Uncertainty or Prediction Error in Dopamine Ramps

Yael Niv
Interdisciplinary Center for Neural Computation
The Hebrew University
Jerusalem, 91904, Israel
yaelniv@alice.nc.huji.ac.il

Michael Duff Peter Dayan
Gatsby Computational Neuroscience Unit
17 Queen Square
London WC1N 3AR. England
{duff,dayan}@gatsby.ucl.ac.uk

Abstract

Substantial evidence suggests that phasic activities of midbrain dopaminergic neurons represent reward prediction errors. This interpretation has been challenged by recent experimental results, that show ramps in dopamine activity which are related to reward uncertainty. We show that these ramps result from reward prediction errors alone, given differential scaling of positive and negative errors, thus rendering uncertainty moot.

A seminal series of experiments by Wolfram Schultz and his colleagues[1] has persuasively suggested that the phasic activities of midbrain dopamine neurons represent a temporal difference (TD) error (written $\delta(t)$) in the predictions of future reward.[2–5] This prediction error signal occurs at the time of the delivery of unpredicted rewards, or stimuli predicting future rewards, and can be used to guide prediction learning. TD offers a computationally compelling account of a role for DA in appetitive classical and instrumental conditioning, and a precise and parsimonious computational theory of the generation of DA £ring patterns.

Figure 1a shows data from a recent, fascinating, experiment by Fiorillo, Tobler & Schultz (FTS)[6] that presents a crucial challenge to TD theory of DA. They studied the consequences of presenting stochastic rewards, and argued that DA activity explicitly represents uncertainty. In the task, presentation of visual stimuli to macaques was associated with the delayed, probabilistic delivery of rewards (drops of juice). Five different stimuli were associated with £ve different reward probabilities ($p_r = 0, 0.25, 0.5, 0.75, 1$). The traces in £gure 1a show the activity of DA cells averaged over trials for each p_r, in well-trained monkeys. The £ring patterns in £gure 1a can be separated into three main components, one consistent with the TD account and two at apparent variance with it. The £rst component is the sharp peak just after the time of the predictive stimulus. TD predicts that this response should scale with the probability of reward, and FTS' is the £rst report that this happens.

The second component is the *ramp* in the responses towards the time of the reward, which is largest for $p=0.5$. FTS suggested that the ramp represents the *uncertainty* of the delivery of reward, *instead of* a prediction error. The ramp is problematic for the TD account of DA activity, because there is no apparent reason for its occurrence (as there is no prediction error in the inter-stimulus interval). Furthermore, since TD learning operates by arranging for DA activity at one time in

a trial to be *predicted away* by cues available earlier in that trial, it is not clear how a seemingly predictable ramp in activity could persist without being predicted by the same preceding stimuli which predict away the average activity at the time of the reward.

The third component in £gure 1a is the activity just after the time of the delivery or non-delivery of the reward. For the conventional TD rule, the prediction error at this time should be the difference between the actual reward and the reward that is expected based on the stimulus presented. For $p_r \neq \{0, 1\}$, this should be positive for trials on which a reward is delivered, and negative for those on which it is not (see £gure 1c). Crucially, under TD, the average of these differences, weighted by their probabilities of occurring, should be 0. The data clearly show positive activity on average.

This last component points to a TD account of the ramp. A key issue is that the low baseline rate of activity of DA neurons constrains the coding of $\delta(t)$ such that positive and negative values are *represented* respectively by £ring rates of $\sim 270\%$ *above* baseline, but only $\sim 55\%$ *below* baseline.[6] We modelled this by scaling negative values of $\delta(t)$ by a factor of $d = 1/6$ (see caption) prior to summation of the simulated PSTHs. Down-scaling negative $\delta(t)$ will clearly make the average £ring rate at the time of the reward positive, as in £gure 1a. However, £gures 1b and 1d show that when using the simple tapped-delay-line representation of time between the stimulus and the reward commonly adopted in TD models,[4,5] together with a £xed learning rate, a ramp in the activity emerges just as in the experimental data. Since the task involves inherently unpredictable rewards, non-zero prediction errors $\delta(t)$ still occur at the time of reward delivery or non-delivery, even after substantial training. The ramp is due to these prediction errors propagating backward asymmetrically toward the predictive stimulus, as TD learning continues.[7]

Analytically deriving the average response at the time of the reward in trial T from the TD learning rule, we get:

$$< \delta[T] >= p_r - (1 - (1 - \alpha)^{T-1})(p_r^2 + dp_r(1 - p_r)) \xrightarrow[T \to \infty]{} p_r(1 - p_r)(1 - d) \qquad (1)$$

where d is the scaling factor for negative errors. This response is proportional to the variance of the rewards, and so, in keeping with the data, is maximal at $p_r = 0.5$. Though the ramps are indeed related to uncertainty in FTS' setting, this may not be true more generally, and, in any case, occurs *because of*, rather than *instead of*, their coding of $\delta(t)$. There is, however, a key difference between the uncertainty and TD accounts of the ramps. According to the former, ramps are within-trial phenomena, coding uncertainty; by contrast, the latter suggests they arise only through averaging across multiple trials. Under the TD account, the non-stationarity engendered by constant learning from errors makes the PSTH traces potentially misleading, as averaging proceeds over different trial histories.

FTS,[6] as well as Morris, Arkadir, Nevet, Vaadia and Bergman (personal communication), also tried trace conditioning with uncertain rewards, in which the stimulus is not present throughout the delay and so cannot directly constrain the time of the reward. The positive response at the time of reward was comparable to that in delay conditioning; however the ramping activity was found to be reduced or absent, although a similar uncertainty in rewards exists. The TD model of DA readily explains these data by noting that the breadth of the ramp is determined by the learning rate α (£gure 1e). Trace conditioning is notoriously slow, suggesting a low learning rate, and thus a lower ramp.

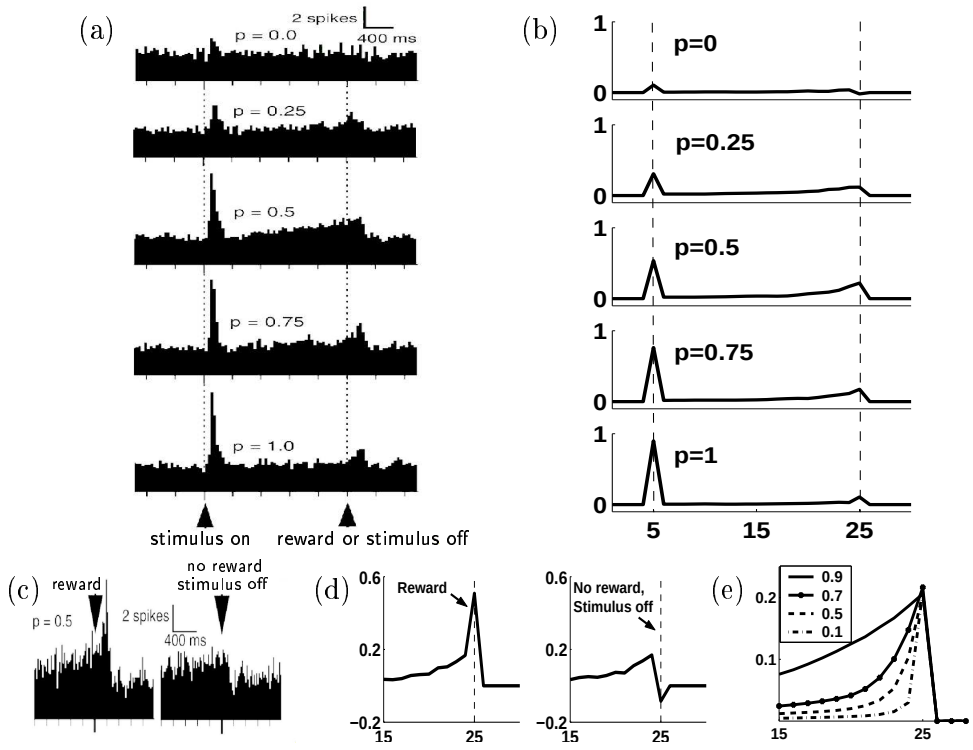

Figure 1: (a) DA response in trials with different reward probabilities, reproduced from Fiorillo *et al.*[6] Population peri-stimulus time histograms (PSTHs) show the summed spiking activity of several DA neurons over many trials, for each p_r (pooling over rewarded and unrewarded cases). Activities are shown time-locked to the stimulus and reward. The predictive stimulus was present throughout the 2-second delay until the delivery (or non-delivery) of reward. (b) TD prediction error with asymmetric scaling. In the simulated task, one of £ve stimuli was randomly chosen and displayed at time $t = 5$. A reward was then given at $t = 25$ with a probability of p_r speci£ed by the stimulus, and the trial ended at $t = 30$. A different set of neurons was used to represent each stimulus across time, using a tapped-delay-line representation. The TD error was $\delta(t) = r(t) + \mathbf{w}(t-1) \cdot \mathbf{x}(t) - \mathbf{w}(t-1) \cdot \mathbf{x}(t-1)$, where $r(t)$ is the reward at time t, and $\mathbf{x}(t)$ and $\mathbf{w}(t)$ are the state and weight vectors for the neurons at this time. The neuron's weights were learned via the standard online TD learning rule with a £xed learning rate $\mathbf{w}(t) = \mathbf{w}(t-1) + \alpha\delta(t)\mathbf{x}(t-1)$, so each weight represented an expected future reward value. With FTS, we depict the prediction error $\delta(t)$ over many trials, after the task has been learned. To account for asymmetric £ring rates about the base rate, negative values of $\delta(t)$ have been scaled by $1/6$ prior to summation of the simulated PSTH, though learning proceeds normally. Finally, to account for the small positive responses at the time of the stimulus for $p_r = 0$ and at the time of the (predicted) reward for $p_r = 1$ seen in (a), we assumed throughout the simulation a small (8%) chance that a predictive stimulus is misidenti£ed as a randomly chosen alternative stimulus. (c) DA response in $p_r = 0.5$ trials, separated into rewarded (left) and unrewarded (right) trials. (d) TD Model of (c). (e) Ramping is ordered by learning rate.

In sum, we have shown that the ramping effect is a straightforward result of TD learning of uncertain rewards, given a neural substrate with an asymmetric representation of positive and negative prediction errors. TD also accounts well for the other aspects of the activity evident in FTS' data. Most importantly, our analysis suggests that uncertainty is playing no explicit part in determining DA activity. Of course, we are not claiming that the ramp cannot have downstream in¤uences, let alone that animals do not learn about and represent uncertainty. Indeed, there is substantial evidence for the sophisticated processing of different aspects of uncertainty by other neuromodulators.[8]

Acknowledgements

We are very grateful to Hagai Bergman, Nathaniel Daw, Daphna Joel, Christopher Fiorillo, Genela Morris, Wolfram Schultz, Peter Shizgal and Philippe Tobler for bene£cial discussions. This work was funded by the EC Thematic Network (YN) and the Gatsby Charitable Foundation.

References

[1] Schultz W. Predictive reward signal of dopamine neurons. *J. Neurophys.*, 80:1–27, 1998.

[2] Sutton RS. Learning to predict by the method of temporal difference. *Machine Learning*, 3:9–44, 1988.

[3] Sutton RS and Barto AG. Reinforcement learning: An introduction. 1998. MIT Press.

[4] Montague PR, Dayan P, and Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, 16:1936–1947, 1996.

[5] Schultz W, Dayan P, and Montague PR. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997.

[6] Fiorillo CD, Tobler PN, and Schultz W. Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science*, 299(5614):1898–1902, 2003.

[7] Kaye, H and Pearce JM The strength of the orienting response during Pavlovian conditioning. *J. Exp. Psychol. Anim. Behav. Process.*, 10:90-109, 1984.

[8] Dayan P, and Yu AJ. Ach, uncertainty, and cortical inference. In TG Dietterich, S Becker, and Z Ghahramani, eds., *Advances in NIPS 14*, 189–196, Cambridge, MA, 2002. MIT Press.