The formation of habits: a computational model mixing reinforcement and Hebbian learning

Meropi Topalidou, Daisuke Kase, Thomas Boraud, Nicolas Rougier

The formation of habits: a computational model mixing reinforcement and Hebbian learning

Meropi Topalidou
INRIA Bordeaux Sud-Ouest, Bordeaux, France
Université de Bordeaux, CNRS UMR 5293, IMN, France
LaBRI, Université de Bordeaux, IPB, CNRS, UMR 5800, Talence, France
meropi.topalidou@inria.fr

Daisuke Kase
Université de Bordeaux, CNRS UMR 5293, IMN, France
Laboratoire Franco-Israélien de Neurosciences, CNRS Bordeaux, France
daisuke.kase@u-bordeaux.fr

Thomas Boraud
Université de Bordeaux, CNRS UMR 5293, IMN, France
thomas.boraud@u-bordeaux.fr

Nicolas Rougier
INRIA Bordeaux Sud-Ouest, Bordeaux, France
Université de Bordeaux, CNRS UMR 5293, IMN, France
LaBRI, Université de Bordeaux, IPB, CNRS, UMR 5800, Talence, France
nicolas.rougier@inria.fr

Abstract

If basal ganglia are widely accepted to participate in the high-level cognitive function of decision-making, their role is less clear regarding the formation of habits. One of the biggest problem is to understand how goal-directed actions are transformed into habitual responses, or, said differently, how an animal can shift from an action-outcome (A-O) system to a stimulus-response (S-R) one while keeping a consistent behaviour.

We introduce a computational model (basal ganglia, thalamus and cortex) that can solve a simple two arm-bandit task using reinforcement learning and explicit valuation of the outcome (Guthrie et al. (2013)). Hebbian learning has been added at the cortical level such that the model learns each time a move is issued, rewarded or not. Then, by inhibiting the output nuclei of the model (GPi), we show how learning has been transferred from the basal ganglia to the cortex, simply as a consequence of the statistics of the choice. Because best (in the sense of most rewarded) actions are chosen more often, this directly impacts the amount of Hebbian learning and lead to the formation of habits within the cortex.

These results have been confirmed in monkeys (unpublished data at the time of writing) doing the same tasks where the BG has been inactivated using muscimol. This tends to show that the basal ganglia implicitely teach the cortex in order for it to learn the values of new options. In the end, the cortex is able to solve the task perfectly, even if it exhibits slower reaction times.

Keywords: basal ganglia, decision making, habits, Hebbian learning, reinforcement learning

1 Introduction

According to Schneider and Shiffrin (1977), a behaviour is automatic (i.e. becomes an habit) if a sensory events always elicit the same behaviour, even if the subject is doing something else. Think, for example, of someone entering a dark room while talking on the phone and switching on the light without ever really thinking about it. How this behaviour is acquired in the first place? How do we learn such habits? Seger and Spiering (2011), characterized habit learning using five definitional features: inflexible, incremental, unconscious, automatic, and insensitive to reinforcer devaluation. This tentative definition seems to be clearly in opposition with decision making that we could think of as flexible and highly sensitive to reinforcer devaluation. However, there are more and more evidences these two types of learning are somehow linked together (Yin and Knowlton (2006)), the question being how?

In a recent (unpublished) study, monkeys have been tested on a two-armed bandit task using a pharmacological approach, combining both decision making and procedural learning. Primates have been daily accustomed with the setup which is composed of four buttons placed on different directions (0, 90, 180 and 270) and another one on central position which detects the contact with monkeys hands. These buttons correspond to the four possible appearance directions of a cursor on a perpendicular screen. The monkeys initiated a trial keeping their hand on the central button, which induced the appearance of the cursor in the central position of the screen. After a random delay (0.5-1.5 s), two cues appeared in two (out of four) different positions determined randomly for each trial. Two experimental conditions were alternated by blocks of ten trials. On Habitual Condition (HC), the two cues (HC1 and HC2) are the ones with which the animals have been trained. Each cue had a fixed probability for monkeys to be rewarded (PHC1=0.75 and PHC2=0.25). The nature and the probability of each cue remained the same during each working session and between each working session. On Novelty Condition (NC), the two cues presented are new (NC1 and NC2). Each cue had a fixed probability for monkeys to be rewarded (PNC1=0.75 and PNC2=0.25). Once the cues are shown, the monkeys had a delay period (0.5-1.5s) to press the button according to one cue during a random time (0.5-1.5s). The cursor appeared on the chosen cue. An "end-of-trial" signal corresponding to the disappearance of the cursor indicated to the monkeys that the trial is actually finished. Monkeys were rewarded (0.3 ml of water) or not according to the reward probability of the chosen target. They could then start a new trial after an inter-trial interval included between 500ms and 1.5 ms. The procedure is summarised in 1 In one condition, the internal part of the Globus Pallidus (the main output structure of the BG) is injected with a saline solution (no effect) and in the other condition, it is injected with muscimol (inactivation). Results tend to show that performances related to familiar cues stay unchanged, independently of GPi inactivation, while learning of new cues is deeply impacted when GPi has been inactivated. This tends to suggest BG might be critical in learning decision and this learning can be later transferred to the cortex. In accordance with this hypothesis, we build a computational model whose architecture is described in the next section.

2 Methods

The architectural basis of the model has been originally described in Guthrie et al. (2013) where authors introduced a biophysical model of action selection that can solve a two-arm bandit task, such as the one describe above. Two parallel action selection pathways compose the model with inputs from distinct areas of the cortex: one for handling the cognitive action selection, and the other for the motor selection. The model includes the cortex (Cx), the thalamus (Th) and several nuclei of the BG: striatum (Str), the subthalamic nuclei (STN) and the globus pallidus internal and external (GPi, GPe). Each module is made of a closed- loop positive feedback direct pathway (Cx-Str-GPi-Th-Cx) and two closed-loop negative feedbacks, indirect pathway (Cx-Str-GPe-STN-GPi-Th-Cx) and hyperdirect pathway (Cx-STN-GPi-Th-Cx), and is based on the center-surround architecture of Mink (1996). The interactions between these three pathways are able to induce an action selection at the motor level. However, the task requires first the actual selection of the best cue before performing the corresponding motor action. In the Guthrie et al. (2013) model, this is implemented at the striatal level where a dopamine reward signal is used to implement a simple value-based learning. The general architectural of the model is illustrated on Figure 2.

In the original model, the inactivation of the basal ganglia output (GPi) results in the inability of the model to make a decision since there is no competitive mechanism at the cortical level. We thus added lateral connections in each cortical modules (self-excitation and surround inhibition) such that a unique cognitive and motor decision can be made. At this point, there is no guarantee that the motor decision corresponds to the cognitive ones. The model can choose a cue A but moves toward the location of a cue B. To overcome this problem, we also need to establish a cross-talking between the different cortical modules, independently of the BG pathways. This has been made using excitatory connections from and to the associative cortical module. Hebbian learning in cortical level allows the cortex to make a consistent decision in the absence of GPi, even if it does not guarantee to make the optimal decision.

One important property of the cortical decision is that it is significantly slower than the BG decision. This can be shown by measuring the time of motor decision; defined as the time required for the difference between the two most activated units in the motor cortex to become greater than a given threshold (40Hz). Before learning, the BG decision time (GPi is

intact) is around 250ms while the cortical decision time (GPi is disabled) is around 800 ms. This difference in timing is actually critical for the BG to teach the cortex as explained in the results section.

3 Results

We tested the model using 4 different paradigms:

- HC/GPi: Habitual condition using already learned stimuli with intact GPi.
- HC/NoGPi: Habitual condition using already learned stimuli with lesioned GPi.
- NC/GPi: Novel condition using non familiar stimuli with intact GPi.
- NC/NoGPi: Novel condition using non familiar stimuli with lesioned GPi.

Each condition has been tested for 250 experiments where each experiment consists in 120 consecutive trials (presentation of the cues, decision, potential reward and reset). Before starting an new experiment, the model is trained on the familiar stimuli until the performance is over 0.95. Performances were measured as the ratio of optimal choices compared to the number of trials. Response time has been recorded as the time of the motor decision. This time is relative to the stimulus onset (t=500ms).

As shown in 3a, our results are equivalent to the experiments in monkeys (3b). In the habitual condition, performances are optimal with or without lesion, indicating the cortex is able to make the optimal decision without the help of the basal ganglia. However, in the novel condition, things are quite different. For the intact model, the model starts a trial at chance level, giving random choices. Nevertheless, after a few trials (around 15), it reaches a near-optimal performance, indicating the model has learned the respective reward probability associated with each cue. For the lesioned model, the performances stay at the level of chance, indicating the cortex is unable to learn the task "on its own". It is to be noted that due to Hebbian learning, the lesioned model tend to first choose a given random cue and stick to this choice until the end of the experiment. If this is the right cue, the performance can reach 1 for a single experiment, but over the course of the 250 experiments, the mean performance is around 0.5.

4 Conclusion

The aim of this model is to gain a better understanding of the role of the basal ganglia in the formation of habits. It is based on a previous model by Guthrie et al. (2013) that explain the dynamic of action selection in the BG. The model has been further refined with connections at the cortical level which are consistent with neuro-anatomy. We also implemented Hebbian learning at the cortical level, independently of reward. However, since BG helped to choose the best action anytime, this results in having cortical learning to be naturally modulated according to the value of the different cue, simply because the best cue is chosen more often. After learning, the cortex is able to choose the best cue without help of the BG, hence forming a new habit.

References

M. Guthrie, A. Leblois, A. Garenne, and T. Boraud. Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study. *Journal of Neurophysiology*, 2013.

J. Mink. The basal ganglia: focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 1996.

W. Schneider and R.M. Shiffrin. Controlled and automatic human information processing: I. detection, search, and attention. *Psychological Review*, 1977.

C.A. Seger and B. J. Spiering. A critical review of habit learning and the basal ganglia. *Frontiers in Systems Neuroscience*, 2011.

H.H. Yin and B. J. Knowlton. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 2006.

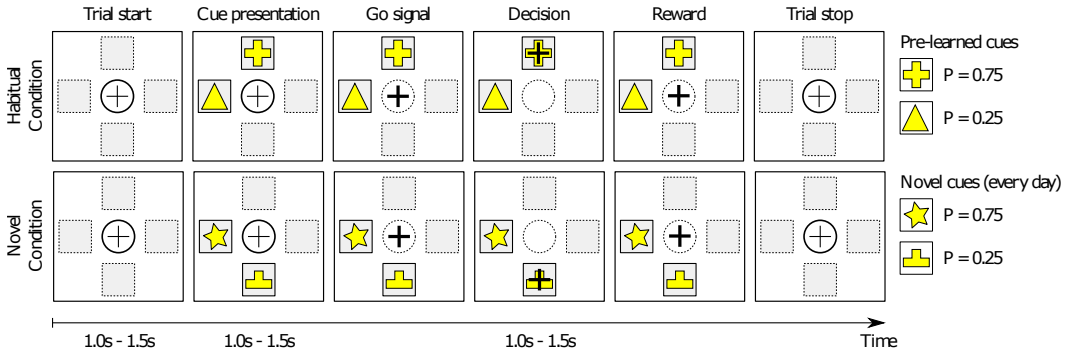


Figure 1: Behavioural paradigm. A session is made up of at least 250 trials broken down into alternate blocks of 10 trials in Habitual (top) or Novelty (bottom) Conditions. In each trial, two cues were displayed simultaneously in two out of four randomly chosen possible positions on the screen. The monkey signalled its choice by moving the cursor to one of the cues and was rewarded by 300 μl of fruit juice with a predefined fixed probability that depends on the choice. In the Habitual Conditions (top) the cues (HC1, P=0.75 and HC2, P=0.25) are the one with which the monkeys has been trained and therefore are familiar with. In the Novelty Condition (bottom), the cues (NC1 and NC2) have the same values (P=0,75 and P=0,25 respectively), but the pairs are changed (new shape and colours) for each session.

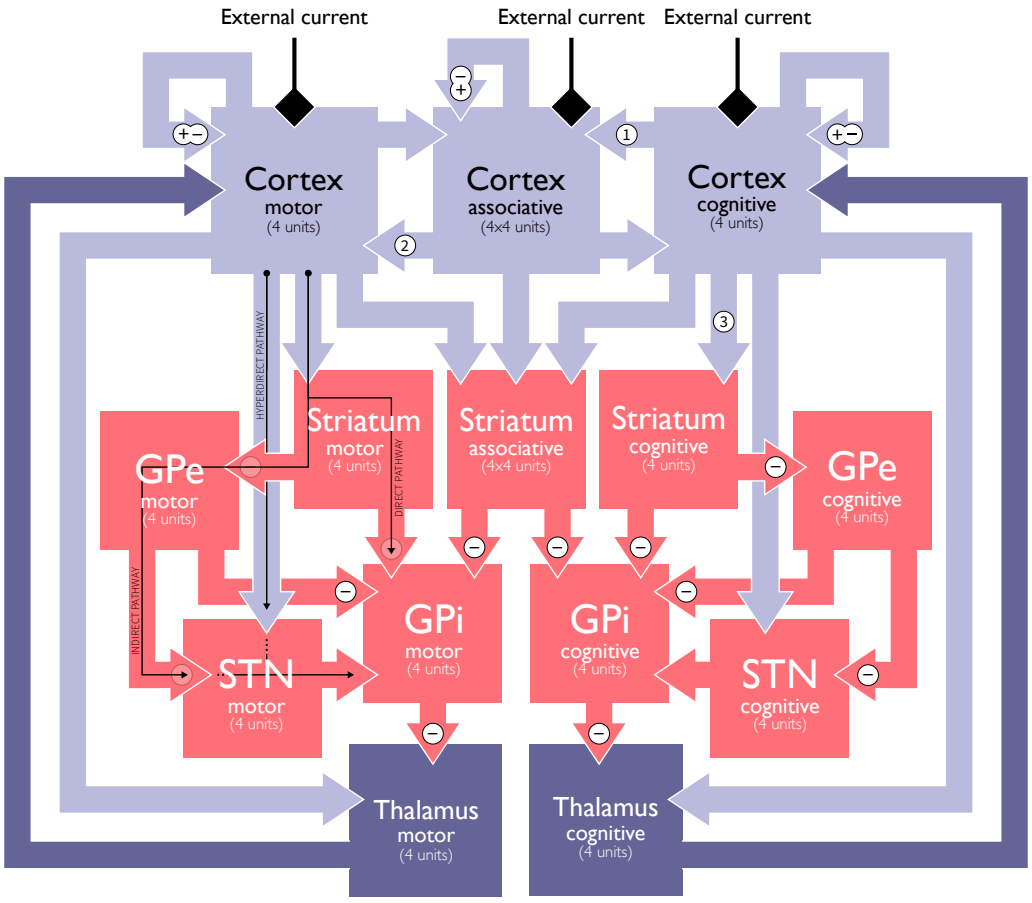
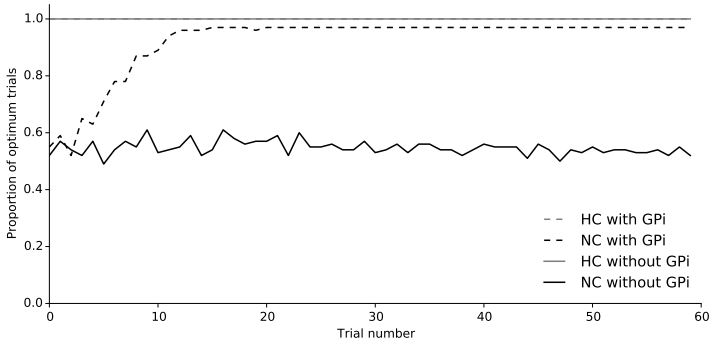
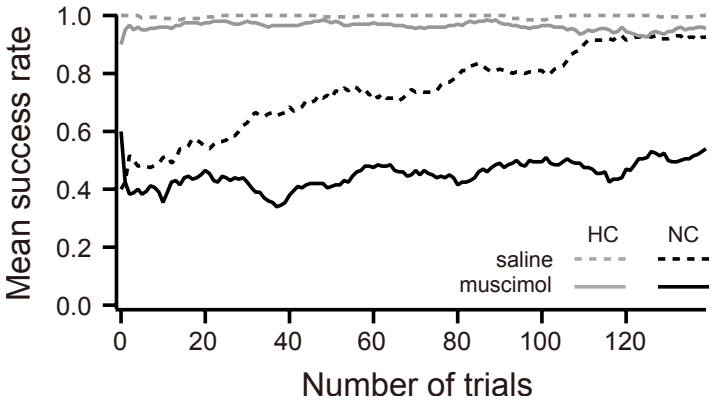


Figure 2: The main pathways in the model are the direct pathway (Cortex-Striatum-GPi-Thalamus-Cortex), the indirect pathway (Cortex-Striatum-GPe-STN-GPi-Thalamus-Cortex), and the hyperdirect pathway (Cortex-STN-GPi-Thalamus-Cortex). Learning occurs at two different levels: Hebbian learning from cognitive to associative cortex (1) and from associative to motor cortex (2), and reinforcement from cognitive cortex to cognitive striatum (3).



(a) Mean performances of model



(b) Mean performances of monkeys

Figure 3: Results averaged over 250 simulations. 3a) In HC, performances of the model are optimal (1), with or without GPi (the dashed line is not shown, because it coexists with the straight one). In NC, only the intact model (with GPi) is able to learn the new stimuli while lesioned model performances stay at the level of chance. 3b) Monkeys' performances are analogous to the models.