HAL
open science

Habits That Contradict Rewards

Fabien C y Benureau, Thomas Boraud, Nicolas P Rougier

To cite this version:

Fabien C y Benureau, Thomas Boraud, Nicolas P Rougier. Habits That Contradict Rewards. 7th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics, Sep 2017, Lisbon, Portugal. hal-01535484

HAL Id: hal-01535484

https://inria.hal.science/hal-01535484

Submitted on 9 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

Nicolas P. Rougier
Inria Bordeaux Sud-Ouest, France
Univ. of Bordeaux, UMR 5293, IMN, France
LaBRI, CNRS, UMR 5800, France
nicolas.rougier@inria.fr

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 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. Unlike the cortex which is organized in layers, the basal ganglia is organized in nucleus which interact with each-other (Figure 1). These interactions form functional loops with the cortex, 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 toward 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.

In the following, we consider a two-armed bandit task where two visual stimuli, A and B, are presented. This task is used on a computational model of decision and on a real-world setup with non-human primates. The stimuli appear in two out of

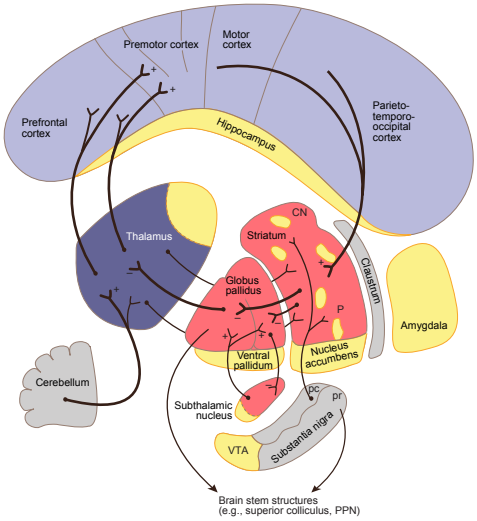

Fig. 1. A schematic representation of the main structures and connections of the basal ganglia in primates, as well as its interaction with the cortex and other structures. The colors match the colors of the model shown in Figure 2, except for yellow (regions involved in limbic-system functions) and grey, both absent in the model. Adapted from Graybiel, 2008.

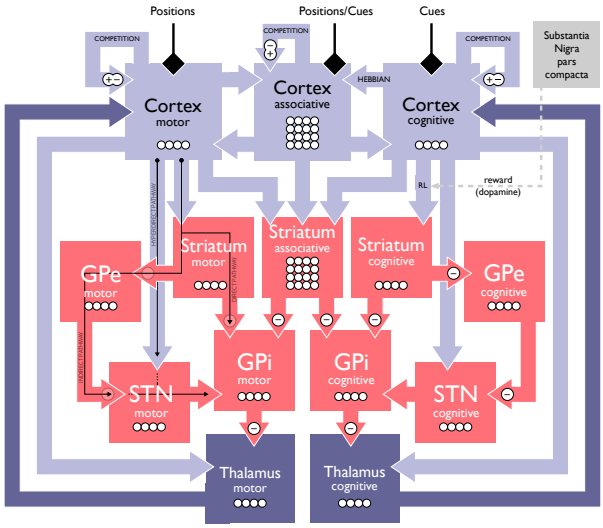

Fig. 2. A schematic representation of the computational model. STN: subthalamic nucleus, GPi/GPe: globus palidus internal/external. For a complete description of the model, see Topalidou et al., 2016. The code source for reproducting results is available at github.com/benureau/basal-ganglia.

four positions, the positions being chosen randomly at each trial. Stimuli are rewarded probabilistically, with probability r_A and $1 - r_A$ respectively. Without loss of generality, we assume $r_A \geq 0.5$: A is more rewarded than B.

II. Computational Model

We previously created a neurocomputational model of the basal ganglia (Topalidou et al., 2016, Figure 2). 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.

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 the two-armed bandit task. For the first 20 trials however, we forced the model to choose a stimuli 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 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 at each trial, which is able to choose freely between them. Our hypothesis if that 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 3, the model is indeed able to display such a behavior. The interpretation of this result is that each time a choice is made, the stimulus chosen is reinforced through Hebbian learning, in the connection from the cognitive to the associative cortex. Being Hebbian, this reinforcement is independent of the presence or absence of reward. Under this mechanism, the more a choice is made, the easier it is to make

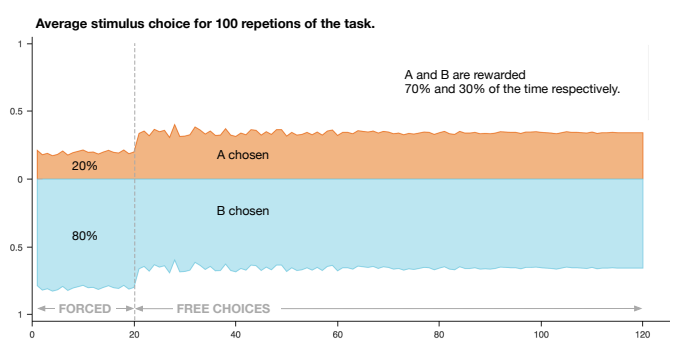

Fig. 3. In the model, choices can go against rewards, if the stimulus with the lower has been sufficiently reinforced enough through Hebbian learning. In this instance, $r_A = 0.7$ and $P_A = 0.2$. The figure shows averages over 100 repetitions.

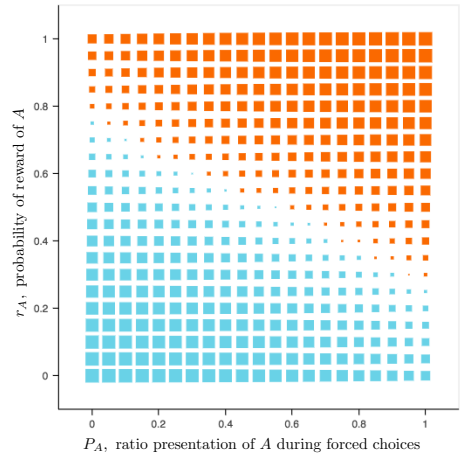

Fig. 4. The frontier between choosing A more than B is not independent of P_A. In this diagram, the area of a square represent the difference between the number of times A versus B has been chosen in the first ten trials of free choices (each time, averaged over 100 repetitions of the task). If the difference 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.

in the future, *regardless of consequences*. When the Hebbian reinforcement does not correspond to the one acquired through RL, there is a competition between the influence of those two learning processes; under some circumstances (Figure 4), the Hebbian influence can prevail. Interestingly, this hypothesis predicts opposite results from many novelty-based intrinsic motivation models: the decision process favors familiarity.

III. Behavioral Experiments

To test the predictions of the model, we are applying the same two-armed bandit protocol to non-human primates (*macaca mulata*). 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, advices, and insights.

References

Graybiel, Ann M. (2008). "Habits, Rituals, and the Evaluative Brain". In: *Annual Review of Neuroscience* 31.1, pp. 359–387. DOI: 10.1146/annurev.neuro.29.051605.112851.

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.