Dopamine neurons activity in a multi-choice task: reward prediction error or value function?

Jean Bellot[1,2], Olivier Sigaud[1,2], Matthew R Roesch[3,4], Geoffrey Schoenbaum[5,6], Benoît Girard[1,2], Mehdi Khamassi[1,2] *

1. Institut des Systèmes Intelligents et de Robotique (ISIR), Université Pierre et Marie Curie (UPMC), 4 place Jussieu, 75005 Paris, France; 2. UMR 7222, Centre National de la Recherche, France Scientifique (CNRS), France; 3. Department of Psychology, University of Maryland College Park, College Park, Maryland, USA; 4. Program in Neuroscience and Cognitive Science, University of Maryland College Park, College Park, Maryland, USA; 5. Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, Maryland, USA; 6. National Institute on Drug Abuse Intramural Research Program, Baltimore, Maryland, USA.

Abstract

Traditionally, dopamine neurons are hypothesized to encode a reward prediction error which is used in *temporal difference learning* algorithms. In previous work we studied the ability of the reward prediction error (RPE) calculated by these algorithms to reproduce dopamine activity recorded in a multi-choice task. It reveals an apparent dissociation between the signal encoded by dopamine neurons and behavioral adaption of the animals. Moreover, this activity seems to be only partly consistent with an RPE. In this work we further investigate the nature of the information encoded by dopamine neurons by analyzing the evolution of dopamine neurons activity across learning. Our results indicate that, complementarily to the RPE, the value function fits well with dopamine neurons activity in this task and could contribute to the information conveyed.

*This research is funded by the HABOT project (Emergence(s) Ville de Paris program)

1 Introduction

During the 90's, Schultz and colleagues discovered that the phasic activity of dopamine (DA) neurons shows strong similarity with an RPE (Schultz, Dayan, & Montague, 1997) during pavlovian conditioning. Since then, *Temporal Difference (TD) learning* algorithms (Sutton & Barto, 1998) have been extensively used to explain the role of DA in learning (see (Glimcher, 2011) for review), even if some alternative models exist (O'Reilly, Frank, Hazy, & Watz, 2007). The RPE is used in *TD learning* to update a value function that predicts the sum of future expected reward. These algorithms learn to choose actions that maximize future rewards on the basis of such value function. But the exact nature of the signal encoded by DA neurons is still unclear, especially when the animal needs to choose between different actions. What does the DA signal encode when the animal has to perform a choice?

Several neurophysiological studies (Roesch, Calu, & Schoenbaum, 2007; Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006) investigated this issue by recording DA neurons during multi-choice tasks. In (Morris et al., 2006), the information encoded by DA at the time of the presentation of the different possibilities reflected an RPE dependent on the future choice of the animal. These results are consistent with an RPE calculated by the *SARSA* algorithm. In (Roesch et al., 2007), the animals were trained to choose between two adjacent wells (the right and the left well). During the free choice trials – indicated by a specific odor –, both well led to differently delayed (short and long delay; referred as delay case) or sized rewards (small and big reward; referred as size case). They found that DA neurons reflected an RPE based on the value of the best available option, suggested to be consistent with the Q-learning algorithm.

Based on this discrepancy, in our previous work (Bellot, Sigaud, & Khamassi, 2012) we quantitatively simulated and compared different *TD-learning* algorithms on the task used in (Roesch et al., 2007). We found that the RPE calculated by the algorithms seemed to converge too fast to explain the DA activity recorded in (Roesch et al., 2007) under the constraint of explaining the behavior. This suggested a dissociation between the behavior of the animals and the RPE signal encoded by DA neurons. This highlighted the necessity to precisely analyze DA signals at different timings to be able to relate such information to the observed dynamics of behavioral adaptation.

In this work, we first re-analyze DA neurons activity so as to distinguish

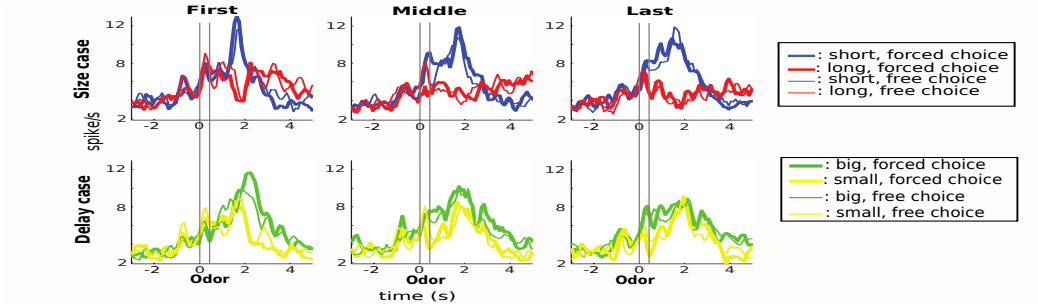

Figure 1: DA activity during the first, middle and last 7 trials extracted from Roesch et al. 2007.

this activity during first, middle and last trials of each learning block (Fig.1). This first shows that the DA signal at the time of the stimulus (*the odor*) now appears qualitatively compatible with SARSA by responding more for the previously best option during first trials (red above blue and yellow above green in fig 1 left), and more for the currently best option during last trials (blue above red and green above yellow in fig 1 right). We then present our simulations of *TD* learning algorithms, trying to fit DA activity during these different temporal windows across learning. We find that none of the tested algorithms can be reasonably ruled out in this case, while DA activity studied here might reflect the combined encoding of both RPE and value signals.

2 Method

To better model the dissociation between the behavior and the calculation of the RPE, we fix the actor by directly matching the behavior of the animals: We use the data from (Roesch et al., 2007) – fig.1(e,f) in the original article – which define the probability that the animals choose the best action at different moments within trial blocks.

To model the task we use a Markov Decision Process (MDP). This allows us to model the different states that the animals experienced, and more specifically the states where DA neuron activity of the animals seems to respond to specific stimuli. These states represent the moment where the animal makes the nosepoke, when the animals perceives the odor - indicating a free-choice trial - and the time of the trial outcome. In order to compare the information carried by DA neurons with the RPE calculated in simulation at these states, we extract the DA activity at these different moments.

We compare three different kinds of critic that were discussed in the original study: *Q-learning*, *SARSA* and *V-learning*. These algorithms update their value function according to the RPE, δ_t: $Q(s_t, a_t) = Q(s_t, a_t) + \alpha\,\delta_t$. The computation of the RPE differs depending on the algorithm:

- *V-learning* : $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$
- *Q-learning* : $\delta_t = r_{t+1} + \gamma \max_a [Q(s_{t+1}, a)] - Q(s_t, a_t)$
- *SARSA* : $\delta_t = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$

These different RPEs predict different responses in a multi-choice task. *V-learning* predicts a response based on an average value of the different options. *SARSA* responds in function of the choosen action and *Q-learning* in function of the value of the best available option.

We then optimize the meta parameters in order to fit DA activity. To do so, we explore the meta parameters space with $\alpha \in [0.01, ..., 0.9]$ and $\gamma \in [0.6, ..., 0.9]$. As the DA activity and the RPE do not share a common scale, we minimize the difference $||(a\delta_s + b) - DA_s||^2$ where DA_s is the experimental DA activity in state s and δ_s is the average RPE computed in s over the different trials. Thus we have: $\delta_s = \frac{1}{n}\Sigma_{e=0}^{n}\delta_s(e)$, where n is the number of considered trials and $\delta_s(e)$ is the RPE computed from the e^{th} trial in s. The (a, b) pair is determined with the least square method, used for the three different set of trials.

3 Results

	RPE Q-learning		*RPE SARSA*		*RPE V-learning*		*Value function*	
	α	error	α	error	α	error	α	error
delay	0.05	16.9	0.075	23.8	0.01	19.2	0.05	5.51
size	0.01	34.7	0.01	35.7	0.01	43.0	0.01	16.26

Table 1: Fitting error when we take into account 21 trials (i.e. 7 for each temporal window).

We can see on Table 1 that the learning rate that optimizes the reproduction of DA activity is very low for the three algorithms. This confirms the hypothesis that the RPE encoded by DA seems to converge slowly. When we compare the ability of the RPE of the algorithms to reproduce DA neurons activity, one can see that even though *Q-learning* is the algorithm that

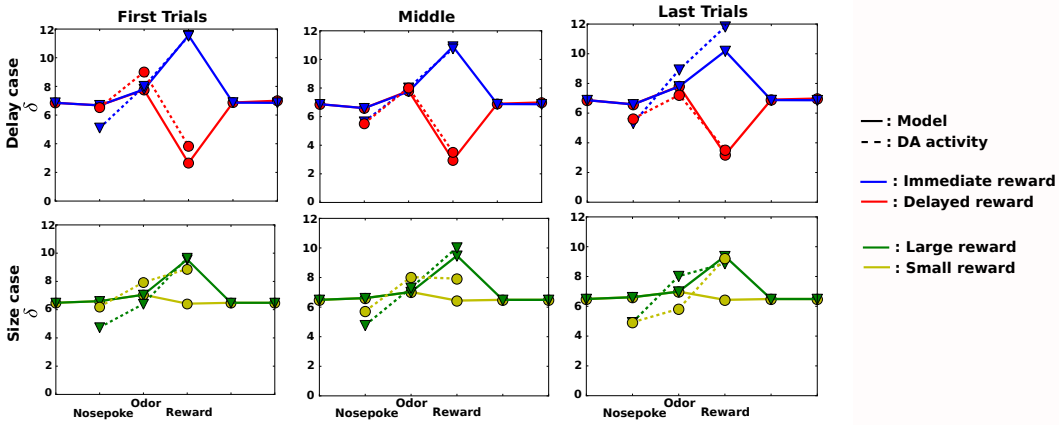

Figure 2: Q-learning reproducing DA activity at the different temporal windows.

quantitatively best reproduces this activity in the delay case (see Table 1, illustrated in Fig.2), *SARSA* and *V-learning* are able to explain this activity better than in previous work (Bellot et al., 2012). Thus, modeling the DA signal at different stages of learning more clearly emphasizes that the difference in the performance of the algorithms on these data is thin. Figure 2 illustrates the fit of Q-learning's RPE on dopamine activity for the delay and size cases. This shows that although Q-learning gets a good quantitative fit, it does not reproduce the above-mentioned qualitative characteristics of DA activity which a priori looked compatible with SARSA.

An important characteristics of DA activity which prevents the algorithms from performing well – especially in the size case – is the persistent response to the reward even during the last trials (Fig. 2), where learning is supposed to have converged. This is different from Schultz' original recordings where DA neurons stopped to respond to expected rewards (Schultz et al., 1997). This could appear more similar to a value function which becomes higher and higher as the agent gets closer to the reward. Thus we also try to fit the value function of the *V-learning* algorithms on DA activity. Interestingly, we find that it gets a smaller fitting error on these data (Table 1).

This work thus partly confirms the results of our previous work by showing that the RPE of *Q-learning* can better explain the activity of DA neurons recorded in (Roesch et al., 2007), comparatively to the RPE of *SARSA* and *V-learning*. But, it also confirms that even though we did not try to fit the

behavior, none of the critics could explain the size case. Then it is reasonable to ask whether this DA activity can really be compared to a pure RPE signal or whether it also incorporates information relative to the value function.

4 Discussion

The starting hypothesis of this work was based on numerous studies that showed that DA neurons encode an RPE signal comparable to the one used in *TD learning* algorithms (Schultz et al., 1997; Bayer & Glimcher, 2005; Flagel et al., 2010). Moreover, (Roesch et al., 2007)'s study presents a number of control analyses showing that the recorded DA activity in the task presented here indeed encodes an RPE. We thus wanted here to precisely characterize the information encoded by such RPE signal and compare it with different TD algorithms.

Our previous work (Bellot et al., 2012) had shown that we cannot reproduce this activity with the same parameters than those that fit the rats' behavior during the task. Hence, in this work we fixed the actor and compared the ability of different critics to reproduce the information encoded by DA neurons. These results confirm that *Q-learning* is best suited to reproduce DA activity. However, we cannot rule out *SARSA* because it is the only algorithm that predicts a response that depends on the future action of the animal. We also show more unexpectedly that a value function encoding the sum of the immediate and future reward can quantitatively better explain the illustrated activity than an RPE. This raises the question whether DA signal only encodes an RPE or whether it also incorporates other kinds of information such as a value function. Further investigation are needed in order to elucidate this issue.

References

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129141.

Bellot, J., Sigaud, O., & Khamassi, M. (2012). Which temporal difference learning algorithm best reproduces dopamine activity in a multi-choice task? In T. Ziemke, C. Balkenius, & J. Hallam (Eds.), *From animals to animats 12* (Vol. 7426, p. 289-298). Springer Berlin / Heidelberg.

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., et al. (2010). A selective role for dopamine in stimulus-reward learning. *Nature*, *469*(7328), 5357.

Glimcher, P. W. (2011, September). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 3*, 15647–54.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nat Neurosci*, *9*(8), 1057–1063.

O'Reilly, R. C., Frank, M. J., Hazy, T. E., & Watz, B. (2007, February). PVLV: the primary value and learned value Pavlovian learning algorithm. *Behavioral neuroscience*, *121*(1), 31–49.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007, December). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci*, *10*(12), 1615–1624.

Schultz, W., Dayan, P., & Montague, P. R. (1997, March). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593 –1599.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. The MIT Press.