition in the cortex) is slower than the cortico-basal selection (see Fig 4) such that the cortex is initially driven by the basal ganglia output (GPi), hence it learns from the statistics provided by the BG selection.

Lesion

Lesion in the model is made through the removal of all the connections between the motor (resp. cognitive) GPi and the motor (resp. cognitive) thalamus (red crosses on Fig 4). This prevents any communication from the basal ganglia to the model but keep intact the communication from the cortex to the basal ganglia.

Data analysis

For statistic analyses, unless stated otherwise, data are shown as the mean ± standard deviation. We used the multi-way repeated measures analysis of variance (ANOVA) to examine relations between session (n=25), periods (beginning: 25 first; end: 25 last trials), sessions (Control or Protocol 1 or Protocol 2). Post-hoc comparisons were conducted by using Bonferroni for the simulations but Kruskal-Wallis for the data (non parametric analysis) when the ANOVA showed significant differences. Significance was set at P<0.001.

Model

The ANOVA revealed a significant difference between the different simulation conditions (F=90.87; DF=349; p<0.001). Bonferroni post-hoc analysis (P<0.001) showed that beginning of control period was significantly different to all other condition except beginning of D1 in protocol 1 (BD1-P1), beginning of D1 in protocol 2 (BD1-P2) and end of D1 protocol 2 (ED1-P2). More specifically, this shows that there is a significant difference in performance between the end of day 1 and the start of of day 2 for protocol 2 and no difference between end of day2 and start of day 3.

Monkeys

The ANOVA revealed a significant difference between the different simulation conditions (F=28.18; DF=117; p<0.001). Kruskall-Wallis post-hoc analysis (P<0.001) showed that: beginning of control period was significantly different to all other condition except beginning of D1, protocol 1 (BD1-P1), beginning of D1, protocol 2 (BD1-P2), and end of D1, protocol 2 (ED1-P2). More specifically, this shows that there is a significant difference in performance between the end of day 1 and the start of of day 2 for protocol 2.

Discussion

These experimental results clearly demonstrate that Hebbian and reinforcement learning can be explicitly dissociated by inactivating the output of the basal ganglia while preserving the learning of the stimuli value in a two-arm bandit task. It suggests that a behavioral decision results from both the cooperation (acquisition) and competition (expression) of two distinct but entangled memory systems. To understand such hypothesis, it is important to consider how the basal ganglia forms a series of parallel loops (motor, associative, limbic) with the cortex and the thalamus [33]. In higher order mammals such as primates, the overall process starts in the sensory cortex, where stimuli are encoded, and ends up preferentially in the motor cortex from where an actual action is sent to the medullar motor neurones. Accordingly, in the previous versions of our model [25], the cortex was considered as a single input/output excitatory layer without intrinsic dynamic properties other than the I/O function of the populations. We then added a thalamic loop which allowed positive feedback but the different channels/populations were still independent [13, 26]. This limited autonomy is reasonable to mimic non-mammal vertebrate 3-layers dorsal mantle (aka pallium), but it was too rudimentary for the more complex architecture of the 6-layers mammal cortex [34]. The latter is remarkable for its organization in functional columns that are able to provide themselves positive feedback and to exert lateral inhibition on their neighbors. This architecture grants the cortex with dynamic properties that are far beyond what we were able to capture with our previous versions of the model: the balance between activation and inhibition allowed to toggle into various states that could be segregated enough to trigger different decisions [35]. We therefore decided to add a cortical module encompassing these properties (see Fig 4 and method section). We also grant it with the capacity to perform Hebbian learning based on the consensual hypothesis that this property is shared by most of the cortical structures [36–40].

This new model captures the very essence of animal behavior in our two-armed bandit like task [28] and provides also a non-intuitive prediction about the occurring of covert learning in the striatum while the output of the BG is disrupted (Fig 2 & 3). An upgraded version of our original task (Fig 1) allowed us to confirm this prediction in 2 animals (Fig 2). It also elegantly solves an old paradox about the absence of behavioral effect of BG surgical disruption/lesion in movement disorders: decision making and pre-learned motor task are not altered while learning of new paradigm should be disrupted (for discussion see [28]).

Most of the literature about habits stems from rodent research. It stands that goal directed behaviors and habits rely on 2 different systems that competed in order to generate a behavior in response to a given stimulus [41]. Probably influenced by the triune brain theory [42], still pregnant in experimental psychology despite it has been refuted by evolution biologists [43, 44], this model stands that the prefrontal cortical territories, belonging only to the mammals, are supposed to perform the “noble task”: the goal oriented behavior, while the BG, which appeared earlier in the evolution stored the “low level” task: the habits. The data supporting this hypothesis are indeed very clear concerning the need of an intact basal ganglia system in order to develop habits, but much less concerning the fact that they are need to perform in a habitual way [45]. In fact, there is also a growing body of evidences showing that learning occurred first in the BG and then is secondary transferred to the cortex [7, 46, 47]. Our model provides a plausible architecture to the latter hypothesis.

Beside conciliating diverging view, our hypothesis has also the advantage of being more ecological by saving cognitive resources. The dual theory needs a comparative mechanism/structure that arbitrates between goal oriented and habitual behaviors, while our architecture is based on cooperative/competitive parallel mechanisms that are entangled and adjust themselves without the need for an umpire.

It may worth notice that our model shares similarities with the so-called standard theory of systems-level memory consolidation in which the prefrontal cortex needs the hippocampus in order to create new memory traces, but once they are consolidated, the latter disengages itself and the former is able to retrieve remotes memories alone [48, 49]. We believe, that the two main memory systems (i.e. episodic and procedural) share similar mechanisms is a good index of plausibility for ecological reasons (but certainly not a proof *per se*).

Despite this promising start, our model needs further experiments to be confirmed. It is nevertheless interesting to notice that it reverses the old idea that automatism is a sub-cortical feature. Ad variance with the other vertebrates, the cortex of mammals (especially the primates) is both the main sensory input and motor output structure. The fact that automatic input/output association occurred there, bypassing a long sub-cortical journey and therefore saving cognitive resources is another strong ecological argument. If our model is confirmed by further experiments, it opens new questions such as: i) is it a mammal specificity? ii) a primate specificity? iii) how such automatisms are implemented or even are they implemented in other vertebrates? Whatever the answers, we are confident that we planted here the last nail in the coffin of this good old triune brain theory… Shall it rest in peace.

Appendix

Model description. The source code for the model as well as all the scripts for running each of the experiments are available from `https://github.com/rougier/basal-ganglia`. We provide below the tabular description of the model as following the prescription of [50].

References

[1] M Mishkin, B Malamut, and J Bachevalier. “Memories and habits: Two neural systems”. In: *Neurobiology of human learning and memory*. Ed. by G Lynch, J L McGaugh, and N M Weinberger. 1984.

[2] M. Mishkin and H. L. Petri. “Memories and habits: Some implications for the analysis of learning and retention.” In: *Neuropsychology of Memory*. Ed. by L. R. Squire and N. Butters. Guilford, 1984.

[3] A Dickinson. “Actions and Habits: The Development of Behavioural Autonomy”. In: *Philosophical Transactions of the Royal Society of London. Series B* 308.1 (Feb. 1985), pp. 67–78.

[4] Endel Tulving. “How many memory systems are there?” In: *American psychologist* 40.4 (1985), pp. 385–398.

[5] Carol A Seger and Brian J Spiering. “A critical review of habit learning and the Basal Ganglia”. In: *Frontiers in systems neuroscience* 5 (2011).

[6] Hisham E Atallah et al. “Separate neural substrates for skill learning and performance in the ventral and dorsal striatum”. In: *Nature Neuroscience* 10.1 (Dec. 2006), pp. 126–131.

A Model Summary

Populations	Twelve: Cortex (motor, associative & cognitive), Striatum (motor, associative & cognitive), GPi (motor & cognitive), STN (motor & cognitive), Thalamus (motor & cognitive)
Topology	–
Connectivity	one to one, one to many (divergent), many to one (convergent)
Neuron model	Dynamic rate model
Channel model	–
Synapse model	Linear synapse
Plasticity	Reinforcement learning rule
Input	External current in cortical areas (motor, associative & cognitive)
Measurements	Firing rate

B Populations

Name	**Elements**	**Size**	**Threshold** (h)	**Noise**	**Initial state**
Cortex motor	Linear neuron	1×4	-3	1.0%	0.0
Cortex cognitive	Linear neuron	4×1	-3	1.0%	0.0
Cortex associative	Linear neuron	4×4	-3	1.0%	0.0
Striatum motor	Sigmoidal neuron	1×4	0	0.1%	0.0
Striatum cognitive	Sigmoidal neuron	4×1	0	0.1%	0.0
Striatum associative	Sigmoidal neuron	4×4	0	0.1%	0.0
GPi motor	Linear neuron	1×4	+10	3.0%	0.0
GPi cognitive	Linear neuron	4×1	+10	3.0%	0.0
STN motor	Linear neuron	1×4	-10	0.1%	0.0
STN cognitive	Linear neuron	4×1	-10	0.1%	0.0
Thalamus motor	Linear neuron	1×4	-40	0.1%	0.0
Thalamus cognitive	Linear neuron	4×1	-40	0.1%	0.0
Values (V_i)	Scalar	4	–	–	0.5

C Connectivity

Source	Target	Pattern	Weight	Gain	Plasticity
Cortex motor	Thalamus motor	$(1, i) \rightarrow (1, i)$	1.0	0.4	-
Cortex cognitive	Thalamus cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	0.4	-
Cortex motor	STN motor	$(1, i) \rightarrow (1, i)$	1.0	1.0	-
Cortex cognitive	STN cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	1.0	-
Cortex motor	Striatum motor	$(1, i) \rightarrow (1, i)$	0.5	1.0	-
Cortex cognitive	Striatum cognitive	$(i, 1) \rightarrow (i, 1)$	0.5	1.0	(F1)
Cortex motor	Striatum associative	$(1, i) \rightarrow (*, i)$	0.5	0.2	-
Cortex cognitive	Striatum associative	$(i, 1) \rightarrow (i, *)$	0.5	0.2	-
Cortex associative	Striatum associative	$(i, j) \rightarrow (i, j)$	0.5	1.0	-
Thalamus motor	Cortex motor	$(1, i) \rightarrow (1, i)$	1.0	1.0	-
Thalamus cognitive	Cortex cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	1.0	-
GPi motor	Thalamus motor	$(1, i) \rightarrow (1, i)$	1.0	-0.5	-
GPi cognitive	Thalamus cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	-0.5	-
STN motor	GPi motor	$(1, i) \rightarrow (1, i)$	1.0	1.0	-
STN cognitive	GPi cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	1.0	-
Striatum cognitive	GPi cognitive	$(i, 1) \rightarrow (i, 1)$	1.0	-2.0	-
Striatum motor	GPi motor	$(i, 1) \rightarrow (i, 1)$	1.0	-2.0	-
Striatum associative	GPi motor	$(*, i) \rightarrow (1, i)$	1.0	-2.0	-
Striatum associative	GPi cognitive	$(i, *) \rightarrow (i, 1)$	1.0	-2.0	-
Cortex motor	Cortex motor	$(1, i) \rightarrow (1, *)$	1.0	-0.5	-
Cortex cognitive	Cortex cognitive	$(1, i) \rightarrow (1, *)$	1.0	-0.5	-
Cortex associative	Cortex associative	$(i, j) \rightarrow (*, *)$	1.0	-0.5	-
Cortex motor	Cortex associative	$(1, i) \rightarrow (*, i)$	1.0	0.01	-
Cortex associative	Cortex cognitive	$(i, *) \rightarrow (i, 1)$	1.0	0.01	-
Cortex cognitive	Cortex associative	$(i, 1) \rightarrow (i, *)$	1.0	0.025	(F2)
Cortex associative	Cortex motor	$(*, i) \rightarrow (1, i)$	1.0	0.025	-

D1 Neuron Model

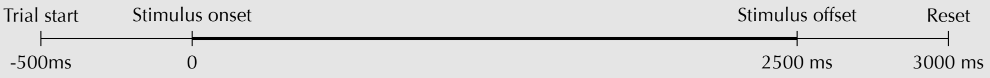
Name	Linear neuron
Type	Rate model
Membrane Potential	$\tau dV/dt = -V + I_{syn} + I_{ext} - h$
	$U = max(V, 0)$

D2 Neuron Model	
Name	Sigmoidal neuron
Type	Rate model
Membrane Potential	$\tau dV/dt = -V + I_{syn} + I_{ext} - h$
	$U = V_{min} - (V_{max} - V_{min})/\left(1 + e^{\frac{V_h - V}{V_c}}\right)$

E Synapse	
Name	Linear synapse
Type	Weighted sum
Output	$I^B_{syn} = \sum_{A \in sources}(G_{A\rightarrow B} W_{A\rightarrow B} U_A)$

F1 Plasticity	
Name	Reinforcement learning
Type	Delta rule
Delta	$\Delta W_{A\rightarrow B} = \alpha \times PE \times U_B$
	$PE = Reward - V_i$
	$\alpha = 0.004$ if $PE < 0$ (LTD), $\alpha = 0.005$ if $PE > 0$ (LTP)
	$\Delta V_i = \beta \times PE, \beta = 0.0125$

F2 Plasticity	
Name	Hebbian learning
Type	Hebb rule
Delta	$\Delta W_{A\rightarrow B} = \alpha \times U_A \times U_B$, $\alpha = 0.00025$

G Input	
Type	Cortical input
Description	A trial is preceded by a settling period (500ms) and followed by a reset period. At time $t = 0$, two shapes are presented in cortical cognitive area ($I_{ext} = 7$ at $\{i_1, i_2\}$) at two different locations in cortical motor area ($I_{ext} = 7$ at $\{j_1, j_2\}$) and the cortical associate area is updated accordingly ($I_{ext} = 7$ at $\{i_1, i_2\} \times \{j_1, j_2\}$).
Timing	 Trial start Stimulus onset Stimulus offset Reset -500ms 0 2500 ms 3000 ms

H Measurements	
Site	Cortical areas
Data	Activity in cognitive and motor cortex
	Cortico-striatal weights

I Environment	
OS	OSX 10.11 (El Capitan)
Language	Python 3.5.1 (brew installation)
Libraries	Numpy 1.10.2 (pip installation)
	Cython 0.23.4 (pip installation)
	Matplotlib 1.5.0 (pip installation)

[7] Ann M Graybiel. "Habits, Rituals, and the Evaluative Brain". In: *Annual review of neuroscience* 31.1 (2008).

[8] Okihide Hikosaka and Masaki Isoda. "Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms". In: *Trends in Cognitive Sciences* 14.4 (Apr. 2010), pp. 154–161.

[9] Ray J Dolan and Peter Dayan. "Goals and Habits in the Brain". In: *Neuron* 80.2 (Oct. 2013), pp. 312–325.

[10] Fiery Cushman and Adam Morris. "Habitual control of goal selection in humans". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.45 (Nov. 2015), pp. 13817–13822.

[11] Ann M Graybiel and Scott T Grafton. "The Striatum: Where Skills and Habits Meet". In: *Cold Spring Harbor Perspectives in Biology* 7.8 (Aug. 2015), a021691–14.

[12] B Pasquereau et al. "Shaping of Motor Responses by Incentive Values through the Basal Ganglia". In: *Journal of Neuroscience* 27.5 (2007).

[13] M Guthrie et al. "Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study". In: *Journal of neurophysiology* 109.12 (2013).

[14] R. J. Herrnstein. "Formal properties of the matching law". In: *J Exp Anal Behav* 21.1 (1974).

[15] D. A. Graft, S. E. Lea, and T. L. Whitworth. "The matching law in and within groups of rats". In: *J Exp Anal Behav* 27.1 (1977).

[16] L. R. Matthews and W. Temple. "Concurrent schedule assessment of food preference in cows". In: *J Exp Anal Behav* 32.2 (1979).

[17] C. M. Bradshaw et al. "The effect of signaled reinforcement availability on concurrent performances in humans". In: *J Exp Anal Behav* 32.1 (1979).

[18] J. D. Dougan, F. K. McSweeney, and V. A. Farmer. "Some parameters of behavioral contrast and allocation of interim behavior in rats". In: *J Exp Anal Behav* 44.3 (1985).

[19] R. J. Herrnstein et al. "Teaching pigeons an abstract relational rule: insideness". In: *Percept Psychophys* 46.1 (1989).

[20] B. Lau and P. W. Glimcher. "Dynamic response-by-response models of matching behavior in rhesus monkeys". In: *J Exp Anal Behav* 84.3 (2005).

[21] B. Lau and P. W. Glimcher. "Value representations in the primate striatum during matching behavior". In: *Neuron* 58.3 (2008).

[22] L. B. Gilbert-Norton, T. A. Shahan, and J. A. Shivik. "Coyotes (Canis latrans) and the matching law". In: *Behav Processes* 82.2 (2009).

[23] Henry H Yin and Barbara J Knowlton. "The role of the basal ganglia in habit formation". In: *Nature Reviews Neuroscience* 7.6 (2006).

[24] Mark G Packard and Barbara J Knowlton. "Learning and Memory Functions of the Basal Ganglia". In: *Annual review of neuroscience* 25.1 (2002).

[25] A Leblois, T Boraud, and W Meissner. "Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia". In: *The Journal of Neuroscience* 26.13 (2006).

[26] Meropi Topalidou and Nicolas P. Rougier. "[Re] Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study". In: *ReScience* 1.1 (2015).

[27] K. Doya. "Complementary roles of basal ganglia and cerebellum in learning and motor control". In: *Current Opinion in Neurobiology* 10.6 (2000).

[28] C Piron et al. "The Globus Pallidus pars interna in goal oriented and routine behaviours. Resolving a long standing paradox". In: *Movement disorders* -.- (2016). in press.

[29] G E Alexander, M R DeLong, and P L Strick. "Parallel organization of functionally segregated circuits linking basal ganglia and cortex". In: *Annual review of neuroscience* 9.1 (1986).

[30] R L Albin, A B Young, and J B Penney. “The functional anatomy of basal ganglia disorders”. In: *Trends in neurosciences* 12.10 (1989).

[31] G E Alexander and M D Crutcher. “Preparation for movement: neural representations of intended direction in three motor areas of the monkey”. In: *Journal of neurophysiology* (1990).

[32] A. Parent and L.N. Hazrati. “Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop”. In: *Brain Research Reviews* 20.1 (1995).

[33] Suzanne N Haber. “The primate basal ganglia: parallel and integrative networks”. In: *Journal of Chemical Neuroanatomy* 26.4 (2003).

[34] Gordon M Shepherd. “The Microcircuit Concept Applied to Cortical Evolution: from Three-Layer to Six-Layer Cortex”. In: *Frontiers in neuroanatomy* 5 (2011), pp. 1–15.

[35] Nicolas P. Rougier and Julien Vitay. “Emergence of Attention within a Neural Population”. In: *Neural Networks* 19.5 (2006), pp. 573–581.

[36] Mark F. Bear and Robert C. Malenka. “Synaptic plasticity: LTP and LTD”. In: *Current Opinion in Neurobiology* 4.3 (1994), pp. 389–399.

[37] Natalia Caporale and Yang Dan. “Spike Timing–Dependent Plasticity: A Hebbian Learning Rule”. In: *Annual Review of Neuroscience* 31.1 (2008), pp. 25–46.

[38] Daniel E. Feldman. “Synaptic Mechanisms for Plasticity in Neocortex”. In: *Annual Review of Neuroscience* 32.1 (2009), pp. 33–55.

[39] F G Ashby, B O Turner, and J C Horvitz. “Cortical and basal ganglia contributions to habit learning and automaticity”. In: *Trends in Cognitive Sciences* (2010).

[40] Naoki Hiratani and Tomoki Fukai. “Hebbian Wiring Plasticity Generates Efficient Network Structures for Robust Inference with Synaptic Weight Plasticity”. In: *Frontiers in Neural Circuits* 10.41 (2016).

[41] N. D. Daw, Y. Niv, and P. Dayan. “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control”. In: *Nat Neurosci* 8.12 (2005), pp. 1704–11.

[42] P. D. MacLean. “Evolutionary psychiatry and the triune brain”. In: *Psychol Med* 15.2 (1985), pp. 219–21.

[43] O. Marin, W. J. Smeets, and A. Gonzalez. “Evolution of the basal ganglia in tetrapods: a new perspective based on recent studies in amphibians”. In: *Trends Neurosci* 21.11 (1998), pp. 487–94.

[44] Z. Molnar. “Evolution of cerebral cortical development”. In: *Brain Behav Evol* 78.1 (2011), pp. 94–107.

[45] B. W. Balleine and J. P. O’Doherty. “Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action”. In: *Neuropsychopharmacology* 35.1 (2010), pp. 48–69.

[46] A. Pasupathy and E. K. Miller. “Different time courses of learning-related activity in the prefrontal cortex and striatum”. In: *Nature* 433.7028 (2005), pp. 873–6. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking).

[47] H. E. Atallah et al. “Separate neural substrates for skill learning and performance in the ventral and dorsal striatum”. In: *Nat Neurosci* 10.1 (2007), pp. 126–31.

[48] Larry R Squire. “Memory systems of the brain: A brief history and current perspective”. In: *Neurobiology of Learning and Memory* 82.3 (2004).

[49] P. W. Frankland and B. Bontempi. “The organization of recent and remote memories”. In: *Nat Rev Neurosci* 6.2 (2005), pp. 119–30.

[50] Eilen Nordlie, Marc-Oliver Gewaltig, and Hans Ekkehard Plesser. “Towards Reproducible Descriptions of Neuronal Network Models”. In: *PLoS computational biology* 5.8 (2009).