A Model of Dopamine and Uncertainty Using Temporal Difference

Angela J. Thurnham* (a.j.thurnham@herts.ac.uk), D. John Done (d.j.done@herts.ac.uk),**
Neil Davey* (n.davey@herts.ac.uk), Ray J. Frank* (r.j.frank@herts.ac.uk)
School of Computer Science,* School of Psychology, **University of Hertfordshire,
College Lane, Hatfield, Hertfordshire. AL10 9AB United Kingdom

Abstract

Does dopamine code for uncertainty (Fiorillo, Tobler & Schultz, 2003; 2005) or is the sustained activation recorded from dopamine neurons a result of Temporal Difference (TD) backpropagating errors (Niv, Duff & Dayan, 2005)? An answer to this question could result in a better understanding of the nature of dopamine signaling, with implications for cognitive disorders, like Schizophrenia. A computer simulation of uncertainty incorporating TD Learning successfully modelled a Reinforcement Learning paradigm and the detailed effects demonstrated in single dopamine neuron recordings by Fiorillo et al. This alternate model provides further evidence that the sustained increase seen in dopamine firing, during uncertainty, is a result of averaging firing from dopamine neurons across trials, and is not normally found within individual trials, supporting the claims of Niv and colleagues.

Keywords: Dopamine; Uncertainty; Single Cell Recordings; Temporal Difference; Computer Simulation.

Dopamine and Uncertainty

Current theories of the effects of dopamine on behaviour focus on the role of dopamine in Reinforcement Learning, where organisms learn to organise their behaviour under the influence of goals, and expected future reward is believed to drive action selection (McClure, Daw & Montague, 2003; Montague, Dayan & Sejnowski, 1996; Schultz, Dayan & Montague, 1997; Suri & Schultz, 1999). Single cell recordings of dopamine neurons have identified a phasic dopamine burst of activity which is posited to be a reward prediction error (Schultz, 1998; Waelti, Dickinson & Schultz, 2001) and Temporal Difference (TD) Learning (Sutton, 1988; Sutton & Barto, 1998), a form of Reinforcement Learning, provides an explicit method of modelling and quantifying this error (Hollerman & Schultz, 1998; Schultz et al., 1997). It is likely that disruption to the dopamine system gives rise to an abnormality in information processing by dopamine and some of the symptoms currently associated with schizophrenia, particularly psychosis and deficits in working memory.

It has been posited that dopamine also codes for uncertainty (Fiorillo, Tobler & Schultz, 2003), as under conditions of maximum uncertainty, observations of single cell recordings have shown a sustained increase in activity from presentation of a conditioned stimulus (CS) to the expected time of a reward. They recorded the activity of neurons in two primates, identified as dopamine neurons from their electrophysiological characteristics, during a delay paradigm of classical conditioning to receive a fixed juice reward, while manipulating the probability of receipt of the reward. Two related but distinct parameters of reward were identified from the activation produced, after learning had taken place: (i) A phasic burst of activity, or reward prediction error, at the time of the expected reward, whose magnitude increased as probability decreased; and (ii) a new slower, sustained activity, above baseline, related to motivationally relevant stimuli, which developed with increasing levels of uncertainty, and varied with reward magnitude. Both effects were found to occur independently within a single population of dopamine neurons.

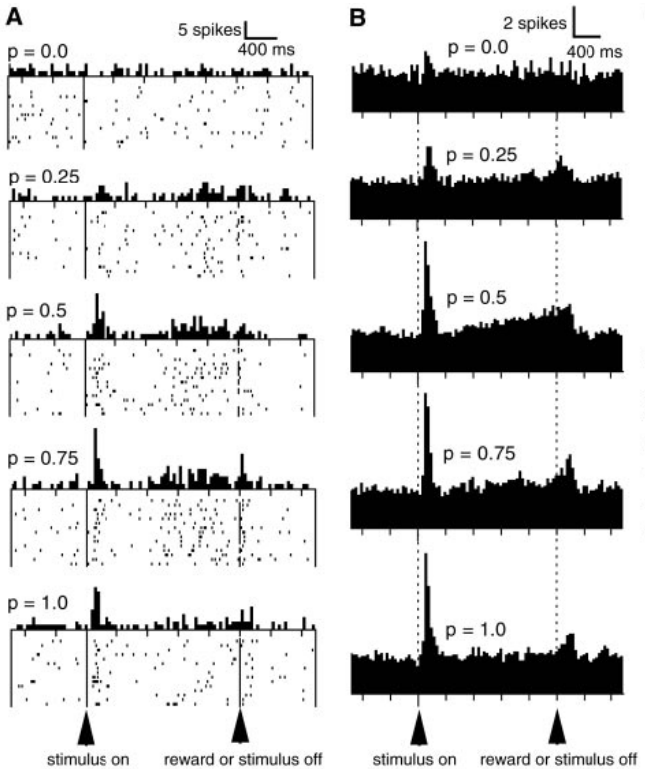

Figure 1: Sustained activation of dopamine neurons with uncertainty taken from Fiorillo et al. (2003) (A) Rasters and histograms of single cell activity (B) Population histograms

With uncertainty, the sustained activation began on presentation of a CS and increased in strength until a reward was due, at which point the activation ceased (Figure 1B, where P = 0.25, 0.5 and 0.75). This activation was greatest

when uncertainty of reward was at a maximum, i.e., when the reward was received on only 50% of occasions and probability (p) was 0.5. Sustained activation was also seen at lower values of uncertainty, when probability was 25% and 75%, but to a lesser extent. No sustained activation was seen when probability was certain at either zero or 1, suggesting that the sustained activation coded for uncertainty.

However, this view is controversial as Niv, Duff and Dayan, (2005) have suggested that the sustained activation, or 'ramping' effect in the delay period, is due to backpropagating TD prediction errors, and not to uncertainty. Specifically, they suggest that it is the asymmetric coding of those prediction errors that give rise to the effects seen in time, over consecutive CS presentations, due to a low baseline rate of activity in dopamine neurons. Firing rates of positive prediction errors typically rise to about 270% above baseline, while negative errors only fall to approximately 55% below baseline (Fiorillo et al. 2003). During uncertainty, these asymmetrical positive and negative errors, when summed, will not cancel each other out, as predicted by the TD algorithm, even after extensive training periods. The overall effect, as seen in Fiorillo et al., will be of (i) a positive response across trials at the expected time of reward, and (ii) a 'ramping' effect from presentation of the CS to the expected time of reward, described by Fiorillo and colleagues as sustained activation. The resulting effects arise as a result of averaging across multiple trials and are not a within trial phenomena.

Using TD, Niv and colleagues successfully modelled both effects identified by Fiorillo et al. (2003) during uncertainty. They also showed that the shape of the ramp depended on the learning rate, and that the difference in the steepness of the ramp between delay and trace conditioning could be accounted for by the low learning rates associated with trace conditioning, resulting in a smaller or even negligible ramp.

In reply to Niv et al., Fiorillo and colleagues defend their original claim that dopamine encodes uncertainty about reward (Fiorillo, Tobler & Schultz, 2005). Three of the five points raised are of particular interest to this study. Firstly, they give two examples as evidence of sustained activation within single trials, which is contrary to the postulations of Niv et al., and secondly, they suggest that activity in the last part of the delay period should reflect the activity of the preceding trial. Finally, they suggest that other ways of using TD to model dopamine as a TD error are more biologically plausible than backpropagating TD errors. It is important, therefore, to look at a range of models in order to understand the limitations of using the TD algorithm to model the role of dopamine.

In the present study a simulation of a 'rat' in a one-armed maze was used to investigate the claims of Fiorillo and colleagues, using an alternative TD model to Niv et al. The maze modelled was similar to that used by McClure et al. (2003) linking the ideas of reward prediction error and incentive salience, but contained an additional 'satiety' state and only allowed travel in one direction. The aim of this investigation was to use TD learning to model the following effects seen in dopamine neuron firing by Fiorillo and colleagues: (a) The phasic activation at the expected time of reward that increased as probability decreased; (b) the sustained increase in activity from the onset of the CS until the expected time of reward, during uncertainty, posited either as uncertainty, or as backpropagating TD prediction errors; and (c) the sustained activation increasing with increasing reward magnitude. In addition, in the discussion an attempt is made to address three of the points raised by Fiorillo et al. (2005) in response to Niv et al. (2005).

Method

Temporal Difference

The maze incorporated an 'actor-critic' architecture (McClure et al., 2003; Montague, Hyman & Cohen, 2004; Sutton & Barto, 1998), a form of reinforcement TD learning where an 'adaptive critic' computes a reward prediction error, which is used by the 'actor' to choose those actions that lead to reward.

The Critic The TD algorithm is designed to learn an estimate of a value function V^*, representing expected total future reward, from any state, s, (Equation 1), where t represents time and subsequent time steps t = 1, t = 2 etc; E is the expected value and r represents the value of the reward. γ is a discounting parameter between 0 and 1 and has the effect of reducing previous estimates of reward exponentially with time, so that a reward of yesterday is not worth as much as a reward of today. Equation 2 is Equation 1 in a recursive form that can be used in the learning process.

$$V^*(s_t) = E[r_t + \gamma\, r_{t+1} + \gamma^2\, r_{t+2} + \gamma^3\, r_{t+3} + \ldots] \quad \text{[Eqn 1]}$$

$$V^*(s_t) = E[r_t + \gamma V^*(s_{t+1})] \quad \text{[Eqn 2]}$$

TD prediction error is a measure of the inconsistency for estimates of value at successive time steps. The error, δ(t), is derived by rearranging Equation 2 into Equation 3, which is a measure of the relationship between two successive states and the current reward. This will give estimates, V, of the value function V^*. The dopamine prediction error signal, δ(t), takes into account the current reward, plus the next prediction multiplied by the discounting parameter γ, minus the current prediction. It is the error δ(t) that is equivalent to the dopamine reward prediction error, or learning signal, to create better estimates of future reward.

$$\delta(t) = r_t + \gamma V(s_{t+1}) - V(s_t) \quad \text{[Eqn 3]}$$

The Actor An extension to the TD model has been made to include a role for dopamine in biasing action selection using the same prediction error signal, δ(t), to teach a system to take the best actions, namely those that are followed by

rewards (McClure et al., 2003; Montague et al., 1996). The way an action is selected is that the actor randomly chooses a possible action, and the anticipated δ(t) is calculated using Equation 3. The probability of taking this action is then calculated from this δ(t) value using the softmax function in Equation 4 (where *m* and *b* are parameters of the softmax curve), which calculates the probability of that action occurring from the anticipated δ(t) value. If no action is selected, time is increased by one step and another random action is considered.

$$P \text{ (of taking action)} = (1 + e^{-m(\delta(t) - b)})^{-1} \qquad \text{[Eqn 4]}$$

Actions are generated with a probability of selection based on the predicted values of their successor states, preferring those actions that give a high burst of dopamine, or TD error signal. There is a greater probability of remaining at the same state and not making a move when the error signal is low as all states become increasingly probable.

Learning takes place in the model according to Equation 5, where α is a learning rate parameter.

$$V(s_i) \leftarrow V(s_i) + \alpha\, \delta(t) \qquad \text{[Eqn 5]}$$

The Maze

A computer simulation was constructed of a 'rat' learning to traverse a one-arm maze to receive a reward, using the TD algorithm with an 'actor-critic' architecture. Figure 2 shows a maze with positions modelled as five states, starting at State 0 (the CS) and progressing through intermediate states to receive a simulated reward in State 4 (the reward state). In order to model the breaks between maze runs in real rats, it was necessary to insert a 'satiety' state (State 5) into the maze, between the goal (State 4) and the start (State 0), where the transition between that state and State 0 remained at zero so that no learning could take place. This had the effect of resetting the value of start State 0 to zero, acting as a 'resting' state and ensuring that the 'rat' was always surprised when starting the maze. Without this additional state, the simulated rat learnt the value of the start state, and in effect, there was no CS. Intermediate states were added and removed, as required to make mazes of different lengths.

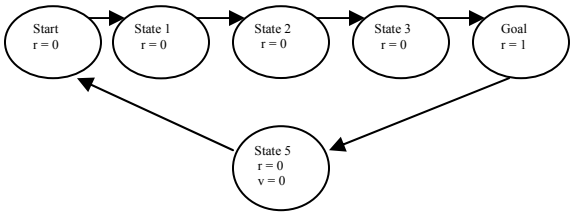

Figure 2: Maze with five states plus 'satiety' state

Simulations

Uncertainty – Degree of Probability The ranges of probabilities used for trials were 0.25, 0.5 (maximum uncertainty), or 0.75. The δ(t) values were recorded for each state transition, for a single probability in each trial. Each trial consisted of 1000 steps through a one-way maze with eight states plus a 'satiety' state, with a step being a transition from one state to the next, and a run being one complete journey through the maze, from start to finish. At the beginning of each trial the values of each state in the maze (V) were set to zero. Movement to the next state in the maze was selected according to the effect of TD learning on different probabilities of receiving a reward for each run.

In keeping with the biology of dopamine, namely the asymmetry in coding of positive and negative errors, any negative prediction errors were scaled by a factor of one sixth, the scaling factor used by Niv et al. (2005). The scaled δ(t) values were then averaged across fifty consecutive runs for each state, where γ = 0.98, and the magnitudes of the scaled values compared. This averaging corresponded to the summing of peri-stimulus-time-histograms (PSTH) of activity over different trials and inter-trial averaging used by Fiorillo et al. (2003).

Reward Magnitude Individual reward magnitudes of 0.5, 1 and 2 were compared in different trials to see the effect on the sustained activation.

An Example of Learning

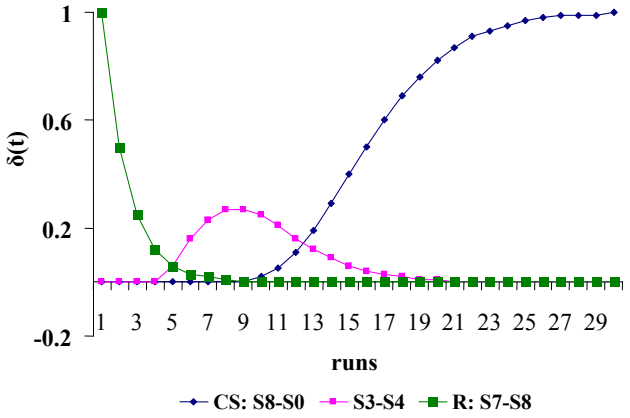

Figure 3: Delta values for each state transition over first thirty runs, p = 1, r = 1

With a probability of 1 and a maze of eight states plus a 'satiety' state, complete learning took place over the first thirty runs (γ = 1). On the first run a large prediction error, δ(t), was recorded at the expected time of the reward (S7-S8), and as runs progressed, this δ(t) was transferred back to the CS (S8-S0). When full learning had taken place only the CS elicited a reward prediction error. This effect is demonstrated in Figure 3, which shows the δ(t) at the expected time of reward beginning at 1 and reducing to zero by run 9 at which point the value of the state is learnt and the reward fully predicted. The δ(t) at the CS begins at zero and increases gradually to 1, from run 10 to run 30. An

intermediate state transition S3-S4 is included which records the δ(t) backpropagated from the reward state by run 5. The error increases until run 8 and then reduces to zero by run 21.

All the following tests with uncertainty were done post training.

Results

Uncertainty – Degree of Probability

Eventually, by chance, actions were selected in trials for the entire range of probabilities (p), 0.25, 0.5 or 0.75, and the 'rat' progressed along the maze towards the reward state receiving the reward (r) of that state, r = 1. On subsequent runs, learning occurred as the value of the reward was propagated backwards, updating earlier states using a proportion of the prediction error signal, δ(t).

The patterns of data obtained show that it is necessary for the history of previous runs to be taken into consideration when analysing reward prediction errors and not just the last trial. Accordingly, consecutive runs should be selected for averaging in order to preserve the backward chaining effect of the TD algorithm. The TD algorithm uses rewards obtained in the past to make predictions about future expected reward, affecting the values of all the states in the maze, which are continually being updated as the rat progresses along the maze. With uncertainty, the particular course a rat takes on a particular trial is novel in each trial, as it depends on the exact order of rewarded and non-rewarded runs, which are delivered randomly by the computer program. The δ(t) values are then propagated backwards, in order, from later states to earlier states, as time progresses.

As the probability of obtaining a reward increased, from 25% to 50% to 75%, so did the level of phasic activation at the CS (S8-S0) (Figure 5), with average δ(t) values of 0.23, 0.57 and 0.70 respectively.

(a) The phasic activations at the expected time of reward Without scaling the δ(t) values recorded for each state transition to compensate for the biologically asymmetric coding of positive and negative prediction errors, no average positive phasic activation was seen at the expected time of reward (Figure 4 S7-S8). However, after scaling δ(t) values by a factor of one sixth and averaging δ(t) values over consecutive trials, positive phasic activation was seen at the expected time of reward (Figure 5).

When comparing the average scaled δ(t) values across trials with probabilities of 0.25, 0.5 and 0.75, similar averaged, scaled δ(t) values were recorded of 0.16, 0.16 and 0.14 respectively. However, if averages were taken over rewarded trials only, as suggested in Figure 2A in Fiorillo et al. (2003), δ(t) values would be positive at the expected time of reward as all negative values would be removed. In addition, there would be less non-rewarded runs to be removed at higher probabilities, resulting in the phasic activation varying monotonically with reward probability. There would also be less of an effect on the phasic activation seen at the CS as the effect of reward would take longer to reach that state. As suggested above, difficulties arise when the backpropagating chain of reward prediction errors is broken and runs are taken out of context of the trial history.

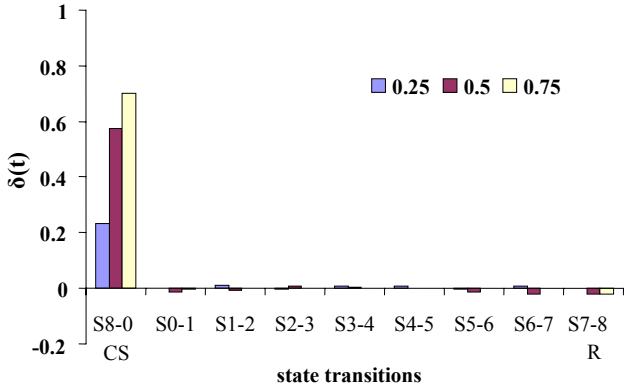

Figure 4: Average δ(t) values, **before scaling**, for each state transition over 50 runs, when p = 0.25, 0.5 and 0.75, α = 0.9

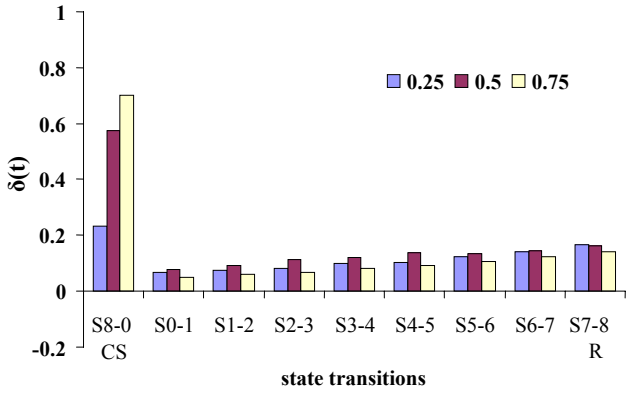

Figure 5: Average δ(t) values, **after scaling**, for each state transition over 50 runs, when p = 0.25, 0.5 and 0.75, α = 0.9

In conclusion, phasic activations were seen at the expected time of reward, in accordance with the findings of Fiorillo et al., when δ(t) values were scaled to compensate for the asymmetric coding. However, unless rewarded only trials were averaged, the phasic activation did not vary monotonically with reward probability.

(b) The sustained increase in activity Figure 4 shows that no 'ramping' effect is seen from plotting the average δ(t) values obtained for probabilities of 0.25, 0.5 and 0.75 for each state transition. Here the symmetrical positive and negative errors effectively cancel each other out, in accordance with the TD algorithm. However, when the δ(t) values were scaled by a factor of one sixth to compensate for the biological asymmetric coding of positive and

negative errors, and averaged across consecutive runs, positive δ(t) values were seen that corresponded to the sustained activation and 'ramping' effects reported in Fiorillo et al. (2003) and Niv et al. (2005) respectively (Figure 5). The magnitude of the ramping effect is marginally greater for maximum probability, p = 0.5 than for the lower probabilities of p = 0.25 and p = 0.75, in accordance with the findings of Fiorillo et al. However, the difference seen between the two trials with probabilities of 0.25 and 0.75, which are comparable levels of uncertainty, could be accounted for by the different reward history for each. This difference should be negligible if more trials were taken into account.

(c) Reward Magnitude The value of the reward was manipulated across different trials, with rewards given of 0.5, 1 and 2. The size of the reward had an effect on the range of δ(t) values available for each state. With a larger reward comes a larger range of possible δ(t) values, and, accordingly, larger 'ramping' effects (Figure 6). Therefore, the sustained activation increased with increasing reward magnitude, in accordance with Fiorillo et al. (2003).

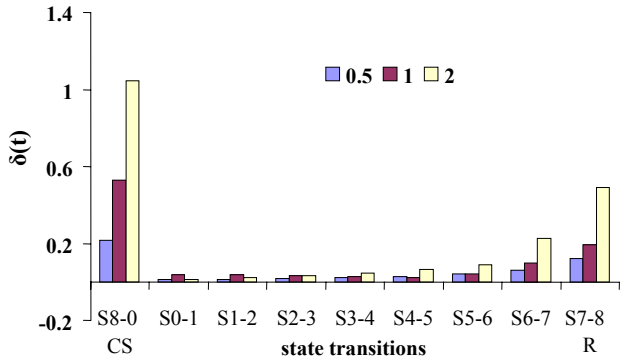

Figure 6: Scaled average δ(t) values over 30 runs for reward values of 0.5, 1 and 2, p = 0.5, α = 0.5

Discussion

A simulation of reinforcement learning, incorporating an 'actor-critic' architecture of TD learning, successfully modelled the following properties of dopamine demonstrated by Fiorillo et al. (2003): (a) The phasic activations at the expected time of reward; (b) the sustained increase in activity from the onset of the conditioned stimulus until the expected time of reward, during uncertainty; and (c) the sustained activation increasing with increasing reward magnitude. This supports the argument by Niv et al. (2005) that the ramping effect seen during uncertainty is a result of backpropagating TD errors and not a within-trial encoding of uncertainty.

In response to the claims of Niv and colleagues, Fiorillo et al. (2005) raised several points in support of their original argument, three of which are relevant to this study. Firstly, they refer to the difficulty of determining whether or not activity increases on single trials as Niv et al. (2005) did not specify what a single trial increase in delay-period activity should look like. In our simulations, Figure 7 is an example of a single trial (or a single run in this simulation), and is represented by recording the scaled prediction errors, δ(t), for each state transition, for one run through the maze. This run is analogous to the activity of a single neuron in a single trial over time and is simply a snapshot of the δ(t) values for each state, which may be either positive or negative with respect to baseline firing, depending on the history of previous runs.

The single run in Figure 7 is taken from actual run 6 in Figure 8 and represents non-rewarded run N preceded by ….RNRNR, where R is a rewarded run. The preceding RNR can be clearly identified in the δ(t) values seen for state transitions S4-S5, S5-S6 and S6-S7 respectively, but the results of earlier runs are harder to make out further back in time, as the TD algorithm ensures rewards or non-rewards in the past are not worth as much as those of the present.

Examination of many single runs through the maze did not reveal a ramping effect. Fiorillo et al. (2005) provided two examples of possible sustained activation in single trials, but these effects could have occurred quite by chance due to the order of rewarded and non-rewarded trials, as explained above, and not necessarily be examples of uncertainty. Indeed, if this within trial ramping effect were a regular occurrence then there would be many examples of single trials in support of the uncertainty hypothesis.

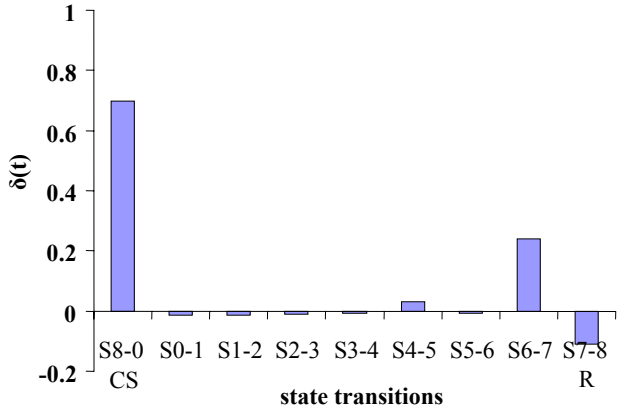

Figure 7: An example of a single trial:
Scaled δ(t) values for a single run, p = 0.5, r = 1

Secondly, Fiorillo and colleagues claimed that if activity during the delay period is due to backpropagating error signals that originated on previous trials, then the activity in the last part of the delay period of each individual trial should reflect the last reward outcome. Specifically, they suggest that if the preceding trial was rewarded, there should be more activity at the end of the delay period, and less activity if it was not rewarded, but they found no

dependence of neural activity on the outcome of preceding trials.

Our results show that it is necessary for more of the history of previous runs to be taken into consideration than just the last reward outcome, when analysing reward prediction errors. For example, Figure 8 shows a history of rewarded and non-rewarded runs RNRNRNNNNN. After scaling, large δ(t) values were seen for runs 1-6 because alternate rewards and non-rewards were given, but runs 7-10 were not rewarded and, consequently, gradual extinction of the negative prediction error occurred. This example shows that it is not always the case that less activity will be seen if a trial is not rewarded (and vice versa), as runs 8-10 show an increase in firing (towards baseline) following non-rewarded runs.

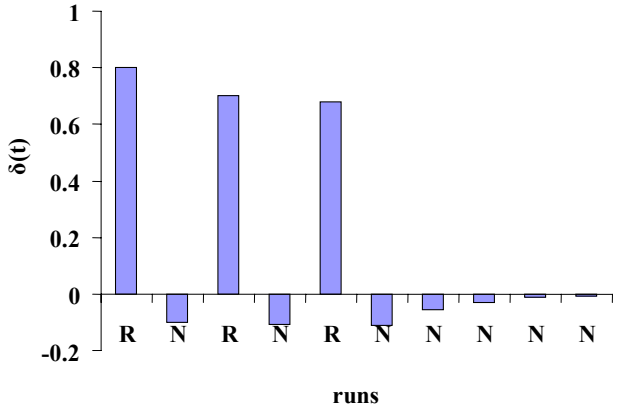

Figure 8: Scaled average δ(t) values at expected time of reward (S7-S8) recorded over 10 runs, p = 0.5, r = 1

Finally, Fiorillo et al. (2005) raise the argument that other TD models of dopamine are more biologically plausible than backpropagating TD errors, for example Suri & Schultz (1999), and it is important, therefore, to look at a range of models in order to understand the limitations of using the TD algorithm to model the role of dopamine. However, our work has shown that the predictions of Niv et al. (2005) are robust in the sense that they transfer to another type of model, albeit still using the same TD algorithm.

Conclusion

This alternate TD model to Niv et al. (2005) has effectively simulated conditioning in a Reinforcement Learning paradigm and successfully modelled the effects demonstrated in single dopamine neuron recordings, suggested to be coding for uncertainty, by Fiorillo et al. (2003). In addition, we have demonstrated what a single trial in TD Learning might look like and provide further evidence that ramping of the reward prediction error, δ(t), is not normally found within a trial of a single dopamine firing, but instead arises from averaging across trials.

Our simulations add further weight to the criticisms of Niv et al. that the effects demonstrated by Fiorillo and colleagues are due to backpropagating TD errors, and not a within-trial encoding of uncertainty. We support the claims by Niv et al. (2005) that the ramping signal is the best evidence yet for the nature of the learning mechanism of a shift in dopamine activity from expected time of reward to the CS.

References

Fiorillo, C.D., Tobler, P.N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299*, 1989-1902.

Fiorillo, C.D., Tobler, P.N., & Schultz, W. (2005). Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors. *Behavioral and brain Functions*, *1:7.*

Hollerman, J.R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1,* 304-309.

McClure, S.M., Daw, N.D., & Montague, P.R. (2003). A computational substrate for incentive salience. *Trends in Neuroscience 26(8),* 423-428.

Montague, P.R., Dayan, P., & Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian Learning. *Journal of neuroscience, 16(5),* 1936-1947.

Montague, P.R., Hyman, S.E., & Cohen, J.D. (2004). Computational roles for dopamine in behavioral control. *Nature 431,* 760-767.

Niv, Y., Duff , M.O., & Dayan, P. (2005). Dopamine, uncertainty and TD Learning. *Behavioral and brain Functions, 1:6* 1-9.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80,* 1-27.

Schultz, W., Dayan, P., Montague, P.R. (1997). A Neural substrate of prediction and reward. *Science*, *275:* 5306, 1593-1599.

Suri, R.E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91(3)*, 871-890.

Sutton, R.S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning 3,* 9-44.

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

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412*, 43-48.