Dopamine enables dynamic regulation of exploration

François Cinotti
Institut des Systèmes Intelligents et de Robotique
Université Pierre et Marie Curie, CNRS
4 place Jussieu, 75005, Paris, FRANCE
francois.cinotti@isir.upmc.fr

Virginie Fresno
Institut de Neurosciences Cognitives et Intégratives d'Aquitaine
Université de Bordeaux, CNRS
146 rue Leo Saignat

Nassim Aklil
Institut des Systèmes Intelligents et de Robotique
nassim.aklil@isir.upmc.fr

Etienne Coutureau
Institut de Neurosciences Cognitives et Intégratives d'Aquitaine
etienne.coutureau@u-bordeaux.fr

Benoît Girard
Institut des Systèmes Intelligents et de Robotique
benoit.girard@isir.upmc.fr

Alain Marchand
Institut de Neurosciences Cognitives et Intégratives d'Aquitaine
alain.marchand@u-bordeaux.fr

Mehdi Khamassi
Institut des Systèmes Intelligents et de Robotique
mehdi.khamassi@isir.upmc.fr

Abstract

We present rat behavioural data in a non-stationary three-armed bandit task where observed long-term improvements in performance and decline in exploration levels suggest that rats are capable of some sort of meta-learning, i.e. the regulation of learning and decision-making parameters underlying behaviour. This initial observation is followed by a proposal for a reinforcement learning model with an added meta-learning mechanism regulating the inverse temperature β of the action selection function. More specifically, this mechanism is designed in such a way that accumulation of positive "reward prediction errors" (RPE) leads to increased exploitation of what is perceived as the best action, whereas a drop in the rate of RPEs entails increased adaptive exploration of potentially better options. This model is capable of reproducing a range of experimental results, which a series of rival models cannot, and allows predictions which could then be verified. In a second part of the experiment, inhibition of dopamine through a systemic injection of D1/D2 receptor antagonist flupenthixol is shown to increase exploration levels without affecting performance and learning quite as strongly, thus supporting the hypothesis that, in addition to the well established role of phasic dopamine in signalling individual RPEs necessary for the updating of action values, dopamine also controls the balance between exploration and exploitation as reported by Humphries et al. (2012). This possibility is mirrored in our model by the average RPE signal used to regulate β, which may be construed as a long-term or tonic component of dopaminergic activity. Indeed, applying a filter to this signal of our model allows us to recapture the data obtained in the various pharmacological conditions.

Keywords: Dopamine, meta-learning, exploration-exploitation trade-off, multi-armed bandits

Acknowledgements

We are deeply indebted to the Agence Nationale de la Recherche for funding the LU2 project from which this study originally stems.

Experiment description

The experiment consisted in a non-stationary 3-armed bandit task (fig. 1) in which rats (n=24) were required to press a lever at each trial in the hope of getting a reward. The best lever was either associated to a very high reward probability of 7/8 compared to the two other levers (1/16 each), or to a reward probability only slightly higher (5/8) than the other levers (3/16 each). These two possible situations are called low uncertainty and high uncertainty situations respectively. Unsignalled changes in the position of the best lever and sometimes in the uncertainty condition occurred after a fixed number of 24 trials defining one block, forcing rats to constantly adapt their behaviour to a changing environment. This first stage of the experiment proceeded for 20 sessions of 6 blocks each laid out pseudo-randomly so that each target $\times$ uncertainty level combination was encountered once in the session, and 4 "double" sessions of 12 blocks each. In a second stage, the same rats were injected systemically with different concentrations of flupenthixol (0, 0.1 mg/kg, 0.2 mg/kg and 0.3 mg/kg) – an antagonist of D1 and D2 dopamine receptors –, on two different sessions, interspersed with sessions in which a saline injection was used.

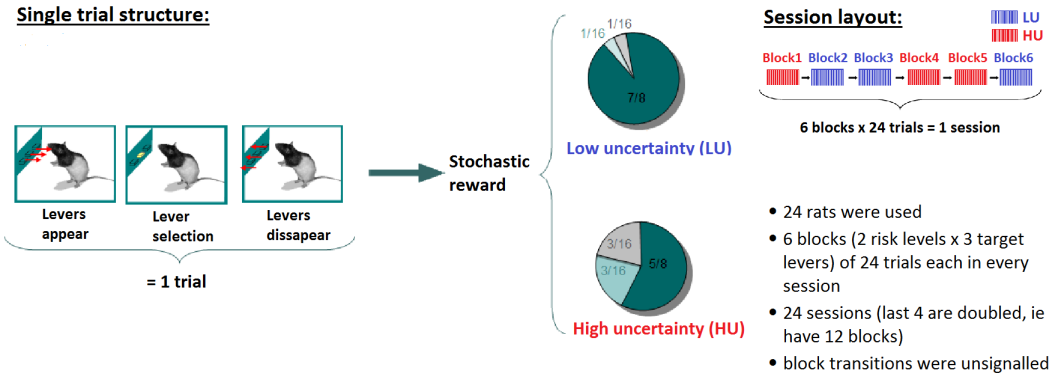

Figure 1: Experimental setup

The long term evolution of the behaviour of rats in the first stage of the experiment

When looking at the proportion of correct choices made throughout blocks of different sessions (fig. 2), we found, as might be expected, that the rats were able to perform the task correctly and that their ability to do so was dependent on the uncertainty level, but also that the final performance level was dependent on which session the blocks belonged to. In the first sessions, correct choices were made on respectively 45% and 40% of the last trials of low and high uncertainty blocks belonging to the first six sessions versus 68% and 48% in blocks belonging to the last six sessions.

Additionally, looking at the evolution of win-shift across sessions, i.e. the proportion of trials in which rats after being rewarded chose to nonetheless shift from this lever on the next trial, a long-term trend – namely a decrease as the experiment proceeds – is also evident. These long-term dynamics might be reflecting an adaptation of learning and decision-making parameters, a process known as meta-learning, which in this case manifests itself most obviously by a modification of the exploration-exploitation trade-off. In the computational paradigm of reinforcement learning (eq. 1 and 2), this lends itself to the interpretation that it is β, the inverse temperature of the decision-making function (eq. 2) that is targeted. Indeed, when β is high, agents exacerbate differences in action values, making them prone to exploitation, whereas a low β, by attenuating these differences, will entail more exploration.

$$Q_{t+1}(a_t) = Q_t(a_t) + \alpha.(r_t - Q_t(a_t)) \quad (1)$$

$$P(a_{t+1} = a_i) = \frac{e^{\beta_{t+1}.Q_t(a_i)}}{\sum_j e^{\beta.Q_t(a_j)}} \quad (2)$$

Proposed model

We designed a computational model of the behaviour of the rats capable of adjusting the β parameter given current performance levels. Indeed, if positive RPEs start accumulating, then the rat has probably hit upon the correct target and should start exploiting it by increasing β; conversely, when positive RPEs become scarcer it seems like a good idea to try something different by reducing β and exploring other options. One way of implementing this intuition is by estimating the current RPE average R_t through a low pass filter of the RPEs δ_t generated at each trial:

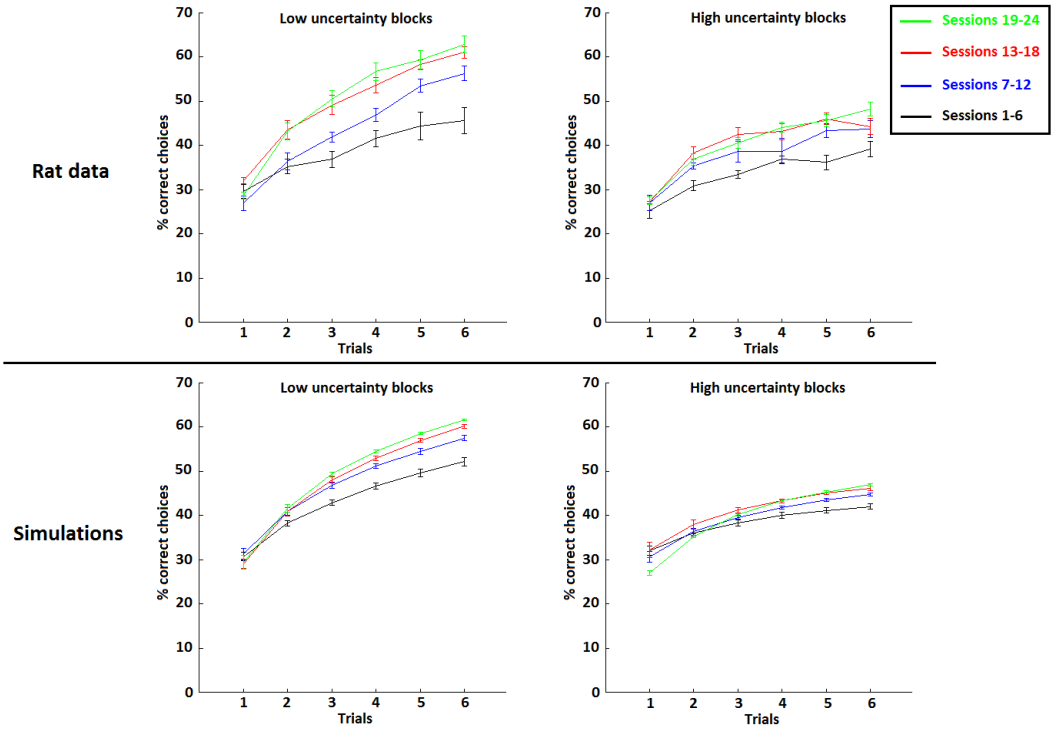

Figure 2: Within-block evolution of performance in low and high uncertainty blocks of rats and of the simulated meta-learning model. Trials were binned into groups of four before averaging the number of correct choice. Sessions were binned by groups of six. A repeated-measures ANOVA with three factors (trial, session and uncertainty) uncovered very highly significant effects ($p < 0.0001$) of all these factors.

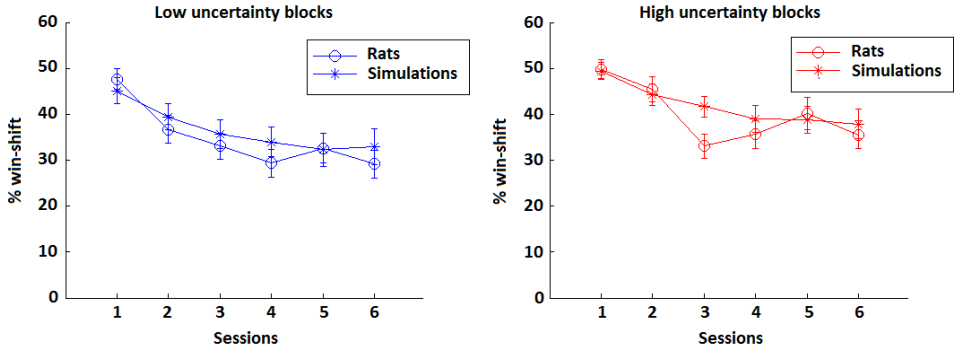

Figure 3: Inter-session evolution of win-shift of rats and of the meta-learning model. Sessions were binned into groups of four. Very highly significant effects ($p < 0.0001$) of uncertainty and session were found using a repeated-measures ANOVA.

$$R_t = R_{t-1} + \alpha_R.(\delta_t - R_{t-1}) \tag{3}$$

The parameter α_R is analogue to the learning rate determining the update of individual action values (eq. 1) a high α_r makes the agents more sensitive to recent outcomes, and a low α_R give agents a more resilient and long-term vision. This average RPE signal is then used to determine the inverse temperature used for the next trial:

$$\beta_{t+1} = \beta_0 + m.R_t \tag{4}$$

The parameter β_0 represents a baseline value of β for the cases where $R_t = 0$, and m determines how sensitive β actually is to the RPE average. The parameters of this model were optimized separately for each individual by maximizing the likelihood of the choices the rats really made using equation 2. Once optimized, this model could be simulated as if it had been confronted to the same sequence of blocks as the original subjects so as to generate its own behaviour. Analysis of these simulations reveal that this model more or less adequately captures the experimental observation (fig. 2 and 3), contrary to multiple other rival models such as simple Q-learning, meta-learning on the learning rate α, selective attraction to uncertain actions and other models (not presented here due to space limitation).

Dopamine inhibits exploration

Given the design of the model, it is likely that dopamine, which has long been associated to reward prediction errors (Schulz et al. (1997) to name just one) is somehow involved in the regulation of the exploration-exploitation trade-off through β. Indeed, previous work by Humphries et al.(2012) has already shown in a computational model of decision-making in the striatum that low levels of tonic dopamine favour exploration and high levels exploitation. This is verified in the results of the second part of the experiment in which we see a relatively modest effect of the injected dose of dopamine antagonist on learning and performance (fig. 4), only visible for the low uncertainty blocks, and a much greater effect on win-shift (fig. 4) which is indeed enhanced as dopamine is more strongly inhibited. These experimental findings are replicated by applying a simple filter f to the calculation of the RPE average in our model:

$$R_t = R_{t-1} + \alpha_R.(f.\delta_t - R_{t-1}) \tag{5}$$

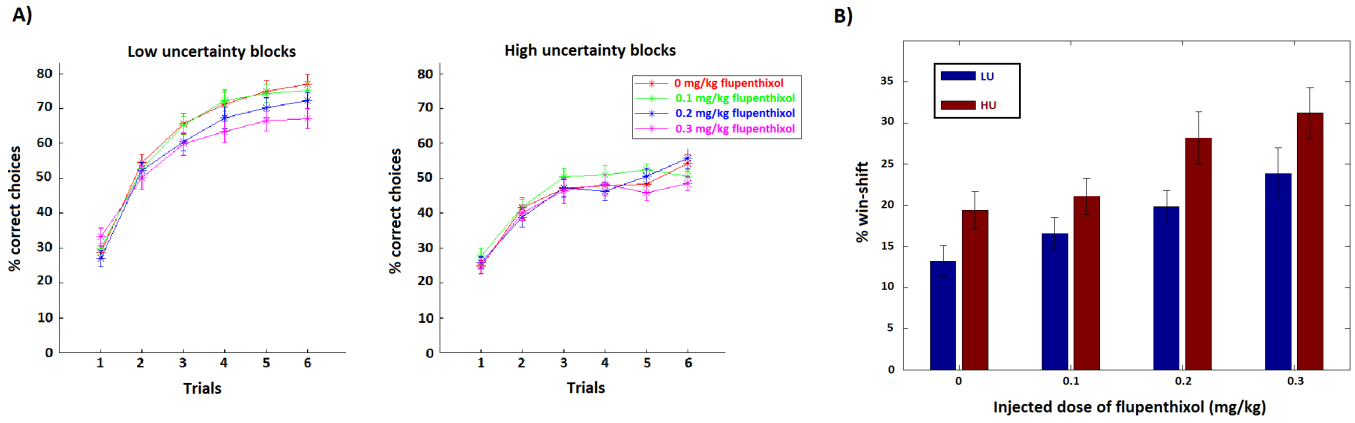

Figure 4: Impact of flupenthixol on the behaviour of rats: A) Within-block evolution of the performance of rats for different doses of flupenthixol. A repeated-measures ANOVA on low uncertainty blocks detected a highly significant effect (p=0.008) of dose and of trials ($p < 0.0001$) on performance, while the same test applied on the high uncertainty blocks detected no effect of dose(p=0.48). B) Win-shift proportions for different uncertainty levels and doses of flupenthixol. Both uncertainty and dose effects are very highly significant ($p < 0.0001$)

Discussion

One of the more surprising results of this study is the relatively modest effect of inhibiting dopamine on performance and learning. Indeed, given the much stronger effect on win-shift, it is very possible that the effect of flupenthixol on performance in low uncertainty blocks is mainly a consequence of this increase in win-shift rather than a deterioration of learning *per se*. This does not necessarily challenge the established view that dopamine supports learning itself, but it certainly points to the fact that there exist other decision-making functions in which it is more crucially involved, namely the ability to "express" learning by selecting what is known to be the best option (Eisenegger et al. 2012). This ability could in an extended sense be included in the multi-faceted concept of motivation because it ensures that actions are not merely known to reap rewards, but are also actually "wanted" (McClure et al. 2003). This would correspond to an "orienting" aspect of motivation, but another aspect, not touched upon here, is the "energizing" aspect, which encompasses such things as the amount of effort an animal is ready to spend to get a reward (Niv et al. 2007) or the speed of action (Shiner et al. 2012). Whether this facet of motivation is also regulated by an average RPE signal remains an open question.

References

- Beeler, J. a, Daw, N., Frazier, C. R. M., & Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. Frontiers in Behavioral Neuroscience, 4(November), 170. http://doi.org/10.3389/fnbeh.2010.00170
- Eisenegger, C., Naef, M., Linssen, A., Clark, L., Gandamaneni, P. K., Mller, U., & Robbins, T. W. (2014). Role of dopamine D2 receptors in human reinforcement learning. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 39(10), 236675. http://doi.org/10.1038/npp.2014.84
- Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. Frontiers in Neuroscience, 6(FEB), 114. http://doi.org/10.3389/fnins.2012.00009
- McClure, S. M., Daw, N. D., & Read Montague, P. (2003). A computational substrate for incentive salience. Trends in Neurosciences, 26(8), 423428. http://doi.org/10.1016/S0166-2236(03)00177-2
- Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. Psychopharmacology, 191(3), 507520. http://doi.org/10.1007/s00213-006-0502-4
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275, 15931599. http://doi.org/10.1126/science.275.5306.1593
- Shiner, T., Seymour, B., Symmonds, M., Dayan, P., Bhatia, K. P., & Dolan, R. J. (2012). The Effect of Motivation on Movement: A Study of Bradykinesia in Parkinsons Disease. PLoS ONE, 7(10), 17. http://doi.org/10.1371/journal.pone.0047138