Model-based value in midbrain dopamine signals

Marta Blanco-Pozo* (marta.blancopozo@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford, 13 Mansfield Rd
Oxford OX1 3SR, UK

Thomas Akam* (thomas.akam@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford, 13 Mansfield Rd
Oxford OX1 3SR, UK

Timothy E. Behrens (behrens@fmrib.ox.ac.uk)
Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital
Oxford OX3 9DU, UK

Mark E. Walton (mark.walton@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford, 13 Mansfield Rd
Oxford OX1 3SR, UK

* Equal contribution

Abstract

Midbrain dopamine activity is thought to represent reward prediction errors (RPEs) used to update the value of stimuli and/or actions. However, it remains unclear what sources of value information are available to dopamine neurons, and to what extent values derived from internal models inform dopaminergic RPEs. To assess how midbrain dopamine activity is influenced by internal models of task structure, we trained mice in a multi-step probabilistic decision-making task with changing reward contingencies, and performed photometry recordings from dopamine neurons in the ventral tegmental area (VTA) and dopamine axons in the nucleus accumbens (NAc) and dorsomedial striatum (DMS). Our results indicate that dopamine activity in VTA and NAc terminals is influenced by value information derived from models of task structure. By contrast, value information was absent from activity in DMS dopamine axons, which instead is strongly modulated when making choices towards the option contralateral to the recording site.

Keywords: dopamine, reinforcement, model-free, model-based, striatum

Introduction

Changing environments require animals to flexibly adapt their actions to changes in the world's contingencies. Such behavioural flexibility is thought to be aided by rich internal models of the rules and statistical relationships between external events and actions, which allow animals to predict the consequences of action and update these predictions when actual outcomes differ from the predictions (Tolman, 1948; Daw & Dayan, 2014; Doll, Duncan, Simon, Shohamy, & Daw, 2015).

However, a simpler, though less flexible, strategy involves just repeating those actions that were previously rewarded. It just requires storing - or 'catching' - the value of actions and updating this value when it differs from the predicted one using a reward prediction error (RPE). This is what underlies model-free behaviour (Sutton & Barto, 1998).

Classically, activity in dopamine neurons has been reported to reflect a cached value of actions and to convey a signal consistent for a model-free RPE (Schultz, Dayan, & Montague, 1997; Eshel, Tian, Bukwich, & Uchida, 2016), informing and guiding behaviour (Steinberg et al., 2013; Hamid et al., 2016). However, some recent studies have suggested the presence of higher dimensional signals in dopamine activity (Sadacca, Jones, & Schoenbaum, 2016; Takahashi et al., 2017; Engelhard et al., 2019) which can allow for stimulus-stimulus associations to be learned (Sharpe et al., 2017, 2019).

Previous work in humans presented a sequential decision-making task in which model-free and model-based behaviour could be dissociated, the 'two-step' task (Gläscher, Daw, Dayan, & O'Doherty, 2010; Daw, Gershman, Seymour, Dayan, & Dolan, 2011). Using fMRI, Daw et al. (2011) showed that activation in the NAc could not be explained by a pure model-free computation, but instead reflected both model-free and model-based predictions weighted by their influence on choice behaviour. However, it is not possible to directly relate BOLD signal changes to dopamine and so it remains unclear the extent to which dopamine activity is itself directly influenced by such model-based predictions.

To investigate this issue, here we used a version of this task adapted for behaving mice (Akam et al., 2017) and employed fibre photometry to determine whether bulk activity of genetically-defined midbrain dopamine cells can also reflect model-based computations. In addition, given the evidence that the computations supported by dopamine cell firing and release in terminal regions may differ (Berke, 2018), we compared the activity in VTA dopamine cells to that in axons in target regions in the NAc and DMS respectively.

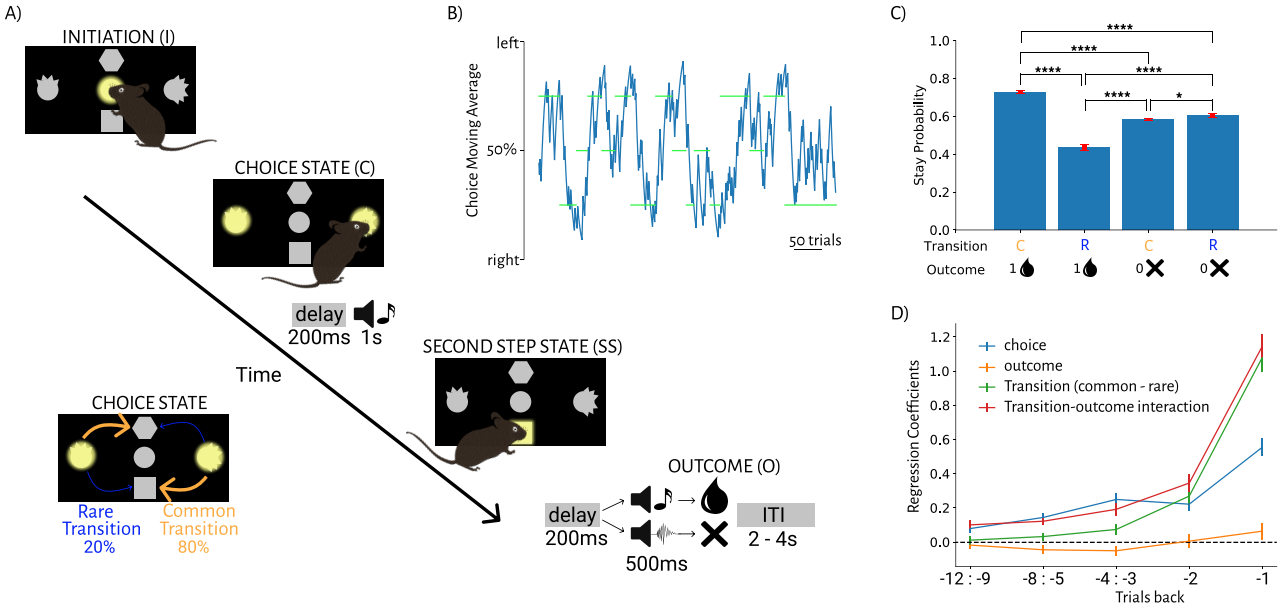

Figure 1: Two-step task. A) Task diagram. B) Example behavioural session. Blue trace shows exponential moving average of choices. Green horizontal bars show reward probability blocks, indicating by their vertical position whether the reward probability was higher for the state commonly reached from the left or right choice, or a neutral block. C) Stay probability analysis showing the probability a choice was repeated as a function of the subsequent state transition (common or rare) and trial outcome (rewarded or not). D) Lagged logistic regression analysis showing how the history of trial outcomes, state transitions, and their interaction affects choice as a function of the number of intervening trials. n = 7 animals, 42943 trials.

Methods

We trained mice on a two-step decision task, adapted from that developed for humans by Daw et al. (2011). The apparatus comprised a central initiation port flanked left and right by 'choice ports' and above and below by 'second-step' ports in which the mice could receive rewards (Fig 1A).

Subjects initiated a trial in the central port then chose between the left and right ports. Each choice port commonly (80% of trials) caused one of the second-step ports to light up, and rarely (20% of trials) caused the other second-step port to light up. Poking the illuminated second-step port delivered reward with probabilities that changed in blocks. In non-neutral blocks, one second-step port had 80% reward probability and the other 20%, while in neutral blocks both seconds-step ports had 50% reward probability. Mice therefore had to learn to choose the choice port that commonly led to the second-step port with high reward probability. Once mice consistently selected the correct choice a block transition was triggered following a random delay and the second step reward contingencies changed.

We recorded bulk calcium activity in midbrain dopamine neurons and their projections to NAc and DMS using fibre photometry. DAT-cre mice were injected bilaterally in VTA with AAV viruses expressing GCaMP6f and TdTomato. Three optic fibres were implanted in each mouse targeting VTA, NAc and DMS.

Results

Behaviour

Subjects learned to track which option was currently best, performing ∼400 trials and >8 reversal blocks in each session (Fig 1B).

Choice behaviour was consistent with a model-based reinforcement learning strategy (Daw et al., 2011), with trial outcome (rewarded or not) and state transition (common or rare) interacting to determine subsequent choice; i.e. subjects tended to repeat choices following rewarded common transitions and non-rewarded rare transitions (Fig 1C). Logistic regression using the trial history to predict choice showed a strong effect of both the transition-outcome interaction and state transition on choices over multiple subsequent trials, but minimal direct influence of the trial outcome (Fig 1D).

Dopamine activity

As expected, calcium activity in VTA and NAc increased at the time of reward, and decreased on reward omission (Fig. 2A). Surprisingly, in DMS the opposite modulation was observed, with lower calcium activity following reward than reward omission.

Dopamine activity in each region was not only modulated at the time of reward delivery, but presented a rich pattern of activity across the different trial stages. In order to disentangle what behavioural variables were driving dopamine activity

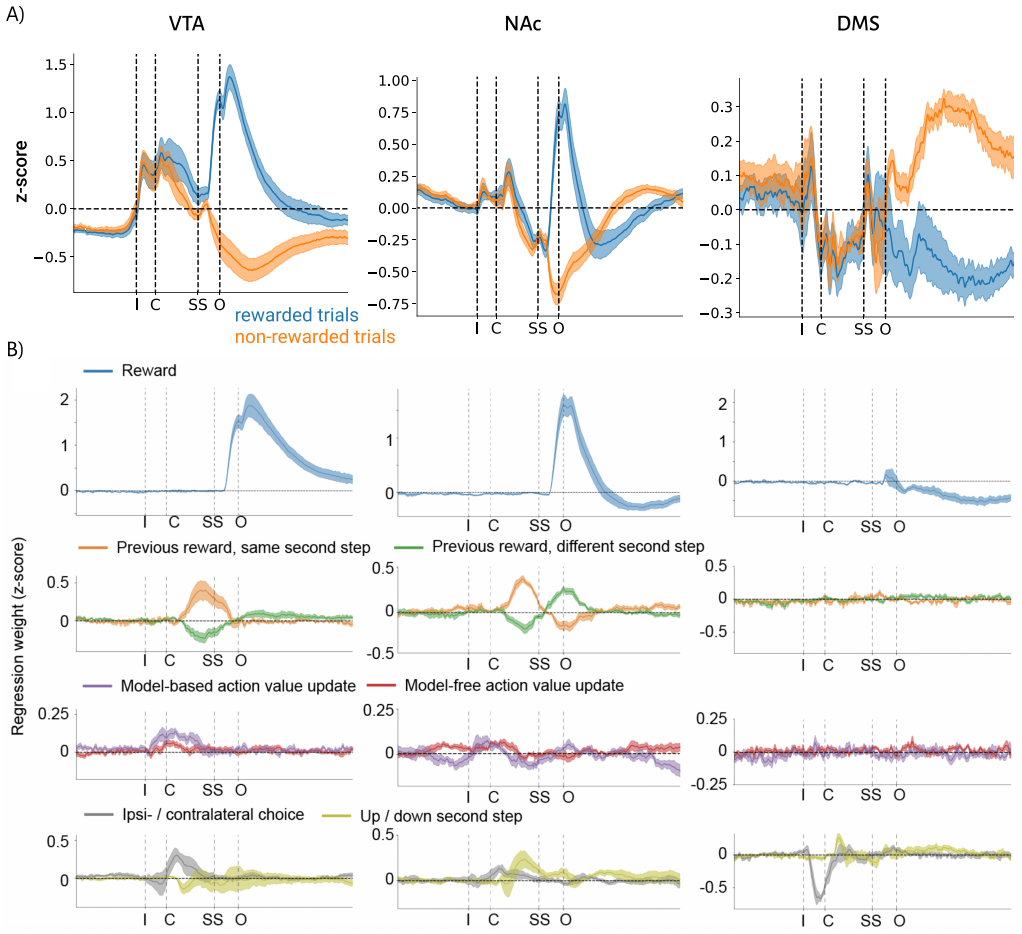

Figure 2: Dopamine photometry. A) Dopamine response on rewarded and non-rewarded trials in VTA, NAc and DMS showing the mean and cross subject standard error. Trials were time-warped to align the initiation poke (I), choice poke (C), second-step poke (SS) and outcome delivery (O). B) Predictor weights in a linear regression analysis predicting trial by trial dopamine activity. Predictors are plotted on separate axes for clarity, but a single regression analysis including all predictors was run for each region. n = 7 animals, VTA:16158 trials; NAc: 15549 trials; DMS: 11236 trials.

at different time points, we used a linear regression analysis predicting trial by trial calcium activity as:

$$y(i,t) = \sum_p \beta_p(t) X_p(i) + c(t) + \varepsilon(i,t) \quad (1)$$

where $y(i,t)$ is the calcium activity on trial i at time-point t, $\beta_p(t)$ is the weight for predictor p at time-point t, $X_p(i)$ is the value of predictor p on trial i, $c(t)$ is the intercept at time-point t, and $\varepsilon(i,t)$ is the residual unexplained variance. Fig. 2B shows the predictor weights $\beta_p(t)$ obtained by fitting the model to activity in each region.

Consistent with the average traces, reward on the current trial strongly increased dopamine activity in VTA and NAc when reward information became available, with a faster timescale in NAc than VTA. Reward had a negative and slower influence on calcium activity in DMS terminals.

We next asked how the previous trial's outcome (rewarded or not) affected dopamine activity as a function of whether the second-step reached on the current trial was the same or different to the previous trial. When the second-step state was the same, reward on the previous trial increased dopamine activity in VTA and NAc at the time when the second-step state was revealed, consistent with an RPE driven by the value of the second-step state. In NAc but not VTA this influence reversed at outcome time, consistent with the second-step state value's influence on the outcome-time RPE. However, crucially, if the second-step state was different from the previous trial, the previous reward had the opposite effect, reducing dopamine activity when the second step state was revealed. This is consistent with subjects understanding the negative correlation between the reward probabilities and inferring that reward in one second-step state reduces the likelihood that the other state has high reward probability, i.e. that mice were inferring a single latent variable about the state of the reward

probabilities rather than independent values for each second-step state. No modulation by previous reward was observed in DMS.

We also constructed predictors which coded the direction in which a model-based and a model-free value update on the previous trial would affect the value of the action chosen on the current trial. Dopamine activity in VTA, though less clearly in NAc, was increased at choice time when the model-based value update was positive, consistent with model-based action value estimates contributing to an RPE once the choice is made. The direction of model-free action value updates also influenced VTA activity weakly at the same time-point. These action value update predictor loadings in NAc showed a complex temporal pattern which it is unclear how to interpret. Neither of these predictors explained activity in DMS terminals.

Finally, we looked at how direction of movement influenced population activity during the trial. Activity in all three areas showed some modulation by whether the choice required an ipsi- or contralateral movement relative to the recording site. This was particularly striking in DMS terminals, which showed a strong increase in activity between trial initiation and choice when mice chose the contralateral poke, suggesting that activity in DMS encoded initial action choice in a lateralised way, independently of action or state values.

Conclusion

We have presented data from dopamine population recordings during a multi-step probabilistic reversal learning task in mice. Mice were able to track the best option across reversals, exhibiting choice behaviour consistent with model-based reinforcement learning. Photometry recordings from midbrain dopamine neurons and projections to NAc showed evidence of value information which respected the task structure, including the anti-correlated nature of the reward probabilities, and transition structure linking actions and states. By contrast, dopamine in DMS was primarily influenced by the direction of the animals' initial chosen action. Together, this demonstrates that dopamine contains multiple representations beyond model-free RPEs.

Acknowledgments

We thank the Wellcome Trust for funding this study.

References

Akam, T., Rodrigues-Vaz, I., Zhang, X., Pereira, M., Oliveira, R., Dayan, P., & Costa, R. M. (2017). Single-trial inhibition of anterior cingulate disrupts model-based reinforcement learning in a two-step decision task. *bioRxiv*, 126292.

Berke, J. D. (2018). What does dopamine mean? *Nature neuroscience*, 1.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130478.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature neuroscience*, *18*(5), 767.

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H. J., Ornelas, S., ... Witten, I. B. (2019). Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons. *Nature*, 1.

Eshel, N., Tian, J., Bukwich, M., & Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nature neuroscience*, *19*(3), 479.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., ... Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature neuroscience*, *19*(1), 117.

Sadacca, B. F., Jones, J. L., & Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *Elife*, *5*, e13665.

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

Sharpe, M. J., Batchelor, H. M., Mueller, L. E., Chang, C. Y., Maes, E. J., Niv, Y., & Schoenbaum, G. (2019). Dopamine transients delivered in learning contexts do not act as model-free prediction errors. *bioRxiv*, 574541.

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., ... Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, *20*(5), 735.

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience*, *16*(7), 966.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press Cambridge.

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron*, *95*(6), 1395–1405.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.