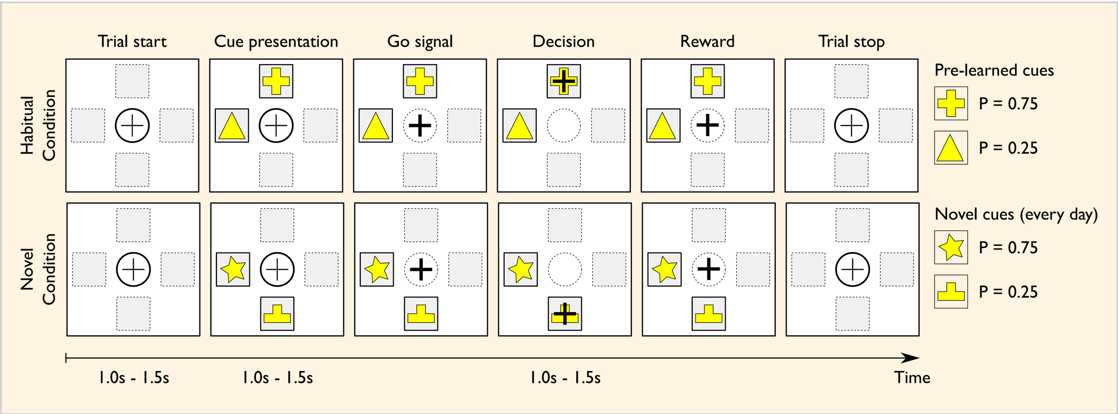
The formation of habits: A computational model mixing reinforcement learning and Hebbian learning

Meropi Topalidou, Camille Piron, Daisuke Kaise, Thomas Boraud, Nicolas P. Rougier

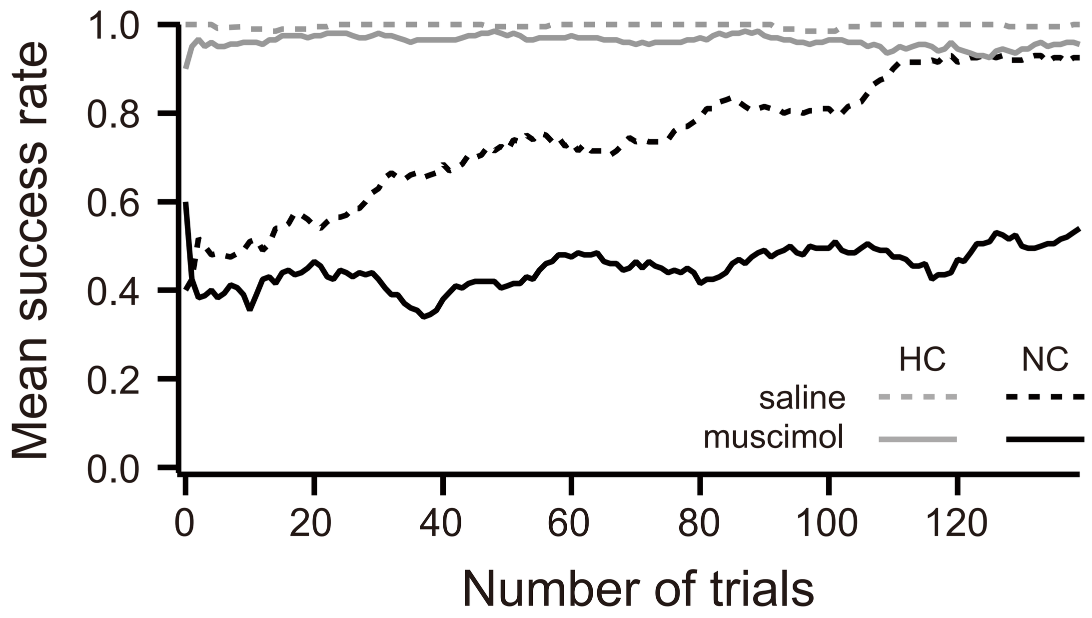
If basal ganglia are widely accepted to participate in the high-level cognitive function of decision-making, their role is less clear regarding the formation of habits. One of the hardest problem is to understand how goal-directed actions are transformed into habitual responses, or, said differently,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



Experiments Two monkeys were tested on a two-armed bandit task using 20 sessions in control conditions and 20 in muscimol conditions (10 for each monkey in each condition). We defined as success rate, the number of trials in which the animals chose the optimal target.

In *saline* conditions, the animals maximize their choice in the habitual condition and learn progressively the difference between the two cues in the novelty condition. They choose randomly at the beginning of training to finally display a clear preference for the target associated to the best reward.

In *muscimol* conditions (inhibition of the internal globus pallidus), the animals are still able to make the optimal choice in habitual conditions (with slower reaction time) but are unable to learn in novelty condition and make random choices from start to end of the session.

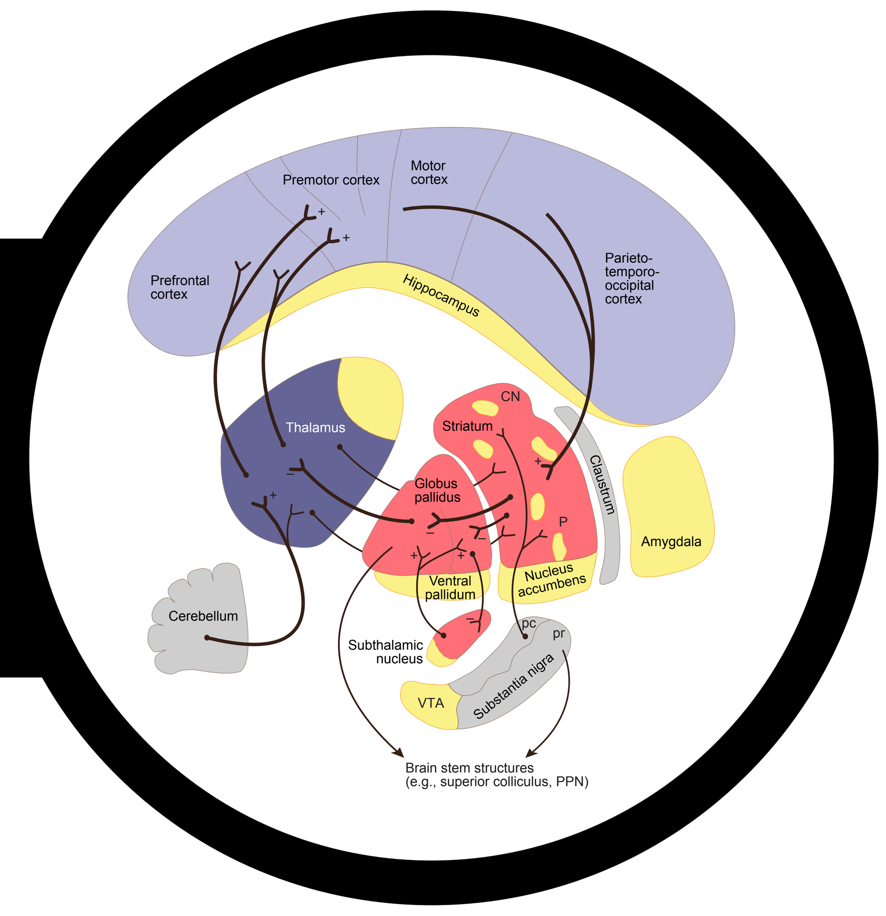


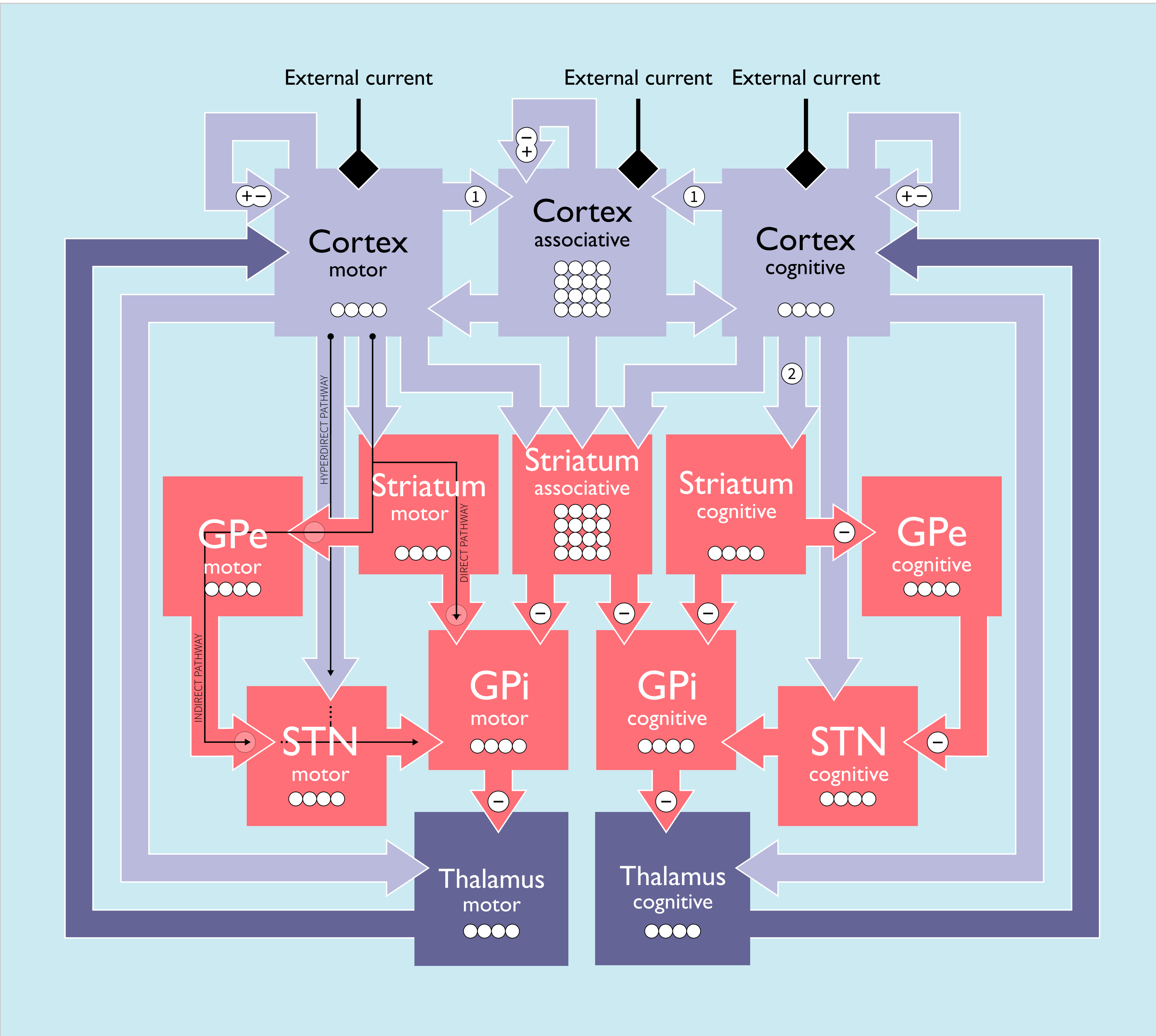
THE FORMATION OF HABITS

The implicit supervision of the basal ganglia

MEROPI TOPALIDOU - CAMILLE PIRON - DAISUKE KASE - THOMAS BORAUD - NICOLAS ROUGIER

Institut des Maladies Neurodégénératives - INRIA - CNRS - Université de Bordeaux





$$\tau\frac{dV}{dt} = -V + I_{ext} + I_{syn}, U = f(V)$$

$$I_{syn}^{A\rightarrow B} = \text{gain}\sum \times W_{A\rightarrow B} \times U_A$$

$$\text{Hebbian (1)}: \Delta W_{A\rightarrow B} = U_A \times U_B \times (W_{A\rightarrow B} - W_{min}) \times (W_{max} - W_{A\rightarrow B})$$

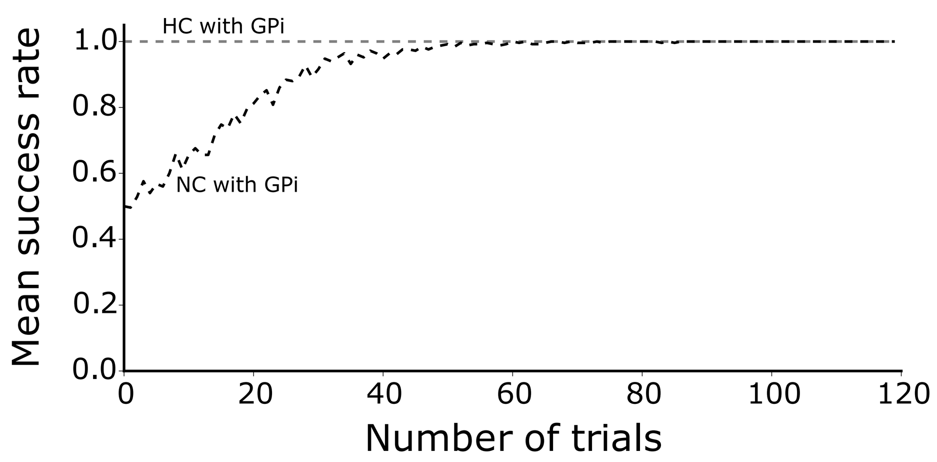
$$\text{Reinforcement (2)}: \Delta W_{A\rightarrow B} = \alpha \times PE \times U_B$$

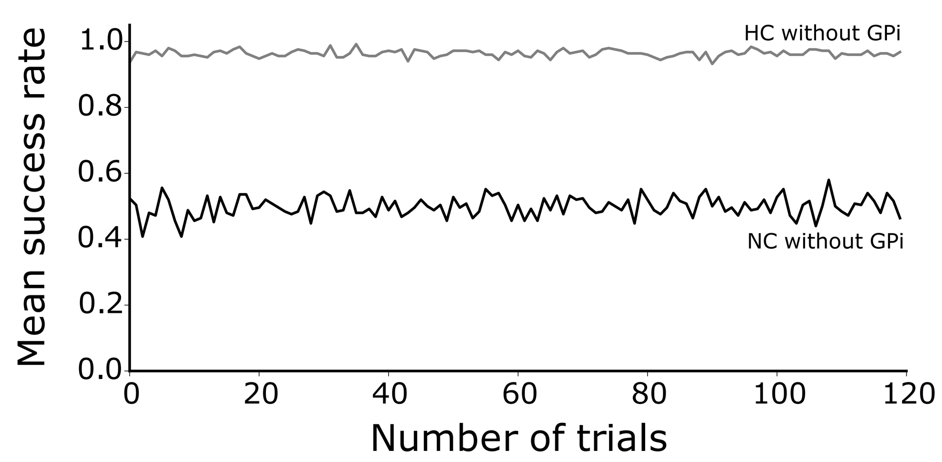
$$PE = \text{Reward} - V_i$$

The model is based on the model introduced in [1,2]. This former model introduces an action selection mechanism that is based upon the competition between a positive feedback through the direct pathway and a negative feedback through the hyperdirect pathway. The model has been further extended and exploits the parallel organization of circuits between the basal ganglia and the cortex [3] using segregated loops: one for making the selection between the two presented cues, and the other for making the selection between the two possible movement directions. To solve the task, it is necessary for the model to choose the cue shape and to select the right movement direction which depends upon the chosen cue. The model has been further refined such as to have a competition mechanisms within each cortical group. Using short range excitation and long range inhibitions, this competition ensures that a unique cognitive and motor decision eventually emerges, even if these decisions might be unrelated at this stage.

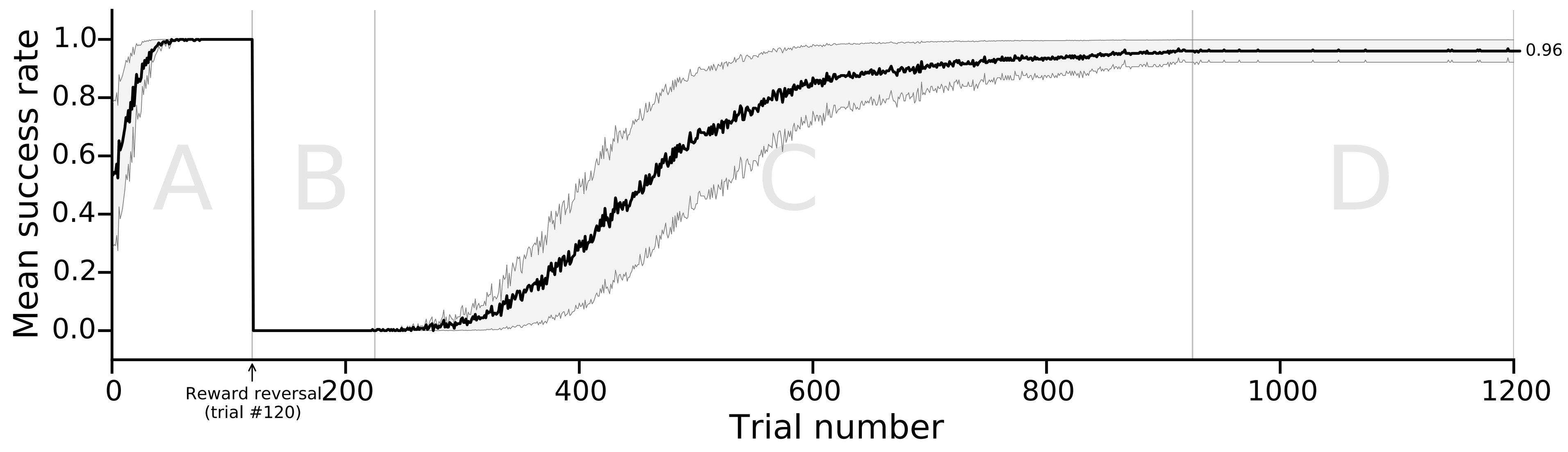
Learning occurs between the cognitive cortex and the cognitive striatum using a simple reinforcement learning where the value of the different cues are updated after each decision (see [2] for details). We added Hebbian learning (LTP) at the cortical level between the cognitive/motor cortical groups and the associative cortical group. This learning is enforced once per trial, at the time a move is made and independently of the actual reward. In habitual (resp. novelty) condition, the model is trained using cues 1 (resp. 3) & 2 (resp. 4) which are presented simultaneously at random positions. Cue 1 (resp. 3) is associated with a reward probability of 75% while cue 2 (resp. 4) is associated with a reward probability of 25%. In habitual condition, the model is trained until it achieves a mean performance of 0.95. This takes between 40 and 50 trials depending on the initial conditions (noise) and whether first cues are rewarded or not. This training impacts significantly Hebbian learning at the cortical level because cue 1 is chosen most of the time and consequently, the associative link relative to cue 1 is strengthened compared to associative link relative to cue 2.

[1] 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, 109:3025–3040, 2013.
[2] Leblois A., Boraud T., Meissner W., Bergmann H., Hansel D. Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. J Neurosci 26: 3567–3583, 2006.
[3] Piron C., Daisuke K., Topalidou M., Goillandeau M., N'guyen T., Orignac H., Rougier N.P., Boraud T. The role of the basal ganglia in the formation of habits in monkeys, *submitted*.

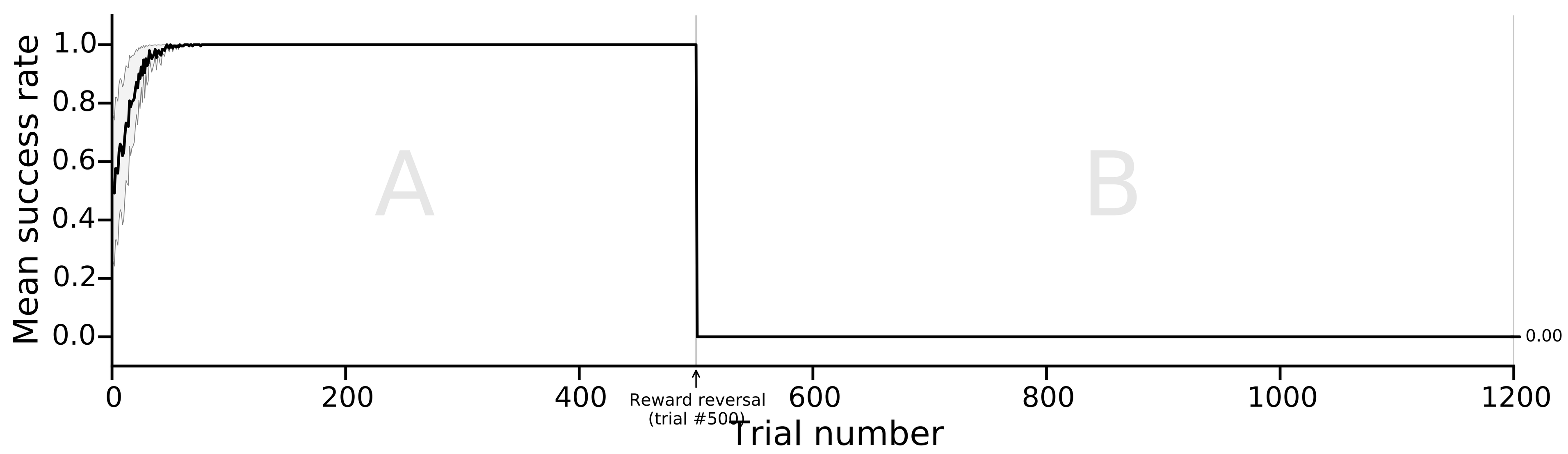




Results are in accordance with the experiments in monkeys. In the habitual condition, performances are optimal with or without lesion, indicating the cortex is able to make the optimal decision without the help of the basal ganglia if it has been learned previously. In novel condition, performances of the intact model are initially at chance level but after a few trials, it reaches a near-optimal performance, indicating the model has learned the respective reward probability associated with each novel cues. However, for the lesioned model, performances stay at chance level, indicating the cortex is unable to learn the new task without the help of the basal ganglia.



Habits are better expressed following extensive training on the two armed-bandit task, when the reward probability is reversed. If behavior were goal directed, then the model should react to the devaluation of reward. However, there is period (B) during which model persists in old behavior and seems to be inflexible to such devaluation. But, if extensive training is not too strong, model can recover and switch back to a goal-directed behavior that ultimately overcome habits.



Modelling | Experiments | Experiments | Experiments | Modelling

Meropi Topalidou | Camille Piron, PhD | Daisuke Kase, PhD | Thomas Boraud, MD, PhD | Nicolas Rougier, PhD