Modeling the sensory roles of noradrenaline in action selection

Maxime Carrere, Frédéric Alexandre

Modeling the sensory roles of noradrenaline in action selection

Maxime Carrere[1,2,3] and Frédéric Alexandre[2,1,3]

Abstract— **Noradrenaline participates in the neuromodulation of brain activity to modify the trade-off between exploration and exploitation when sensory contingencies have changed. Accordingly, attentional models of noradrenaline acting on sensory representations have been proposed. In this paper, we explore another possible action of this neuromodulator in the decision making process and report simulation results that illustrate that its role is concerned with different aspects of sensory processing. This is made possible by the extension of a classical model of action selection, to render it able to detect and to adapt to sudden changes in sensory contingencies, which is a major characteristic of autonomous learning.**

I. INTRODUCTION

Reinforcement Learning is among the major learning paradigms used to model decision making [1]. Whereas bayesian approaches have been successfully proposed to address noisy environments [2], no fully satisfactory solution has been proposed to date, concerning unstationary and changing environments [3] which are nevertheless classical environments for humans and animals and represent a major paradigm to be addressed in autonomous learning.

Neuromodulation has been presented as a clever way to modulate the main parameters of classical reinforcement learning algorithms [4]. Among the four main ascendant neuromodulators, noradrenaline is proposed in that paper to control the randomness in action selection or, to tell it differently [5], to control the trade-off between exploration and exploitation, depending on the level of uncertainty. The neuronal model presented in [5] consists in two sensory units representing alternative choices, the gain of which can be modulated by noradrenaline. This model is very interesting because it is very simple (two accumulators in mutual inhibition) and also because it has nice mathematical properties [6]. Nevertheless it also has important limitations, the first one about the fact it is limited to binary decisions [7]. Another important limitation is about the fact that this model has been only presented and tested in isolation, focusing on attentional effects of noradrenaline on the sensory cortex. In natural conditions, dealing with evaluation of the level of uncertainty and triggering the exploration of alternative solutions are performed in a complex network, involving not only sensory representation implemented in [5] but also a more systemic network, particularly involving the prefrontal cortex, like it has been observed in experimental studies [8].

[1]LaBRI, Université de Bordeaux, Bordeaux INP, CNRS, UMR 5800, Talence, France

[2]Inria Bordeaux Sud-Ouest, 200 Avenue de la Vieille Tour, 33405 Talence, France

[3]IMN, Université de Bordeaux, CNRS, UMR 5293, Bordeaux, France

Relying on biological data, we present here a wider model of the effects of noradrenaline on brain functioning, extending the regions in consideration to the network between the basal ganglia and the prefrontal cortex, known to be central in the domain of action selection [9]. To better present the main features of this extension, we first present some important characteristics of reinforcement learning and of neuromodulation related to uncertainty, before presenting these new features and their evaluations in simulations involving a systemic model of the cerebral network.

II. REINFORCEMENT LEARNING

Reinforcement Learning (RL, [1]) has ambiguous relations with Cognitive Computational Neuroscience [10]. In the process of modeling the brain and particularly cognitive and behavioral functions, it is often mentioned that RL is a too simple and too schematic process to capture the complexity of the brain. Nevertheless, principles, concepts and algorithms elaborated in RL researches are often evoked because they remain the best way to introduce some brain mechanisms. In this paper, we will continue to conform to this dual relation.

In the process of action selection, state-values and action-values are among the most central concepts in RL and emphasize two critical evaluations: the value of the state of the environment and the value of the action to select, as a function of some kind of cumulative reward. The actor-critic architecture [11], widely used in RL, exploits these complementary concepts and is also reminiscent of underlying behavioral paradigms [12], pavlovian conditioning (learning the value of a stimulus) and operant conditioning (learning in some context the value of the action). The cerebral implementation of this architecture has often been discussed, particularly comparing the architecture of the basal ganglia with the actor module [13], whereas several hypotheses have been proposed concerning the critic, localized in the ventral striatum [13] or in the projections between dopaminergic region VTA and the striatum [9].

The limitation of the analogy between the actor-critic architecture and the corresponding cerebral implementation arises from the fact that the architecture doesn't integrate enough details about the underlying learning mechanisms. The temporal difference learning algorithm [1] used in the actor-critic architecture refers to model-free learning algorithms directly selecting an action with no explicit a priori evaluation. This is indeed the case in habitual behaviors but not in goal-directed learning, another important learning paradigm in operant conditioning corresponding to model-based learning. Both kinds of learning have been localized

in respectively dorsal and medial regions of the striatum [14] but efficient model-based algorithms remain to be proposed. In natural learning, goal-directed learning is very powerful because, in addition to selecting an action, it allows to explicitly evoke the consequences of the action and can explain phenomena like devaluation but can also trigger prospective evaluations [15].

In fact, interestingly, whereas this latter issue is widely discussed about the actor, it is hardly remarked that the same could be said about the critic, also learned in a model-free way whereas data suggest that it can also be sensitive to devaluation [16]. Accordingly, some authors have proposed the same distinction in the ventral striatum, with a region, the core of the nucleus accumbens, standing for the model-free version of the critic, whereas the shell of the nucleus accumbens would correspond to a model-based learning [17]. In behavioral studies, this could correspond to the distinction between, respectively, the preparatory and the consummatory phases of pavlovian conditioning [15], with the consummatory phase automatically triggering actions to consume the outcome, explicitly evoked when the reward is close, whereas the preparatory phase is not specific of the outcome and simply acts to favor the approach when the reward is still distant.

Another striking difference between natural and artificial studies is the fact that artificial studies generally consider one paradigm at a time, whereas they are in fact coexisting. [18] shows for example that goal-directed and habitual responses correspond to different phases of training but that both are available if needed. [19] has shown in rats that the decision of the best paradigm to employ is taken in the medial prefrontal cortex, massively projecting to the striatum. The arbitration between computationally simple but rigid habits and flexible and efficient but costly model-based approaches has also been shown to depend on the level of uncertainty of the environment [20].

III. COPING WITH UNCERTAINTY

Uncertainty in the sensory world is an important dimension to take into account, to adapt to noisy and changing conditions. Neuromodulation has been presented as a major way to adapt neuronal processing and learning to these conditions [21], with acetylcholine signaling expected uncertainty (or stochasticity) and noradrenaline signaling unexpected uncertainty (or changes in sensory criteria asking for the exploration of new sensory rules). Other more detailed models have presented the influence of acetylcholine [22] and noradrenaline (also called norepinephrine, NE) [5] as acting on the sensitivity of neurons in sensory cortical regions, modifying their signal-to-noise ratio. Concerning noradrenaline, other known effects in other regions of the brain have not been investigated in models, and particularly an inhibitory effect in the shell of the striatum [23] that we would like to study in more details here.

In a systemic view of brain functioning, wondering how the kind of (expected or unexpected) uncertainty is estimated and how the more adapted neuromodulator is released accordingly is also of prime importance. It is proposed in [24] that important criteria in that aim are about the contrast between long and short term histories of conflicts of responses and about history of rewards, with all these pieces of information encoded in the medial prefrontal cortex.

This is another good reason to study the consequences of noradrenergic modulation of activity in the shell, in the ventral striatum, which is also known to be a major target of medial prefrontal cortex projections [12].

IV. OUR MODEL

In order to study this effect, we have adapted and extended an existing model of action selection, described in [25]. In short, this model reproduces a classical experiment where monkeys learn to get rewards by manipulating levers, depending on visual cues presented beforehand. This is carried out in a model implementing several loops between the basal ganglia and the cortex, including a cognitive loop (cf the green section in figure 1) featuring afferences from the medial prefrontal cortex (region OFC known to represent the value of the stimuli involved in the task [26]) to the ventral striatum, modulated by a learning driven by reward prediction error, proposed to be implemented by dopaminergic projections as in most biologically inspired models of decision making; see [25] for details.

In addition to stochastic associative rules between cues and rewards studied in [25], we would like to consider here cases where the associative rules have changed and new solutions must be explored. To do so, we have to introduce new mechanisms for detecting the occurrence of this unstationarity and for triggering exploration of new rules, instead of exploiting a no longer valid rule. Reconsidering the implementation of the shell in the ventral striatum was also an opportunity for us to improve the computation of the dopaminergic reward prediction error by taking the values computed in the shell as the prediction to be confronted to the actual reward in the dopaminergic nucleus VTA, as it is proposed in more realistic dual-pathway learning rules [27].

Our model uses the DANA library for neuronal representation and computation [28]. All the code for the model and parameters are open-source and available online at https://github.com/carreremax/basal-ganglia-ne. We only describe and discuss here changes made from the Guthrie model, shown in figure 1. Figure 2 reports new parameters that have been introduced by the extension of the model and existing parameters that have been modified to keep a functioning model, even when the shell is inhibited by NE.

In order to detect an uncertain situation, we compute the difference between reward arrival s_reward in the most recent trials and reward arrival l_reward in a larger number of (less) recent trials.

s_reward is the average reward in the last n_recent_trials trials :

$$s\_reward = (\sum_{k \in n\_recent\_trials} reward_k)/n\_recent\_trials$$

with $reward_k$ the reward received at trial k. Similarly, l_reward is the average reward in the last n_long_trials.

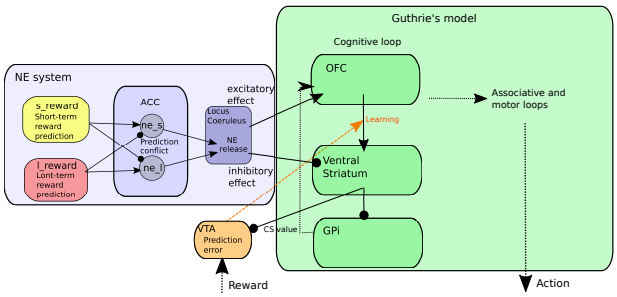

Fig. 1. Main differences between our model and [25]. ST-PRED and LT-PRED are respectively short-term predictor and long-term predictor which predict reward arrival as the average reward from a short and long sequence of trials. Each prediction inhibits the excitatory influence of the other in ACC (modeling a region of the medial prefrontal cortex, known to be involved in conflict monitoring [29]), thus performing an "XOR" function between the two predictions. Only if these predictions are different, NE release will be performed, and will trigger exploration by facilitation of cortical excitation and inhibition of striatal inputs. The value of the stimulus to be considered, CS selected in the ventral striatum, is used to compute the reward prediction error in the dopaminergic nucleus VTA, used to train the connection between OFC and the ventral striatum.

The difference between s_reward and l_reward is computed in two neuronal populations n_l and n_s, as a model of another region of the medial prefrontal cortex, ACC, known to monitor such kinds of conflicts [29] and introduced in the present extension of the model:

$$\frac{dU_{ne\_s}}{dt} = \tau * (-U_{ne\_s} + s\_reward - l\_reward)$$

$$\frac{dU_{ne\_l}}{dt} = \tau * (-U_{ne\_l} + l\_reward - s\_reward)$$

The neuronal population U_{ne_l} computes the above zero value of $|s\_reward|$ minus $|l\_reward|$, while U_{ne_s} computes the other one, $|l\_reward|$ minus $|s\_reward|$. Thus the level of NE release, ne, taken as the sum of ne_s and ne_l activities, is high if s_reward and l_reward are different, when the rule is changing, and correlates with unexpected uncertainty. In case of expected uncertainty, the levels of rewards are the same at short and long terms and the corresponding level of stochasticity can be deduced from comparison to the maximal rewarding situation [24].

Consistent with classical models of NE [5], NE effect at the cortical level is an excitatory gain :

$$\frac{dV_{ctx}}{dt} = f(U_{ctx} * (1 + 1 + ne) * (1 + noise))$$

where V_{ctx} and U_{ctx} are respectively the firing rate and membrane potential of cortical neurons, f and $noise$ respectively the sigmoid function and activation noise used in [25]. Because of the threshold effect of f, already salient stimuli do not have a huge increase in activity, while stimuli with initially low salience may increase a lot, thus facilitating alternative choice and exploration, as proposed in [5].

As an original mechanism introduced in the model, NE has also an inhibitory effect at the striatal level, which impacts the output gain of projection from cortex to shell. This inhibitory effect lessens the impact of learnt weights between cortex and striatum in the decision loop, and thus favors exploration-driven decision.

$$gain = g\_ctx\_cog\_str\_cog * ne\_modulation$$

with $g\_ctx\_cog\_str\_cog$ the constant gain between cortex and striatum in the cognitive loop, and $ne\_modulation$ the modulatory effect of NE.

$$ne\_modulation = max(0.5, 1 - ne\_efficiency * ne)$$

NE modulatory effect is limited to halving excitatory projections from cortex to shell, consistent with the effect of NE observed in [23]. $ne\_efficiency$ is a constant set to 0.8, so that only maximum values of ne will provoke a minimum value of $ne\_modulation$.

The other difference between our model and [25] is the computation of the prediction error. In both models, values of stimuli to be considered, CS, are learnt in the connections between OFC (cognitive cortex in [25]) and the shell of the striatum (cognitive striatum in [25]). Learning depends on a prediction error

$$dw = ERR * V_{ctx} * V_{str} * \alpha$$

where α is the learning rate, V_{ctx} is the presynaptic cortical activity, V_{str} the post-synaptic striatal activity, and ERR the prediction error, which correlates with the influence of dopamine on learning, and is computed as the difference between the current reward and the expected reward. Previously, this expected reward was computed as the sum of all the previous prediction errors (similar to the Rescorla-Wagner rule). Yet this predicted value was not correlated with the network value of the stimulus. For example, if the network is making the good choice because of habitual behavior (resulting from another loop in the Guthrie model), the critic can learn the value, will not make prediction error anymore and will stop the network learning. Yet the cognitive part of the network may not have learnt the current CS value. To address both this problem and the non-neural way the expected reward was computed, we use the value of the CS in the shell as the reward prediction.

$$ERR = reward - \alpha\_str * V_{str}$$

with α_str a scaling constant. Thus, it both prevents the network from stopping learning before having the relevant

Architectural parameters		
Parameter	Meaning	Value
α_LTP	learning rate for long term potentiation	0.0001
α_LTD	learning rate for long term depression	0.00005
$g\_ctx\_cog\_str\_cog$	gain from cognitive cortex to cognitive striatum	1.2
$g\_ctx\_cog\_str\_ass$	gain from cognitive cortex to associative striatum	0.3
g_ne_exc	gain of excitatory projections in NE populations	1.0
g_ne_inh	gain of inhibitory projections in NE populations	-1.0
n_st_trials	Number of trials taken into account for the short-term predictor	3
n_lt_trials	Number of trials taken into account for the long-term predictor	30
α_str	Gain for striatal activity in the prediction error	0.025

Fig. 2. As compared to [25], the values of the first four parameters have been modified to take into account the fact that the activity of the ventral striatum can be inhibited. The other five parametes are new and have been introduced in the extension of the model.

value in the cognitive striatum and is reminiscent of biological data, with the inhibitory projection from the striatum to VTA exploited in dual-pathway dopaminergic models [27].

V. Experiments

To test specifically the network ability to detect changes in reward contingency and switch from exploitation to exploration, we applied it to a classical task in operant conditioning, reversal learning, also known to be a classical paradigm to observe NE release when the rules are reversed [30]. At each trial, 2 stimuli are simultaneously presented to the network for 2500ms, randomly distributed between 2 positions. As soon as the model performs an action, reward is distributed according to the probability of the chosen CS. If no CS is selected after the 2500ms of presentation, the network will not receive any reward. Then we let the neural activities go down to their initial values and the next trial starts. During the first phase, the acquisition phase, for 80 trials, the 2 stimuli are rewarded respectively 100% and 0% of the time. During the second phase, the reversal phase, the same stimuli are presented to the network with their respective reward probabilities switched, so that the network has to detect the change in reward contingency and switch to the newly better rewarded rule.

In fig.3, we report the average performance and decision time on 100 reversal experiments. Each experiment is performed with a "naive" model. The model correctly learns to choose the best rewarded CS during the acquisition phase and switch to the other during the reversal phase. However, NE release allows to perform random exploration immediately after the first trials of reversal, as shown in fig.3.A by a quick increase of the performance to chance level just after reversal, and then to gradually learn from this exploration. In addition NE release also increases the decision time of the model during the first trials of reversal (fig.3.B). This is in accordance with [31] results, showing that animals with NE depletion respond with greater rapidity when perseverating. Fig.3.C shows the release of noradrenaline during trials, which is indeed proportional to unexpected uncertainty, with a peak at the reversal onset.

VI. CONCLUSIONS

Noradrenaline has been presented for a long time as signaling changes on sensory aspects of the sensorimotor associations to be learned by animals to display an adapted behavior [32], whereas more recently, tonic dopamine has been proposed to have the same role when the changes are concerned with motor aspects [33]. It is consequently not surprising that NE effects have been mainly studied in the sensory regions and particularly in the posterior parietal cortex dedicated to selective attention [32]. Nevertheless NE is also reported to modify activity in a specific striatal region, the shell, and we were consequently curious of better understanding its rationale and we have presented the corresponding study in this paper.

First, whereas the striatum is generally described as a motor region, we have explained above that the shell in the ventral striatum is a specific region possibly corresponding to a model-based version of the critic where sensory values are explicitly manipulated. Consequently, considering now that the shell is also the only region in the striatum to receive NE does not challenge the sensory role of this neuromodulator [32]. Second, whereas the classical role of NE is generally excitatory, increasing the gains of sensory neural activities, another specificity of its action in the shell is that it is inhibitory. Our simulations show that, whereas the classical role in sensory regions favors exploration of new sensory alternatives, its role in the shell would be rather to inhibit the previous rule, no longer valid.

In addition to these major functional results, our model also includes an original mechanism to detect unexpected uncertainty, from estimations of recent and long-term rewards. Whereas we propose, consistent to biological data [29], that the integration might take place in ACC, we should investigate more precisely where and how each term is estimated. We think that they may represent two different reward predictions computed in different brains areas, namely the hippocampus from the recent one and the amygdala or prefrontal cortex for the other one, as we will study in forthcoming work, extending again the systemic aspect of our implementation of decision making.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An introduction*. MIT Press, 1998.

[2] P. Dayan and N. D. Daw, "Decision theory, reinforcement learning, and the brain." *Cognitive, affective & behavioral neuroscience*, vol. 8, no. 4, pp. 429–453, Dec. 2008. [Online]. Available: http://dx.doi.org/10.3758/cabn.8.4.429

[3] J. D. Cohen, S. M. McClure, and A. J. Yu, "Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 362, no. 1481, pp. 933–942, May 2007. [Online]. Available: http://dx.doi.org/10.1098/rstb.2007.2098

[4] K. Doya, "Metalearning and neuromodulation," *Neural Networks*, vol. 15, no. 4-6, pp. 495–506, June 2002. [Online]. Available: http://dx.doi.org/10.1016/s0893-6080(02)00044-8

[5] G. Aston-Jones and J. D. Cohen, "An integrative theory of Locus Coeruleus-Norepinephrine function: Adaptive Gain and Optimal Performance," *Annual Review of Neuroscience*, vol. 28, no. 1, pp. 403–450, 2005. [Online]. Available: http://dx.doi.org/10.1146/annurev.neuro.28.061604.135709

[6] E. Brown, J. Gao, P. Holmes, R. Bogacz, M. Gilzenrat, and J. D. Cohen, "Simple Neural Networks that optimize decisions," *Int. J. Bifurcation Chaos*, vol. 15, no. 03, pp. 803–826, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1142/s0218127405012478

[7] A. Pouget, J. M. Beck, W. J. J. Ma, and P. E. Latham, "Probabilistic brains: knowns and unknowns." *Nature neuroscience*, vol. 16, no. 9, pp. 1170–1178, Sept. 2013. [Online]. Available: http://dx.doi.org/10.1038/nn.3495

[8] N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan, "Cortical substrates for exploratory decisions in humans." *Nature*, vol. 441, no. 7095, pp. 876–879, June 2006. [Online]. Available: http://dx.doi.org/10.1038/nature04766

[9] D. Joel, Y. Niv, and E. Ruppin, "Actor-critic models of the basal ganglia: new anatomical and computational perspectives," *Neural Networks*, vol. 15, no. 4-6, pp. 535–547, July 2002. [Online]. Available: http://dx.doi.org/10.1016/s0893-6080(02)00047-3

[10] R. O'Reilly and Y. Munakata, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA, USA: MIT Press, 2000.

[11] Y. Niv, "Reinforcement learning in the brain," *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 139–154, June 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0022249608001181

[12] R. N. Cardinal, J. A. Parkinson, J. Hall, and B. J. Everitt, "Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex," *Neuroscience & Biobehavioral Reviews*, vol. 26, no. 3, pp. 321–352, May 2002. [Online]. Available: http://dx.doi.org/10.1016/s0149-7634(02)00007-6

[13] F. Mannella, K. Gurney, and G. Baldassarre, "The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis." *Frontiers in behavioral neuroscience*, vol. 7, 2013.

[14] B. W. Balleine, M. Liljeholm, and S. B. Ostlund, "The integrative function of the basal ganglia in instrumental conditioning," *Behav Brain Res*, vol. 199, no. 1, pp. 43–52+, 2009.

[15] B. W. Balleine and S. Killcross, "Parallel incentive processing: an integrated view of amygdala function," *Trends Neurosci*, vol. 29, no. 5, pp. 272–279, 2006.

[16] A. M. Bornstein and N. D. Daw, "Multiplicity of control in the basal ganglia: computational roles of striatal subregions." *Current opinion in neurobiology*, vol. 21, no. 3, pp. 374–380, June 2011. [Online]. Available: http://dx.doi.org/10.1016/j.conb.2011.02.009

[17] M. R. Penner and S. J. Y. Mizumori, "Neural systems analysis of decision making during goal-directed navigation." *Progress in neurobiology*, vol. 96, no. 1, pp. 96–135, 2012.

[18] M. G. Packard and B. J. Knowlton, "Learning and memory functions of the basal ganglia," *Annual review of neuroscience*, vol. 25, no. 1, pp. 563–593, 2002.

[19] S. Killcross and E. Coutureau, "Coordination of actions and habits in the medial prefrontal cortex of rats." *Cerebral cortex*, vol. 13, no. 4, pp. 400–408, Apr. 2003. [Online]. Available: http://dx.doi.org/10.1093/cercor/13.4.400

[20] N. D. Daw, Y. Niv, and P. Dayan, "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control." *Nature neuroscience*, vol. 8, no. 12, pp. 1704–1711, Dec. 2005. [Online]. Available: http://dx.doi.org/10.1038/nn1560

[21] A. J. Yu and P. Dayan, "Uncertainty, Neuromodulation, and Attention," *Neuron*, vol. 46(4), 2005.

[22] W. M. Pauli and R. C. O'Reilly, "Attentional control of associative learning–a possible role of the central cholinergic system," *Brain Research*, vol. 1202, pp. 43–53, Apr. 2008.

[23] S. M. Nicola and R. C. Malenka, "Modulation of synaptic transmission by dopamine and norepinephrine in ventral but not dorsal striatum." *Journal of neurophysiology*, vol. 79, no. 4, pp. 1768–1776, Apr. 1998. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/9535946

[24] S. McClure, M. Gilzenrat, and J. Cohen, "An exploration-exploitation model based on norepinepherine and dopamine activity," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. MIT Press, 2006, pp. 867–874. [Online]. Available: http://www.csbmb.princeton.edu/~smcclure/pdf/MGC_NIPS.pdf

[25] 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*, vol. 109, no. 12, pp. 3025–3040, June 2013. [Online]. Available: http://dx.doi.org/10.1152/jn.00026.2013

[26] M. L. Kringelbach, "The human orbitofrontal cortex: linking reward to hedonic experience," *Nat Rev Neurosci*, vol. 6, no. 9, pp. 691–702, 2005.

[27] R. C. O'Reilly, M. J. Frank, T. E. Hazy, and B. Watz, "PVLV: The primary value and learned value Pavlovian learning algorithm," *Behavioral neuroscience*, vol. 121, no. 1, pp. 31–49, Feb. 2007. [Online]. Available: http://dx.doi.org/10.1037/0735-7044.121.1.31

[28] N. P. Rougier and J. Fix, "DANA: Distributed (asynchronous) Numerical and Adaptive modelling framework," *Network: Computation in Neural Systems*, vol. 23, no. 4, pp. 237–253, Dec. 2012.

[29] M. M. Botvinick, J. D. Cohen, and C. S. Carter, "Conflict monitoring and anterior cingulate cortex: an update." *Trends in cognitive sciences*, vol. 8, no. 12, pp. 539–546, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.tics.2004.10.003

[30] G. Aston-Jones, J. Rajkowski, and P. Kubiak, "Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task," *Neuroscience*, vol. 80, no. 3, pp. 697–715, July 1997. [Online]. Available: http://dx.doi.org/10.1016/s0306-4522(97)00060-2

[31] S. T. Mason and S. D. Iversen, "An investigation of the role of cortical and cerebellar noradrenaline in associative motor learning in the rat," *Brain Research*, vol. 134, no. 3, pp. 513–527, Oct. 1977. [Online]. Available: http://dx.doi.org/10.1016/0006-8993(77)90826-5

[32] S. J. Sara and S. Bouret, "Orienting and Reorienting: The Locus Coeruleus Mediates Cognition through Arousal," *Neuron*, vol. 76, no. 1, pp. 130–141, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2012.09.011

[33] M. D. Humphries, M. Khamassi, and K. Gurney, "Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia," *Frontiers in Neuroscience*, vol. 6, no. 9, 2012. [Online]. Available: http://www.frontiersin.org/decision_neuroscience/10.3389/fnins.2012.00009/abstract

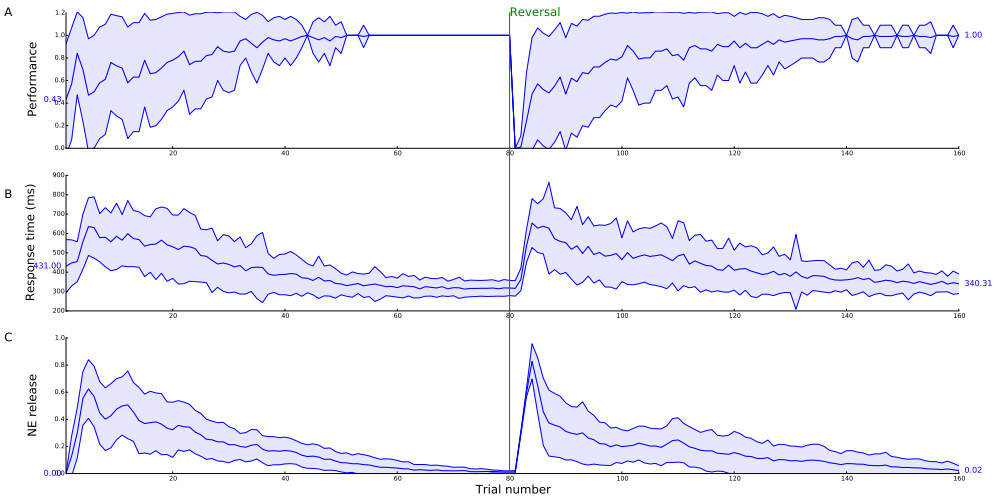

Fig. 3. Reversal experiment for our decision making model. Each curve is the average of 100 experiments performed each time with a "naive" model. Surrounding shaded areas indicate the standard deviation for each curve. (A) Average performance by trials. Network is able to learn reward contingency before and after reversal. At the reversal onset (green line), the NE network detects the change in reward contingency and triggers exploration, which results in a quick increases of the network performance to chance-level, thus exploring the different possibilities and learning the best one. (B) Average convergence time. During the first trials of reversal, exploration by inhibition of the striatum induces a larger response time for NE model. (C) Average release of NE. NE release is important at the beginning of exploration and larger during the first trials of reversal. It correlates with unexpected uncertainty.