# Dopamine, Uncertainty and TD Learning

Yael Niv [a] Michael Duff [b] and Peter Dayan [b]

[a] *ICNC, Hebrew University, Jerusalem – yaelniv@alice.nc.huji.ac.il*

[b] *GCNU, University College London, London – duff,dayan@gatsby.ucl.ac.uk*

**Abstract**

Substantial evidence suggests that phasic activities of dopaminergic neurons in the primate midbrain represent a temporal difference error in predictions of future reward. Recently, recordings in a task involving uncertain rewards have been interpreted as contradicting this interpretation, suggesting that dopamine activity represents uncertainty. We reinterpret these data in terms commensurate with the prediction error hypothesis of dopamine, and further study the effects of temporal uncertainty on prediction-error learning. Although robust under representational noise, prediction-error learning is shown to be very sensitive to timing noise, a finding yet to be reconciled with the neurophysiological data.

A large body of physiological, imaging, and psychopharmacological data regarding the phasic activity of dopamine cells (DA) in the midbrains of monkeys, rats and humans in classical and instrumental conditioning tasks involving predictions of future rewards[7,9] suggests[6,10] that DA neurons represent temporal difference (TD) errors in the predictions of future reward.[11] These errors arise with unpredicted rewards, or stimuli predicting future rewards, and can guide prediction learning. Recent recordings from Fiorillo, Tobler & Schultz (FTS)[3] in a task investigating the effects of uncertainty in reward delivery on the DA signal, pose a crucial challenge to the TD theory of DA. Here we suggest a theoretical interpretation that maintains the integrity of the TD theory, and enriches it in light of the data. We also study the effects of representational uncertainty on TD learning.

## 1 Uncertainty in Reward Occurrence: DA Ramping

FTS[3] associated the presentation of five different visual stimuli to macaques with the delayed, *probabilistic* ($p_r = 0, 0.25, 0.5, 0.75, 1$) delivery of juice rewards. After training, the monkeys' behavioral output indicated their knowledge of the different contingencies. Fig. 1a shows population histograms of DA cell activity for each $p_r$. The decreasing responses to the visual stimuli for decreasing $p_r$ are consistent with the TD model. However, the model seems to offer no account of the apparent *ramping* of the responses towards the time of the reward. Since the ramp is greatest for $p_r = 0.5$, FTS suggest that it reports the *uncertainty* in reward delivery, *rather than* a prediction error, and speculate that this could explain the apparently appetitive properties of uncertainty (as seen in gambling).

This poses a critical challenge to TD. First, absent prediction error in the interstimulus interval, there is no apparent reason for the activity during this period – as a predictable signal, TD learning would cause it to be *predicted away* by the earlier stimulus. Second, a key aspect of TD[1] is that action choice is influenced by the activity of the DA cells at the time of early predictors. The ramping activity of the DA cells is like a constantly surprising reward that is never predicted by earlier cues, and therefore, cannot influence actions such as the decision to gamble.

A further anomalous facet of the recorded DA activity points to the resolution. According to TD, the prediction error just after the delivery or non-delivery of reward should be the difference between the actual reward and the mean reward expected from the conditioned stimulus. For $p_r \neq \{0, 1\}$, this should be positive for trials on which a reward is delivered, negative for those on which it is not (see fig. 1c), and zero when averaged over trials. However, FTS' averaged data clearly show a positive response at the time of reward for intermediate probabilities. We suggest that this results from the low baseline rate of activity of DA neurons (2-4Hz), which constrains the coding of the prediction error such that positive and negative values are *represented* respectively by firing rates of $\sim\ 270\%$ *above* baseline, but only $\sim 55\%$ *below* baseline.[3] We modeled this constraint by scaling negative values of the prediction error $\delta(t)$ by a factor of $d = 1/6$ (see caption) prior to summation of the simulated PSTHs. Down-scaling negative $\delta(t)$ makes the average firing rate at the time of the reward positive, however, figures 1b;d also show that when using the simple tapped-delay-line representation of time between the stimulus and the reward, together with a fixed learning rate, this also results in the emergence of a ramp in the PSTH, comparable to that in the experimental data. The inherent stochasticity in reward delivery implies that non-zero prediction errors persist at the time of expected reward even after substantial training. The ramp is due to these prediction errors propagating backward asymmetrically toward the predictive stimulus, as TD learning continues.

Analytically deriving the average response at the time of the reward in trial $T$ from the TD learning rule, we get $< \delta[T] >= p_r - (1 - (1 - \alpha)^{T-1})(p_r^2 + dp_r(1 - p_r)) \xrightarrow[T\to\infty]{} p_r(1 - p_r)(1 - d)$ where $d$ scales negative errors. This is proportional to the variance of the rewards, and thus maximal at $p_r = 0.5$. However, although the ramps are indeed related to uncertainty in this particular setting, this is *because of*, rather than *instead of*, their coding of $\delta(t)$, and is *not* more generally true. Uncertainty and TD theories can be experimentally distinguished according to whether the ramps are within- (uncertainty) or across- (TD) trial phenomena. Under TD,

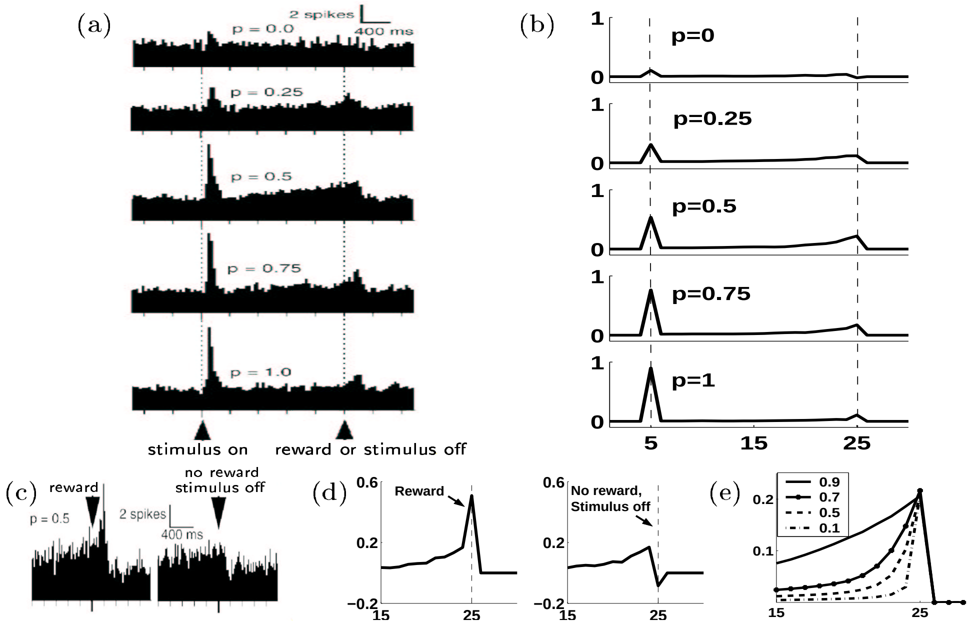

Fig. 1. (a) DA response in trials with different reward probabilities, reproduced from Fiorillo *et al.*[3] Population peri-stimulus time histograms (PSTHs) show the summed spiking activity of several DA neurons over many trials, for each $p_r$, pooling over rewarded and unrewarded trials at intermediate probabilities. 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, in each trial one of five stimuli was randomly chosen and displayed at time $t = 5$. A reward was then given at $t = 25$ with a probability of $p_r$ specified by the stimulus. We used a tapped-delay-line representation of the stimuli, with a different set of neurons representing each of the stimuli across time. 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 $r(t)$ was the reward at time $t$, and $\mathbf{x}(t)$ and $\mathbf{w}(t)$ were the state and weight vectors for the neurons at this time. The neurons' weights were learned via the standard online TD learning rule with a fixed 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 firing rates about the base rate, negative values of $\delta(t)$ have been scaled by $1/6$ prior to summation of the simulated PSTH, although 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 a small ($8\%$) chance that a predictive stimulus is misidentified. (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.

the non-stationarity engendered by constant learning from errors (over different trial histories) renders the PSTH traces potentially misleading, as the ramps appear only as across-trial phenomena. Thus TD can be distinguished experimentally from the uncertainty account as the latter argues for a within-trial phenomenon.

FTS,[3] as well as Morris, Arkadir, Nevet, Vaadia and Bergman (personal communication), also examined the effect of uncertain rewards on DA in "trace conditioning" experiments, in which the predictive stimulus is not present throughout the

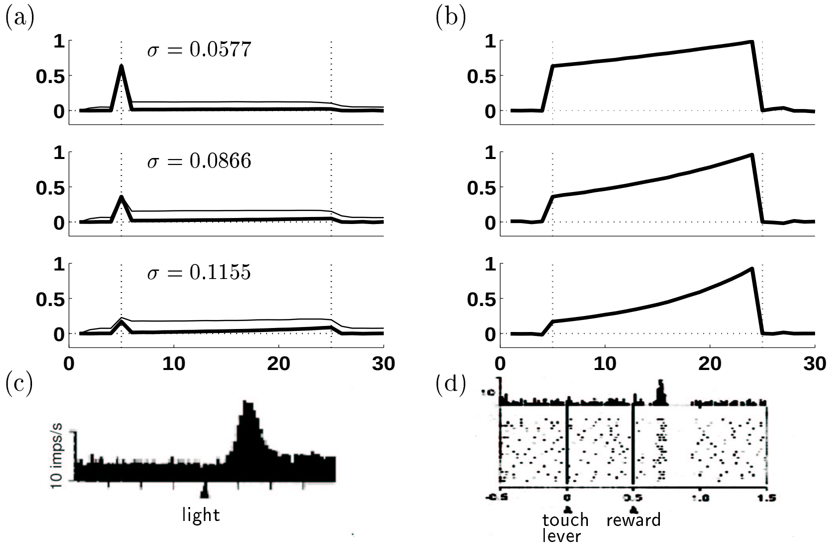

Fig. 2. (a) TD errors with different amounts of uniformly distributed representational noise ($\sigma =$ standard error). Thick line - without scaling of negative values, thin line - with scaling. (b) Respective weight values. (c) Elevated firing rate in the delay interval (after stimulus onset), reproduced from Mirenowicz and Schultz (1996). (d) DA response to the reward at early trials of training reproduced from Schultz et al. (1993).

delay. In these, the positive DA response at the time of reward was comparable to that in delay conditioning; but the ramping activity was reduced or eliminated. The TD model of DA readily explains these data by noting that the breadth of the ramp is determined by the learning rate $\alpha$ (fig. 1e). Trace conditioning is notoriously slow, suggesting a low learning rate, and thus a lower ramp.

## 2 Uncertainty in Representation: Temporal Noise

Another, more pervasive form of uncertainty, is that corrupting activities internal to the predictor. Neural activity is inherently very noisy, with high coefficients of variation; this sort of noise is known to affect learning. Furthermore, tasks such as FTS's involve learning about temporal intervals, and there is ample behavioral evidence[4] that this is particularly prone to noise.

In the tapped-delay line representation, neurons are either inactive or active $(x_i(t) = \{0, 1\})$ at time $t$. We first consider the effect on TD of adding different levels of uniformly distributed random noise to this representation. Fig. 2a shows TD to be robust to this noise, as the overall pattern of responses is maintained. However, the response at the time of the stimulus is reduced, particularly for high noise since the weights (and hence the mean predictions) decline away from the time of the reward as in a form of temporal discounting (fig. 2b). This decline arises from the regularization effect of the noise, favoring small weights.[2] The effective discount factor also results in a mild ramp in $\delta(t)$ during the delay, as seen in fig. 2a (thick

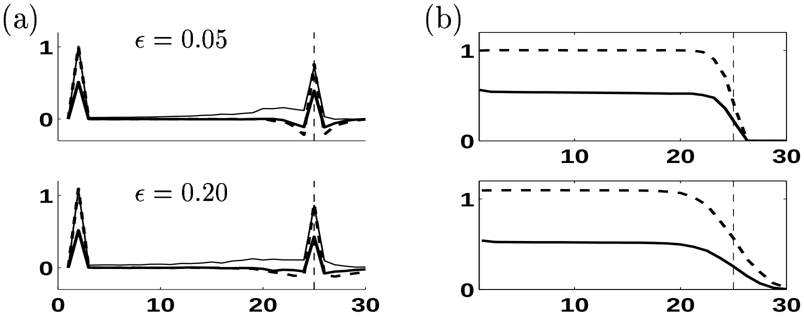

Fig. 3. TD errors (a) and weight values (b) with different amounts of Markov type timing noise, and different probabilities of reward (solid line: $p_r = 0.5$, dashed: $p_r = 1$). Note the response at time of reward. The learning rate has no evident effect on this steady-state response. Thin line in (a) - TD error with scaling of negative values, for the $p_r = 1$ case. In (b), plotted against objective time, is the average weight used at that time.

line). The magnitude of this ramp is proportional to $p_r$ rather than $p_r(1 - p_r)$. Adding scaling (fig. 2a - thin line) amplifies this effect, producing a higher-than-baseline firing during the delay. This elevated baseline after stimulus onset can also be seen experimentally (fig. 2c), again supporting TD.

Finally, interval timing is extremely inaccurate, with the standard deviation of the prediction of an interval proportional to its mean length.[4] Since the DA signal is apparently a sensitive indicator of errors in timing,[9] it is important to understand the consequences of temporal noise on TD. We modeled this by viewing the state of the neurons as a discrete Markov chain, in which being in state $i$ corresponds to the firing of neuron $x_i$, and the noise arises from a probability $\epsilon$ of remaining in the current state at each time step rather than advancing to the next state $i+1$. Fig. 3a depicts the steady-state TD prediction error $\delta(t)$ after the task has been learned, for different values of $\epsilon$, and for different probabilities of reward. Fig. 3b shows the corresponding weight values responsible for this response pattern. TD learning is very sensitive to temporal noise, with small $\epsilon$ resulting in a marked response at the time of reward delivery, even in thoroughly learned and fully predictable ($p_r = 1$) tasks. This response can be seen experimentally only in early phases of training (fig. 2d), but not in well-learned tasks.

## 3 Discussion

In sum, we have studied the effects of implicit and explicit uncertainty on the temporal difference prediction error, and compared our results with the activity of midbrain dopamine cells. We have shown the ramping DA response to be a straightforward consequence of TD learning of uncertain rewards, once the observably differential coding of positive and negative prediction errors is taken correctly into account. Most importantly, our analysis suggests that uncertainty need not play an explicit part in determining DA activity.

One feature of our account is that the learning rate should be maintained over trials. Pearce & Hall's[8] theory of the control of learning by uncertainty suggests exactly this – and there is evidence from partial reinforcement schedules that the learning rate may be higher when there is more uncertainty associated with the reward. A more puzzling aspect is the devastating effect of timing noise on the model DA responses – rather than having the reward responses predicted away, a reward peak arises within negative baseline. The puzzle is whether the high degree of noise in behavioral timing is consistent with the temporal sensitivity of the neural data.

*Acknowledgments*

We are very grateful to H. Bergman, N. Daw, D. Joel, C. Fiorillo, G. Morris, W. Schultz, P. Shizgal and P. Tobler for beneficial discussions. This work was funded by the EC Thematic Network (YN) and the Gatsby Charitable Foundation.

## References

[1] Barto AG, Sutton RS and Watkins CJCH Learning and sequential decision making. In M Gabriel, and J Moore, eds., *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, 539–602, Cambridge, MA, 1990. MIT Press.

[2] Bishop CM. Training with noise is equivalent to Tikhonov regularization. *Neural Comp.*, 7:108–116, 1995.

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

[4] Gallistel CR, and Gibbon J. Time, rate and conditioning. *Psych. Rev.*, 107:289–344, 2000.

[5] Mirenowicz J and Schultz W. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, 379:449–451, 1996.

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

[7] O'Doherty J, Dayan P, Friston K, Critchley H, and Dolan R. Temporal difference learning model accounts for responses in human ventral striatum and orbitofrontal cortex during Pavlovian appetitive learning. *Neuron*, 38:329–337,2003.

[8] Pearce JM and Hall G. A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psych. Rev.*, 87:532–552, 1980.

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

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

[11] Sutton RS and Barto AG. *Reinforcement Learning: An Introduction.* 1998. MIT Press.