The Effects of Uncertainty on TD Learning

Yael Niv[1] **Michael Duff**[2] **Peter Dayan**[2]
[1]ICNC, Hebrew University, Jerusalem [2]GCNU, University College London, London
yaelniv@alice.nc.huji.ac.il {duff,dayan}@gatsby.ucl.ac.uk

Abstract

Substantial evidence suggests that the phasic activities of dopamine neurons in the midbrain of primates report prediction errors in the delivery of reward. Recent recordings from these neurons in a task involving uncertain reward delivery present a crucial challenge to this hypothesis, and pose questions regarding the effects of both external and representational uncertainty on dopamine activity. Here, we analyse this issue, showing that the apparently anomalous activities are in fact what is expected under a standard prediction error account in light of different scalings of positive and negative prediction errors. We also study the implications of certain forms of representational noise on temporal difference learning.

1 Introduction

There is now a large body of neurophysiological, neuroimaging, and pscyhopharmacological data regarding the activity of dopamine cells (DA) in the midbrains of monkeys, rats and humans in tasks involving reward predictions.[19,22] In particular, the temporal patterns of the DA activity in tasks involving delayed prediction suggest[17,24] that DA neurons report a temporal difference (TD) error in the predictions of summed future reward.[26,27] This TD error signal provides a computationally compelling account of the role of DA in appetitive classical and instrumental conditioning.

As a quantitatively precise theory of DA activity, this account has been subjected to concerted experimental test, which in turn has led to its substantial enrichment.[14] A recent experiment by Fiorillo, Tobler & Schultz (FTS)[9] that investigated the effects of experimentally-controlled uncertainty, poses an important challenge to this TD account. In this paper, we suggest a theoretical answer to this challenge that maintains the integrity of the TD theory, and further enriches it in light of the data. We also study the effects of another source of uncertainty – that in the internal representations governing the predictions.

In previous DA experiments, the predictive relationship between stimuli (or actions, in the instrumental case) and receipt of reward was deterministic. However, predictive relationships often have an inherent uncertainty. For example, the appetizing scent of food greeting us as we come home may signal that there is a freshly made dinner waiting for us, or alternatively, for the equally hungry neighbor, and until we enter, we cannot reliably predict the availability of food. FTS[9] studied the consequences of inherently uncertain rewards, by associating the presentation of visual stimuli to monkeys with the delayed delivery of probabilistic rewards (drops of juice). Five different stimuli were associated with five different reward probabilities ($p_r = 0, 0.25, 0.5, 0.75, 1$). The vigor of the anticipatory licking of the juice tube by the monkeys in the two second interval between stimulus and reward indicated that their predictions were sensitive to these different probabilities.

Figure 1a shows the basic results. For each reward probability, a population histogram

depicts the summed and averaged activity of many DA cells across trials. Whereas the decreasing responses to the visual stimuli for decreasing p_r are consistent with the TD model, the TD framework offers no account of the apparent *ramping* of the responses towards the time of the reward. FTS, noting that this ramping is largest for $p = 0.5$, suggest that it reports the *uncertainty* (*e.g.*, entropy or variance) of the delivery of reward. They further speculate that this excess DA release could be a basis for the apparently appetitive properties of uncertainty, as in the attractiveness of gambling.

If true, this account poses a critical challenge to the TD interpretation of DA activity. First, the TD learning rule ensures that DA activity at a particular time in a trial (*e.g.*, at the time of unexpected rewards in early learning trials) comes to be *predicted away* by earlier cues in the same trial, leading to the DA response at the time of the conditioned stimulus. As persistently present activity, the ramp should thus have been predicted away by the preceding cue. This is evidently not the case, and would mean that prediction learning could *not* be based on the DA signal as a prediction error. Second, in the link to behavior, a key aspect of conventional versions of the theory[1] is that it is the activity of the DA cells to early predictors that influences the choice of actions to yield better rewards. The ramping activity of the DA cells is like a constantly surprising reward that is never predicted by earlier cues, and therefore, can never influence the choice of actions (*e.g.*, the decision to gamble) in order to arrange for it to be delivered.

The activities in figure 1a show additional facets that do not arise from the TD model, but point to an underlying resolution. First, since the responses shown have been averaged over many trials, it would be expected that the trial-by-trial positive and negative responses to the occurence and omission of the reward, respectively, would cancel each other (as the mean of these deviations from the mean predicted reward is zero), and on average there would be no response at the time of reward. However, the data clearly shows a positive response at the time of reward for the intermediate probabilities. Second, TD predicts that in the $p_r = 0$ case there would be no response to the stimulus, while in the $p_r = 1$ case there would be no response to the fully predicted reward (as shown in previous experiments[21]). However, in both cases there is a small but clear response.

In section 2, we study the ramping and other anomalous responses from the perspective of the standard TD model. We interpret them as arising from two key facets of DA: different scales of coding for positive and negative prediction errors and a fixed learning rate leading to continuous prediction errors. In section 3, we extend our study to uncertainty arising as a result of noise in the neural representation of the temporal aspects of the task.

2 Uncertainty in reward occurence: Ramping

We modeled the delayed-reward conditioning task of FTS using a tapped-delay line representation[15] of the time elapsed since stimulus onset. Each unit becomes active (i.e., takes the value 1) for one time-step at a specific lag after the stimulus has been presented, so that every time-step after the stimulus onset is represented by the firing of one unit. This representation, although not biologically plausible for long delays, is widely used in TD.[17]

In the simulated task, in every 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, and the trial ended at $t = 30$. A different set of neurons was used to represent each of the different stimuli across time, such that different predictions could be associated with each stimulus. Each neuron's weight represented an expected future reward value, and these weights were learned via an online TD learning rule with a fixed learning rate α:

$$\delta(t) = r(t) + \mathbf{w}(t-1) \cdot \mathbf{x}(t) - \mathbf{w}(t-1) \cdot \mathbf{x}(t-1) \quad (1)$$

$$\mathbf{w}(t) = \mathbf{w}(t-1) + \alpha \delta(t) \mathbf{x}(t-1) \quad (2)$$

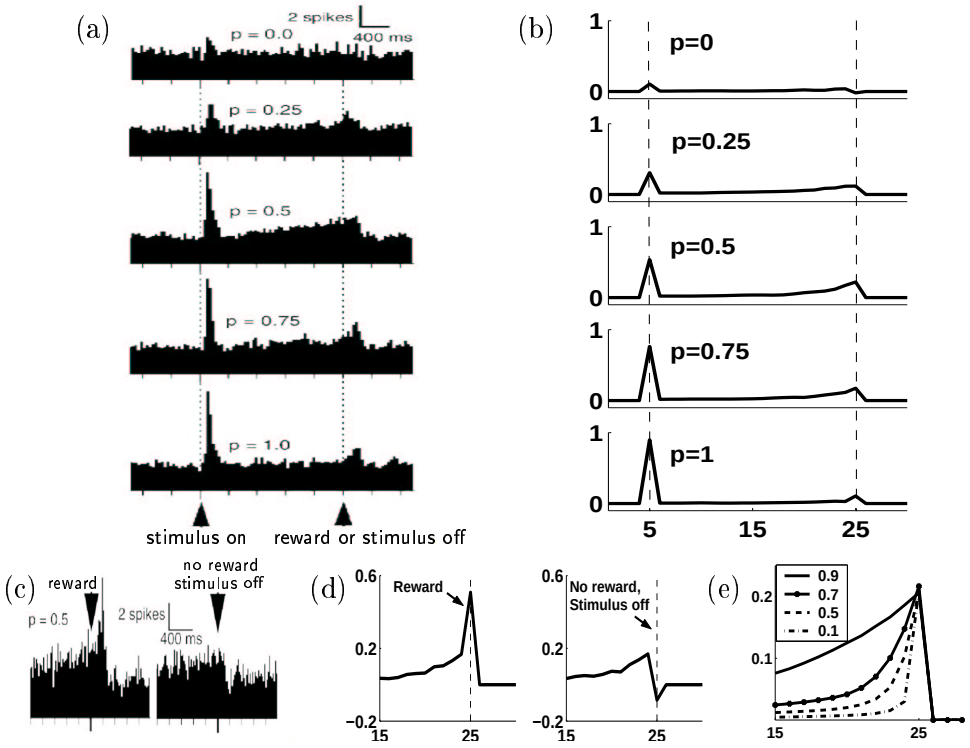

Figure 1: (a) DA response in trials with different reward probabilities, reproduced from Fiorillo *et al.*[9] Rewarded and unrewarded trials are pooled at intermediate probabilities. (b) TD prediction error with asymmetric scaling. (c) Response in $p_r = 0.5$ trials, separated into rewarded (left) and unrewarded (right) trials. (d) Model of (c). (e) Ramping is ordered by learning rate.

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. With FTS, we are interested in the prediction error $\delta(t)$ over many trials, after the task has been learned.

The experimental data in Figure 1(a,c) is presented as a peri-stimulus time histogram (PSTH) obtained by summing the spikes generated by the recorded neurons in discrete time bins. The DA signal is taken to represent the $\delta(t)$ prediction error signal, by considering its values above and below the baseline firing rate as representing positive and negative prediction errors. However, note that the neural signal is not symmetric about its baseline firing rate (which is relatively low at 2-4Hz). The positive signal can be as high as $\sim 270\%$ above baseline, while the minimal negative signal is only $\sim 55\%$ below baseline.[9] This asymmetry of the error signal is a result of the constraint posed by the neural substrate, which can only fire positive, but not negative spikes. We incorporate this asymmetry by "post-processing" the $\delta(t)$ values to scale negative values by a factor of $1/6$ compared with positive values. Crucially, we assume that this post-processing has no effect on the learning of the weights according to eq. (2), as this would lead to learning of incorrect predictions. In the neural mechanism, different scaling for potentiation and depression of the weights should compensate for the asymmetric error signal to ensure that the correct predictions are learned. Alternatively another opponent neurotransmitter may be involved in representing, and therefore learning, the negative prediction errors.[4]

Figure 1b shows the PSTH computed from our simulated data by averaging over the $\delta(t)$

signal in 50 trials for each stimulus type, after applying asymmetric post-processing. Figure 1d shows the results for the $p_r = 0.5$ case, divided into rewarded and unrewarded trials for comparison with figure 1c. The simulated results resemble the experimental data closely in that they replicate the net positive response to the uncertain rewards, as well as the ramping effect, which is highest in the $p_r = 0.5$ case. The shape of the ramping, albeit not the height of its peak, is affected by the learning rate, with higher learning rates producing more gradual ramping (figure 1e). The learning rate in figure 1c was set to 0.8, which is relatively high, but this should not be taken as the literal synaptic learning rate of the neural substate, given our schematic representation of the stimulus. In a more realistic representation in which a population of neurons is active at every time-step, a much lower learning rate would produce similar results. Interestingly, FTS, as well as Morris, Arkadir, Nevet, Vaadia and Bergman (personal communication), report that in experiments in which the stimulus was not present throughout the delay until the reward ("trace" conditioning, as opposed to "delay" conditioning), the ramping effect was not seen, although the positive response at the time of reward was comparable to that in the delay conditioning task. This could simply be a result of a lower learning rate in such experiments, as behavioral data shows that these tasks are more difficult to learn.

The maximum value of the ramping effect can be analytically derived for TD learning with the simplified tapped-delay-line time representation. The weight value at the last time-step in a trial $(t = N)$, as a function of trial number T (with initial values taken to be zero), is $w_N(T) = \alpha \sum_{t=0}^{T-1} (1-\alpha)^t r_{T-t}$ Where r_T is the reward at the end of trial T. The error signal at the last time-step of trial T is simply the difference between the obtained reward r_T, and the value predicting that reward $v[T-1]$. This error is positive with probability p (in which a reward was obtained), and negative with probability (1–p). Scaling the negative errors by a factor of d we get

$$< \delta[T] >= p - (1 - (1-\alpha)^{T-1})(p^2 + dp(1-p)) \xrightarrow[T\to\infty]{} p - (p^2 + dp(1-p)) \quad (3)$$

which is indeed maximized by $p = 0.5$ regardless of the precise scaling factor.

The positive response to the stimulus predicting reward with $p_r = 0$, as well as the response at time of reward for the $p_r = 1$ condition are both also apparent in the simulated results (Figure 1b). These responses result from a different source of uncertainty, namely failures to correctly identify the stimulus in order to predict the forthcoming reward. Behavioral measures show that this identification is not 100% correct even for overtrained tasks – in a similar task requiring an operant response in order to obtain the reward at the specified probability, Morris *et al.* (personal communication) note that the monkeys occasionally make identification mistakes. We assumed a misidentification rate of 8%, so anomalous responses naturally occur on some of the $p_r = 0; 1$ trials. Theories of generalization responses in DA neurons[14] would suggest that the animals might initially make identification errors, but then correct themselves before the time of the reward. This would lead to a specific pattern of expected responses for $p_r = 0$ (which unfortunately cannot be assessed from the data, as a result of the small number of trials), but would not explain the response for $p_r = 1$.

In sum, we have shown that the ramping effect, as well as the net positive response at the time of reward, are straightforward results of TD learning of uncertain rewards, given a neural substrate with an asymmetric representation of positive and negative prediction errors. This explanation of ramping does *not* assume that DA neurons code for anything other than the reward prediction error, and does *not* license a coupling to appetitive aspects of uncertainty. Of course, we are not claiming that the 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.[6]

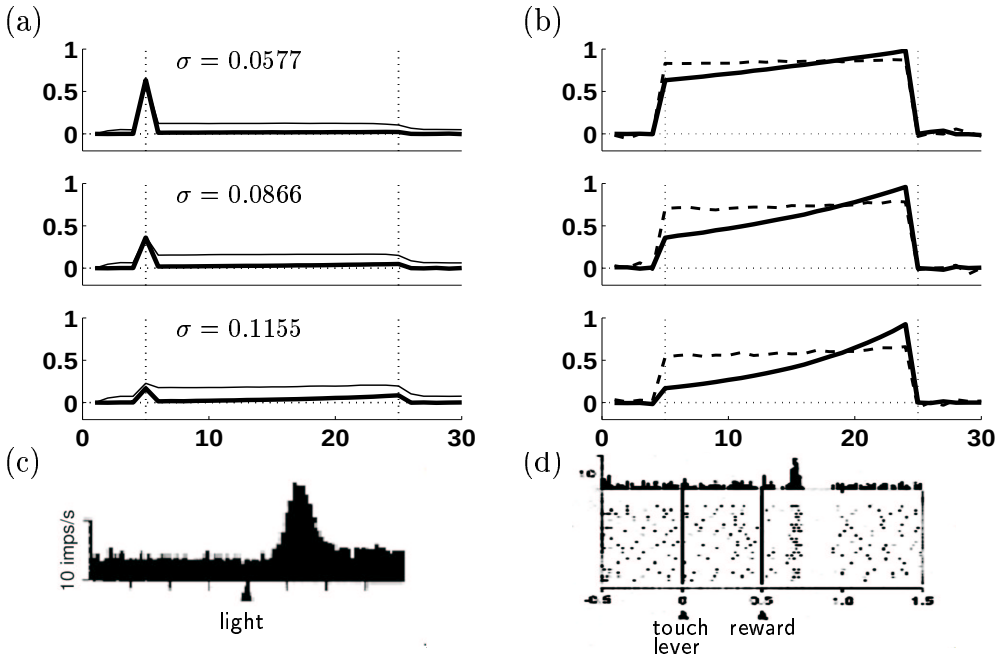

Figure 2: (a) TD errors with different amounts of uniformly distributed representational noise ($\sigma =$ standard error). Thick line - without scaling of negative values, thin line - with scaling. (b) Respective weight values for TD(0) (solid line), and TD(1) (dashed). Note the decaying of the weights for the TD(0) case (and the resulting ramping as well as a smaller response to the stimulus in (a)), in contrast to the flat weights for TD(1). (c) Elevated firing rate in delay interval (after stimulus onset), reproduced from Mirenowicz & Schultz.[16] (d) DA response to the reward at early trials of training, showing a positive response embedded in a trough (reproduced from Schultz *et al.*[23]).

3 Uncertainty in representation: Noise

Fiorillo *et al.*'s experiment concerns uncertainty that is explicit and under control of the experimenter. 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 (summarized in theories such as scalar expectancy theory[10,11]) that interval timing is particularly prone to noise. In this section, we first consider non-temporal and then temporal noise.

3.1 Additive representational noise

In the tapped-delay line representation, a neuron is either active $(x_i(t) = 1)$ or inactive $(x_i(t) = 0)$ at time t. These could be represented in a rate code by the presence or absence of bursts of population activity. This representation inevitably involves noise, affecting both bursting and quiescent states. We model this type of noise by assuming continuous values for $x_i(t)$, adding a uniformly distributed random noise to the $1/0$ representation.

Figure 2a shows the results of adding different levels of noise to the stimulus representation. TD learning proves to be very robust under this type of noise, as the overall pattern of responses was not changed. However, the magnitude of the response at the time of stim-

ulus onset is reduced, with lower responses for higher noise levels. Figure 2b shows the provenance of this effect: the weights (and also the mean predictions) decline away from the time of the reward as in a form of temporal discounting.

This decline can be understood in terms of the regularizing effect of representational noise,[2] which acts as a prior on the weights, encouraging them to be small. Given a standard least mean square rule (*ie* the so-called TD(1) learning rule), this would lead to uniformly reduced values (figure 2b - dashed); but for our TD(0) rule, the shrunken weight at one time between the stimulus and reward leads to a yet further shrunken weight at an earlier time, and thus the apparent discounting (figure 2b - solid). Since the effective discount factor is not a result of changing the learning rule, it also results in a mild ramp in $\delta(t)$ in the delay interval, as can be seen in figure 2a (thick line). However, this ramp is distinct from the ramp discussed in section 2, as it is ordered by the probability of reward, rather than the variance of the uncertainty in the reward, and is thus maximal for $p_r = 1$. Adding scaling (figure 2a - thin line) obviously amplifies this effect, producing a higher-than-baseline firing in the delay interval. Such an elevated baseline after stimulus onset can also be seen experimentally (figure 2c).

3.2 Timing noise

Behavioral results suggest that interval timing is extremely inaccurate, with the standard deviation of the prediction of an interval being proportional to the mean length of the interval (scalar expectancy theory – SET[10,11]). Since the DA signal is apparently a sensitive indicator of errors in timing,[13,22] it is important to understand the consequences we would expect from the TD model. Unfortunately, there is little data about the neural representation of interval timing, making it hard to validate a model.

One critical desirable feature of a temporal representation is *monotonicity*, that is – subjective time representation should only tick forward. We therefore examine the form of the TD prediction error in a model of a monotonic, but noisy, representation of time. In this, we view the state of the timing neurons as a discrete Markov chain, in which being in state i corresponds to the firing of neuron x_i. At every simulated time step, we assume that the probability of staying in the current state is ϵ, and that of transition to state $i+1$ is $(1-\epsilon)$. Thus the mean time counted in the interval between stimulus onset and reward is less than, or equal to the actual length of this interval. This form of timing noise is computationally convenient, but it does not satisfy the scaling relationship in SET,[10,11] as $\mu = t(1-\epsilon)$, while $\sigma = \sqrt{t(1-\epsilon)\epsilon}$). Figure 3a depicts the steady-state TD error $\delta(t)$ after the task has been learned, for different values of ϵ, and for different probabilities of reward. Figure 3b shows the corresponding weight values responsible for this response pattern.

Although monotonicity of the time representation is preserved, TD learning is very sensitive to this type of noise, as even very small amounts of noise result in a marked response at the time of reward delivery, even in thoroughly learned and fully predictable ($p_r = 1$) tasks. This response, which is comparable in size to the response to the stimulus onset, is not seen experimentally. The decay of the weights towards the end of the trial results in a trough from which this positive response rises, with the width of the trough larger, and its minimum higher, for larger amounts of noise. A similar pattern of DA firing has been observed in the early phases of animal training, in which the positive response to the reward can be seen in a trough of DA quiescence (figure 2d). However, this response disappears after more extensive training. Our results show this devastating effect of timing noise on the TD error signal, to rather be a phenomenon of the learned steady state. Indeed, an experiment and associated model involving timing noise which was explicit in the experimental protocol, showed a similar phenomenon.[14]

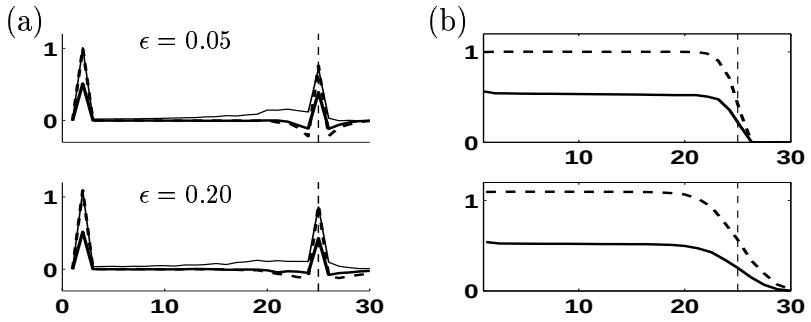

Figure 3: TD errors (a) and weight values (b) with different amounts of Markov type timing noise, and different probabilities of reward (solid line: $p_r = 0.5$, dashed: $p_r = 1$). Note the response at time of reward. The learning rate has no evident effect on this steady-state response. Thin line in (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.

4 Conclusions

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 showed that the ramping response recently observed in DA cells by Fiorillo *et al*[9] arises naturally from standard TD, once the observably differential coding of positive and negative prediction errors is taken correctly into account. This precludes the requirement for abandoning the existing model for the interpretation of phasic DA responses. This differential encoding also accounts for various other features of the response that FTS observed. Of course, in an opponent process model[25] of appetitive and aversive motivational forces, it would be ideal to observe both opponent signals rather than just one. We eagerly await the relevant data on the behavior of potential candidate opponents[4,7] in this task.

One feature of our account is that the learning rate should be maintained over trials. Pearce & Hall's[20] 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. Such an effect would boost ramping for the $p_r = 0.5$ case even more. Of course, if the animal knew that the world was stationary, then it would correctly *reduce* its learning rate over trials to balance old and new information appropriately.[5] FTS's experimental design does not allow a test of this.

The close coupling of uncertainty (in the form of novelty) and reward has previously been noted,[14] and indeed plays an important part in bonus (fictitious reward) theories of reinforcement learning[18] and Bayesian decision theoretic[8,12] accounts of exploration. However, except in very special circumstances (usually to do with short time horizons) predictable uncertainty, as in FTS, ultimately loses its power to inspire fictitious rewards. Perhaps we will have to look elsewhere to understand the lure of gambling.

A distinctly more puzzling aspect of our results is evident in figure 3a. This shows the devastating effect of timing noise in the model DA responses – instead of having reward response that are neatly predicted away (as in the $p_r = 1$ trace in figure 1a), a significant reward peak arises that rides on top of a negative baseline. The puzzle is whether the high degree of noise in behavioral timing is consistent with the temporal sensitivity displayed by the neural data. Daw[3] has suggested a different, and more powerful, account of timing, involving a hidden semi-Markov model. However, at present, this model more poses a similar puzzle rather than answers it.

Acknowledgments

We are very grateful to Hagai Bergman, Daphna Joel, Christopher Fiorillo, Genela Morris, Wolfram Schultz and Philippe Tobler for beneficial discussions, and Daphna Joel for very helpful comments on the manuscript. 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] Daw ND, Courville AC, and Touretzky DS. Timing and partial observability in the dopamine system. In TG Dietterich, S Becker, and Z Ghahramani, eds., *Advances in NIPS 14*, Cambridge, MA, 2002. MIT Press.

[4] Daw ND, Kakade S, and Dayan P. Opponent interactions between serotonin and dopamine. *Neural Networks*, 15(4-6):603–616, 2002.

[5] Dayan P, Kakade S, and Montague PR. Learning and selective attention. *Nat. Neurosci.*, 3:1218–1223, 2000.

[6] 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.

[7] Deakin JFW. Roles of brain serotonergic neurons in escape, avoidance and other behaviors. *J. Psychopharm.*, 43:563–577, 1983.

[8] Duff MO. Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes. Ph.D. dissertation, University of Massachusetts, Amherst, 2002.

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

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

[11] Gibbon J. Scalar expectancy theory and Weber's Law in animal timing. *Psych. Rev.*, 84:279–325, 1977.

[12] Gittins JC. Multi-Armed Bandit Problems. New York, NY:Wiley & Sons, 1989.

[13] Hollerman JR and Schultz W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.*, 1:304–309, 1998.

[14] Kakade S and Dayan P. Dopamine: generalization and bonuses. *Neural Networks*, 15(4-6):549–559, 2002.

[15] Kehoe EJ. Effects of Serial Compound Stimuli on Stimulus Selection in Classical Conditioning of the Rabbit Nictitating Membrane Response. Ph.D. dissertation, University of Iowa, 2002.

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

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

[18] Ng AY, Harada D, and Russel S. Policy invariance under reward transformations: Theory and application to reward shaping. *Proc. of the 16th International Conference on Machine Learning*, 1999.

[19] 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.

[20] 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.

[21] Romo R and Schultz W. Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-intiated arm movements. *J. Neurophys.*, 63:592–606, 1990.

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

[23] Schultz W, Apicella P, and Ljungberg T. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neurosci.*, 13(3):900–913, 1993.

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

[25] Solomon RL and Corbit JD. An opponent process theory of motivation. I. Temporal dynamics of affect. *Psych. Rev.*, 81:19–145, 1974.

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

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