Habits That Contradict Rewards

Fabien C. Y. Benureau
Inria Bordeaux Sud-Ouest, France
Univ. of Bordeaux, UMR 5293, IMN, France
LaBRI, CNRS, UMR 5800, France
fabien.benureau@gmail.com
ORCiD: 0000-0003-4083-4512

Thomas Boraud
Univ. of Bordeaux, UMR 5293, IMN, France
CNRS, French-Israeli Neuroscience Lab, France
CHU de Bordeaux, IMN Clinique, France
thomas.boraud@u-bordeaux.fr
ORCiD: 0000-0002-8942-0129

Nicolas P. Rougier
Inria Bordeaux Sud-Ouest, France
Univ. of Bordeaux, UMR 5293, IMN, France
LaBRI, CNRS, UMR 5800, France
nicolas.rougier@inria.fr
ORCiD: 0000-0002-6972-589X

I. Motivation

Decision-making is a critical skill for animals and autonomous robots alike. Whether you are a rabbit or a driverless car, you constantly need to make appropriate decisions. This work stresses the importance of taking into account habit formation in decision-making and in goal-directed behaviors such as intrinsic motivation, especially as it pertains to sensorimotor learning.

In computational systems, reinforcement learning (RL) (Sutton, 1998) has been a popular framework to describe and explore the issues of decision-making. Interestingly, although the RL framework was not intended to provide a plausible model of reinforcement learning in animals, RL, and in particular temporal difference (TD), has been a popular choice to model and explain observed experimental data in neuroscience (Pan, 2005); this is in great part due to TD providing a temporal model for the Rescorla and Wagner learning rule (Rescorla, Wagner, et al., 1972). The reward prediction error of TD has been proposed to model the firing activity of dopaminergic neurons located in the basal ganglia, a set of neural structures located in the center of the brain that are critically involved in learning to make appropriate decisions.

An important aspect of decision-making is habit formation. Indeed, an action that has been learned through reinforcement towards a specific rewarded outcome (action-outcome, A-O) can progressively become a habit, especially if elicited by a clear stimulus. In that case a previously goal-directed behavior becomes an automatic response to a stimulus (stimulus-response, S-R), characterized by a relative insensitivity to reward devaluation (Yin and Knowlton, 2006).

Here, we do not use the reinforcement/devaluation protocol. Rather, we put forward the hypothesis that habit formation can lead to suboptimal choices even when rewards remain fixed. For this we use a paradigm commonly used in psychology, behavioral neuroscience, and computational science: a two-armed bandit task. This task is used on a computational model of decision and on a real-world setup with non-human primates.

II. Computational Model

Our lab previously created a neurocomputational model of the basal ganglia (Topalidou et al., 2016, Figure 1). The model is implemented as a recurrent neural network with rate-coded neurons reproducing the main structures and interactions found in the basal ganglia. The network is organized as three interactive loops: motor, associative and cognitive. The cognitive loop perceives the visual stimuli, and the motor loop generates the action that pushes the chosen button, and the associative loop encodes the mapping between stimuli and buttons.

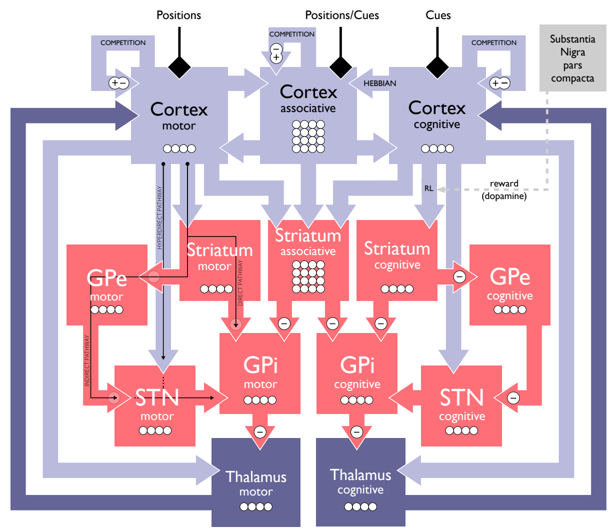

Fig. 1. A schematic representation of the computational model. STN: subthalamic nucleus, GPi/GPe: globus palidus internal/external. For a complete description of the model, see Topalidou et al., 2016. The code source for reproducing results is available at http://dx.doi.org/10.6084/m9.figshare.5203993.

Out of all synaptic connections in the network, only two are plastic. The connection between the cognitive cortex and the cognitive striatum changes its weights according to a reinforcement learning rule, and the one from the cognitive to the associative cortex implements Hebbian learning. While the former is affected by rewards, the latter is not.

We subjected this model to a two-armed bandit task where two visual stimuli, A and B, are presented. Stimuli are rewarded probabilistically, with probability r_A and $1 - r_A$ respectively: if $r_A = 0.8$, A is rewarded 8 out of 10 times, while B only 2 out 10. For the first 20 trials however, we forced the model to choose a stimulus by presenting only one at a time. During this period, A and B are presented with a ratio P_A and $1-P_A$ respectively. For instance, with $P_A = 0.3$, A and B are presented alone, as forced choices, 6 and 14 times respectively in a random order during the first 20 trials. During the rest of the trials, both A and B are presented to the model,

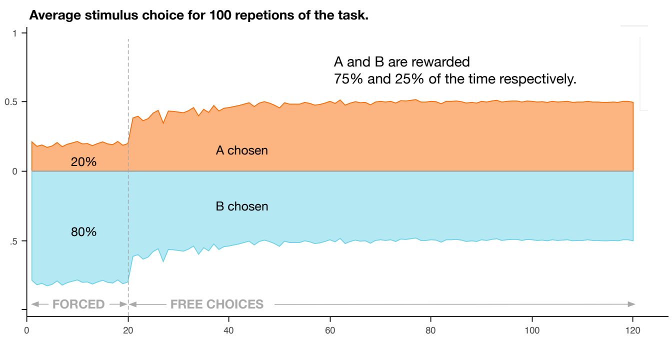

Fig. 2. In the model, choices can go against rewards, if the stimulus with the lower has been sufficiently reinforced enough through Hebbian learning. In this instance, $r_A = 0.75$ and $P_A = 0.2$. The figure shows averages over 100 runs. Despite A being much more rewarded than B, B is initially preferred when free choices are allowed. The choice then reach a balanced equilibrium: around 50 runs have switched to always choosing A (RL was stronger), the other to always choosing B (Hebbian learning was stronger). The behavior of the model is heavily affected by the relative weights of Hebbian learning and RL, as well as the initial RL value for the two stimuli. The parameters were fitted on experimental data from previous monkey experiments.

which is able to choose freely between them. Our hypothesis is that if we force the choice of the less rewarded stimulus B sufficiently more often than A, the model will choose B more than A when able to choose freely.

As shown in Figure 2, the model is indeed able to display such a behavior. The interpretation is that during the forced choice, the Hebbian connection is strongly reinforced towards B. In other words, the model habituates to choosing B: the more a choice is made, the easier it is to make in the future. Under some circumstances (see Figure 3), it allows Hebbian learning to prevail over reinforcement learning. For instance, if B is chosen all the time over a time period, A is not reinforced through RL anymore while Hebbian influence favoring B

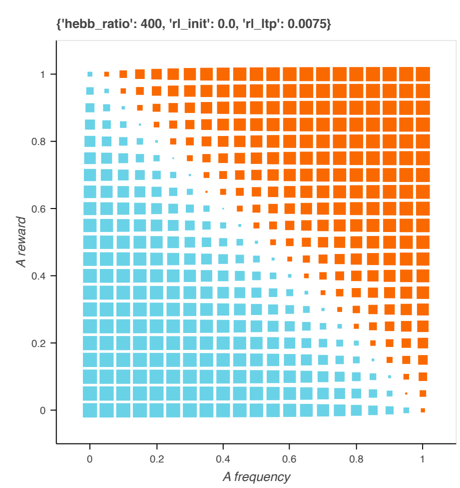

Fig. 3. The frontier between choosing A more than B is not independent of P_A. In this diagram, the area of a square represents the difference between the number of times A versus B has been chosen in the first ten trials of free choices (each time, averaged over 100 repetitions of the task). If the difference is positive (A chosen more than B), the square is orange, else, blue. In a rational agent, the frontier would be horizontal, at $r_A = 0.5$. Here, we can see that P_A can be set to induce suboptimal choices.

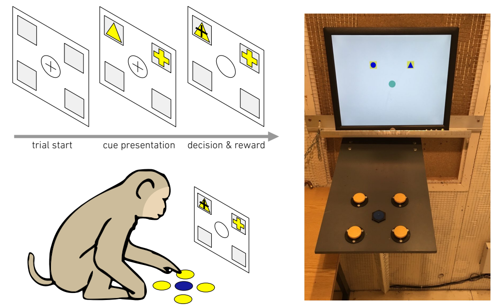

Fig. 4. Three female macaque monkeys were trained on a button/screen setup. The buttons correspond to position on the screen where visual stimuli appear, as shown. The monkeys are sitting in chairs, at 50 cm from the screen. Trials are initiated by holding the central button. After a delay (750-1250 ms), stimuli appear on the screen in two randomly chosen positions out of four. The choice is made by pushing the corresponding button. Reward is water.

grows, making choosing A even less likely in the future. Interestingly, whereas novelty-based intrinsic motivation models predict that (moderately) novel stimuli are more attractive, our model explores an opposite tendency: the decision process favors familiarity of choice.

III. Behavioral Experiments

To test the predictions of the model, we are applying the two-armed bandit protocol to non-human primates (*macaca mulata*, see Figure 4)[1]. Contrary to the model, during the forced choice period, A and B do not appear alone. They appear with a neutral stimulus as an alternative; the neutral stimulus is never rewarded. Additionally, the forced choice period last 50 trials. The experiments are ongoing.

IV. Acknowledgements

We wish to thank Hugues Orignac, Tho-Hai Nguyen, Aurélien Nioche and Brice de la Crompe for their support, advice, and insights.

Source code and data for the figures can be found at: http://dx.doi.org/10.6084/m9.figshare.5203993.

References

Pan, W. -X. (2005). “Dopamine Cells Respond to Predicted Events during Classical Conditioning: Evidence for Eligibility Traces in the Reward-Learning Network”. In: *Journal of Neuroscience* 25.26, pp. 6235–6242. DOI: 10.1523/jneurosci.1478-05.2005.

Rescorla, Robert A, Allan R Wagner, et al. (1972). “A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement”. In: *Classical Conditioning II: Current Research and Theory* 2, pp. 64–99.

Sutton, Richard (1998). *Reinforcement Learning: An Introduction*. Cambridge, Mass: MIT Press. ISBN: 9780262193986.

Topalidou, Meropi et al. (2016). *Dissociation of reinforcement and Hebbian learning induces covert acquisition of value in the basal ganglia*. Tech. rep. DOI: 10.1101/060236.

Yin, Henry H. and Barbara J. Knowlton (2006). “The role of the basal ganglia in habit formation”. In: *Nature Reviews Neuroscience* 7.6, pp. 464–476. DOI: 10.1038/nrn1919.

[1]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).