Proceedings of the Second International Conference on Development and Learning, *Cambridge, MA, June 12–15, 2002.*

Dopamine and Inference About Timing

Nathaniel D. Daw[1,3] Aaron C. Courville[2,3] David S. Touretzky[1,3]

[1]Computer Science Department
[2]Robotics Institute
[3]Center for the Neural Basis of Cognition
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Temporal-difference learning (TD) models explain most responses of primate dopamine neurons in appetitive conditioning. But because existing models are based in the simple formal setting of Markov processes, they do not provide a realistic account of the partial observability of the state of the world, nor of variation in event timing. For instance, the TD model of Montague et al. (1996) mispredicts the dopamine response when an expected reward is delivered early.

We explain such experimental results using a version of TD learning grounded in the richer formalism of partially observable semi-Markov processes. We propose that the brain infers the likely state of the world from limited observations, using a statistical model of how the world's state evolves. Inference is necessary for such judgements as whether an expected reward is merely late, versus having been omitted altogether. The dopamine signal is modeled as a TD error signal for learning to predict future rewards from this inferred state representation.

1 Introduction

Several investigators have suggested that the primate dopamine system carries an error signal for learning to predict future rewards [1, 2, 3]. These models, based on temporal-difference (TD) learning [4], explain most phasic responses of primate dopamine neurons in appetitive conditioning [5]; moreover, they suggest a neurophysiological account of animal conditioning behavior. But because existing models are based in the simple formal setting of Markov processes, they are deficient in at least two areas relevant to physiological and behavioral data. They do not provide a realistic account of the *partial observability* of the state of the world, nor of how the system tracks the *timing* of events. In this paper, we introduce a version of TD learning grounded in a richer formal model to better address both issues and, consequently, to explain some data that challenge existing models.

In a Markov process, the setting for the basic TD algorithm, the state of the world relevant to reward prediction is always fully observable. This property does not hold for many of the key experiments on dopamine neurons. These are better modeled as *partially observable* Markov processes, where the state of the world is hidden and observations may reveal only ambiguous information about it. For instance, in an experiment where a brief flash of light signals that a food reward will be delivered a few seconds later, nothing observable about the world differentiates states in the interim period when reward is imminent. TD models of dopamine handle these situations by augmenting their representation of the state of the world with information about previous observations (such as, in this example, the flash of light) in hopes that the combination will be sufficient to fully predict reward.

This device introduces problems due to how it represents the timing of the previous observations. The state augmentation scheme of previous models treats, for instance, a flash of light five seconds ago and the same stimulus six seconds ago as totally unrelated events. This precludes a satisfactory treatment of *variability* in the intervals between events; as a result such models mispredict the behavior of dopamine neurons when a signaled reward is delivered early. We address this problem using the theory of partially observable *semi*-Markov processes, which explicitly incorporate variability in the timing of events. Such variation interacts with the problem of partial observability in our model: If a flash of light signals reward after some variable delay, as time passes without the reward occurring, the algorithm must determine the chance that the reward is simply late, versus that it was omitted altogether. We treat this as an *inference* problem, and solve it by using a statistical model of how the world's hidden state evolves to infer the state from a series of observations. Because this inferred state is generated from a model of the process's dynamics, it provides a better representation for TD learning than the augmented state of previous models. A similar inference method has recently been proposed to explain a number of behavioral results from animal conditioning [6]. The present paper aims to connect such a theory with known physiology by investigating how state inference and reward prediction systems might interact to generate the dopamine signal.

2 The model

A *Markov process* consists of a set $\mathcal{S}$ of states, and two functions Q and R defined on that set. If the state of the world at time t is $s_t \in \mathcal{S}$, then the transition function $Q(s_t)$ defines a probability distribution over $\mathcal{S}$ from which is drawn the successor state s_{t+1}. Rewards are also delivered: the reward r_t received at time t in the state s_t has magnitude distributed according to the function $R(s_t)$. The TD algorithm learns a third function, the *value function*, which maps each state to the discounted reward expected in the future:

$$V(s_t) = E\left[\sum_{t' \geq t} \gamma^{t'-t} r_{t'}\right] \tag{1}$$

where the parameter $\gamma < 1$ controls the steepness of discounting and the expectation is over variability in reward magnitudes and state transitions.

The TD algorithm [4] follows from a recursive rewriting of equation 1:

$$V(s_t) = E\left[r_t\right] + \gamma \cdot E\left[V(s_{t+1})\right] \tag{2}$$

Using this relation, an estimate, $\widehat{V}$, of V can be improved online. If, in some state s_t, a reward r_t and a successor state s_{t+1} are observed, they can be taken as samples of the distributions over which the expectations are taken in equation 2. Using $\widehat{V(s_{t+1})}$ as an approximation for $E\left[V(s_{t+1})\right]$, a sample of the right side of equation 2 can be computed, and the estimate $\widehat{V}(s_t)$ can be updated in its direction. The change in $\widehat{V}(s_t)$ is proportional to the difference between the approximated left and right sides of equation 2:

$$\delta_t = r_t + \gamma \widehat{V}(s_{t+1}) - \widehat{V}(s_t) \tag{3}$$

The TD models [1, 2, 3] propose that the activity of dopamine neurons reflects this error signal δ_t.

This paper considers TD algorithms for a richer setting, a *partially observable semi-Markov process*. The

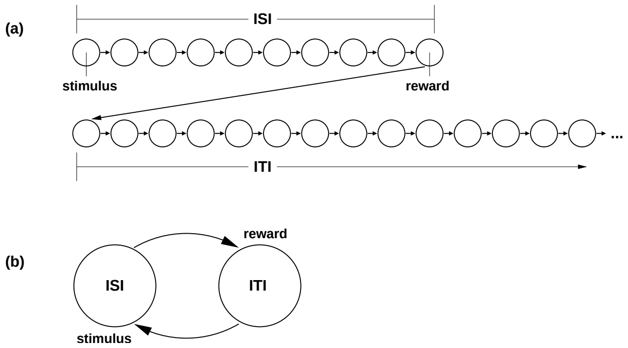

Figure 1: Models of a conditioning task. (a) Markov model. The passage of time is indicated by progression through a series of states. The two rows of states correspond to the interval between stimulus and reward ("ISI") and the interval between trials ("ITI"). Transitions from ITI states back to the beginning are not shown. (b) Semi-Markov model. Events occur on entry to a state, and each state corresponds to the interval between a pair of events. The dwell time in each state varies according to some distribution (not pictured).

first change — partial observability [7] — is that the state s is only observable indirectly through an observation function $O(s)$, which maps s to a distribution over a set of observations $\mathcal{O}$. An observation $o_t \in \mathcal{O}$ may not uniquely identify the underlying state s_t. The second change is that the discrete temporal dynamics, in which the state advances with each "tick" of the clock, are replaced with semi-Markov dynamics [8], in which discrete state transitions take place in continuous time. A function $T(s)$ provides a distribution of the duration of a stay in state s, which allows for explicit modeling of variation in the time between events. As the time t is now continuous, we index the irregular but discrete state transitions by an integer n; e.g., if the process enters state s_n at time t then it enters state s_{n+1} at time $t + \tau_n$ where s_{n+1} is drawn from $Q(s_n)$ and τ_n is drawn from $T(s_n)$.

As a concrete example, figure 1 illustrates Markov and semi-Markov models of the conditioning task we study in this paper. The Markov model tracks the passage of time between events by a procession of intermediate hidden states; in contrast, each interval in the semi-Markov model corresponds to a single state, whose duration can vary continuously according to some distribution. While one can derive a discrete Markov approximation to a semi-Markov system by subdividing the states, the problems of inference about timing that we consider here will persist, albeit in more muddled form.

Our model differs slightly from traditional semi-Markov models with respect to rewards and observations, to better reflect the experimental situation. Traditionally, both occur ongoing during the duration of a stay at a state. We instead assume that all rewards drawn from $R(s_n)$ and observations from $O(s_n)$ are instantaneous and occur only at the moment that the state s_n is entered. Until the state is left, all rewards and observations are empty. So any reward or observation signals that a state change has occurred, but we also assume that state transitions can occur "silently," unsignaled by reward or observation. This feature requires inference to determine whether an unsignaled state transition has occurred.

Neglecting partial observability momentarily, we can redefine the value of a state s_n in the semi-Markov model as the discounted expected reward at the moment the state is entered. For bookkeeping reasons we omit r_n from this value, beginning the sum with r_{n+1}:

$$\begin{aligned} V(s_n) &= E\left[\sum_{N>n} \gamma^{\tau_n+\ldots+\tau_N} r_N\right] && (4) \\ &= E\left[\gamma^{\tau_n}(r_{n+1} + V(s_{n+1}))\right] && (5) \end{aligned}$$

where the expectation is now additionally taken over the dwell durations τ.

Approaches to partial observability in reinforcement learning divide according to whether they use models. Model-free approaches, used in previous TD models of the dopamine system, augment the observable state o_n with previously observed states o_{n-1} etc. in order to try to disambiguate the current hidden state. Unmodified TD for fully observable processes can be used to learn V as a function of the augmented state.

Model-based approaches [7] instead assume that the system learns or is given a *model* of the process — that is, the functions Q, O, R, and T. The value of each hidden state can be derived offline by solving the model (e.g. using value iteration), without any sample trajectories. While traversing the Markov process, the model can be used to compute a probability distribution over the hidden state given a series of observations, and the value is the expectation with respect to this distribution of the hidden state values.

We propose that the dopamine system mixes these model-based and model-free approaches, using methods similar to those of Chrisman [9]. In this proposal, the brain learns a model of the world, but rather than solving the model directly to obtain $\widehat{V}$, it uses the model only to estimate the hidden state during sample trajectories. From these samples, it uses TD to incrementally learn the values of the hidden states.

Were the states and transitions fully observable, we could update $\widehat{V}(s_n)$ at each transition analogously to equation 3:

$$\delta_n = \gamma^{\tau_n}[r_{n+1} + \widehat{V}(s_{n+1})] - \widehat{V}(s_n) \qquad (6)$$

which follows from equation 5 [8]. τ_n is the time since the last transition.

But since the states and transition times are only known probabilistically, through inference, we use a probabilistic form of equation 6, periodically updating each state's value proportionately to how strongly we believe the process has transitioned *out* of that state since the last update. We define $\beta_{s,t} = P(s_t = s, s_{t+\epsilon} \neq s | o_1 \ldots o_{t+\epsilon})$, the probability that the process transitioned out of state s between times t and $t+\epsilon$. Every ϵ timesteps, the model updates its beliefs about β, conditioning on the most recent observation using Bayes' rule:

$$\beta_{s,t} \propto P(o_{t+\epsilon}|s_t = s, s_{t+\epsilon} \neq s) \cdot P(s_t = s, s_{t+\epsilon} \neq s | o_1 \ldots o_t)$$

where the first term is computable from the current observation and the functions O and Q, and the second term follows recursively from its values at previous timesteps together with information about previous observations. (For a full treatment of inference in hidden semi-Markov models, see [10].)

With this information we can update the values. The change in $\widehat{V}(s)$ at time t is proportional to the error:

$$\delta_{s,t} = \beta_{s,t}(E\left[\gamma^{\tau}\right]\left[r_{t+\epsilon} + E\left[\widehat{V}(s_{t+\epsilon})\right]\right] - \widehat{V}(s)) \qquad (7)$$

where $E[\gamma^{\tau}]$ is the expected discounting given the observations and the hypothesis that the process has just transitioned out of s, which depends on the distribution of the system's likely dwell time in s; and $E\left[\widehat{V}(s_{t+\epsilon})\right] = \sum_{s' \neq s} \widehat{V}(s') \cdot P(s_{t+\epsilon} = s' | s_t = s, o_1 \ldots o_{t+\epsilon})$ is the expected value at time $t+\epsilon$, given the observations and the hypothesis that the process has just transitioned out of s.

Unlike the error signal of equation 3, this signal is vector-valued: $\widehat{V}(s)$ is updated differently for every s at every timestep. We could assume that different dopamine neurons code this vector in a distributed manner — which might explain why individual neurons differ in their firing properties even though under previous models they all report the same scalar quantity. Alternatively, we can assume dopamine neurons report a scalar approximation of this error signal: its average over all states weighted by $\beta_{s,t}$. Using β and this scalar signal, dopamine targets could apportion training between states. This approximation works well in our simulations because the distribution of the hidden state is generally sharply peaked; we use it in the results reported below.

To recap the structure of our model, the system is given (or learns, using methods outside the scope of this paper) a "world model" of how the hidden state evolves and produces observations. This model is periodically combined with new observations to update an estimate of β, the likelihood that each state has just been left. This estimate is used to perform a TD backup of information about received and predicted rewards to the likely predecessor states in order to learn the reward prediction function V.

3 Data

Schultz and collaborators have recorded dopamine neurons in primates performing tasks for reward (reviewed in [5]). The neurons' phasic responses share several properties with the TD error signal: They burst in response to unexpected reward. If a reward is *predicted* by a stimulus known to precede it by some fixed interval, the response transfers to the stimulus and there is no response to the reward. If a predicted reward is omitted, the neurons' background firing pauses briefly at the time the reward should have occurred.

These experiments are compatible with TD models using many timing schemes. In practice, most models augment the observable state with extra states corresponding to different fixed intervals after the occurrence of the reward-predicting stimulus [2, 3, 11]. If the reward occurs five seconds after the stimulus, the states corresponding to intervals shorter than five seconds would learn higher values (anticipating the reward) than the states corresponding to intervals longer than five seconds, and transitions from one set of states to the other produce changes in $\widehat{V}(s_t)$ which, together with the presence or absence of observed rewards r_t, produce the appropriate phasic events in δ_t of equation 3.

Further insight into how the system handles timing comes from experiments testing how the neurons treat *variation* in the stimulus-reward interval. In one experiment [12], animals were trained with a fixed stimulus-reward interval of one second, and dopamine neurons were recorded when the rewards were then delivered a half-second early or late. For early rewards, a burst occurred to the rewards, but the neurons did not then pause at the time the reward was originally expected. With late reward, a pause was seen at the time reward was expected, followed by a burst when the reward was delivered.

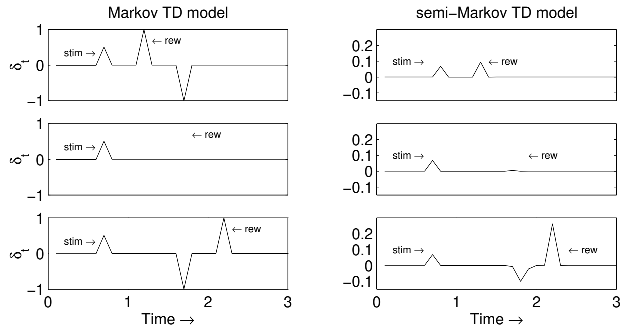

Figure 2: Modeled dopamine signals when reward delivery is varied away from an expected time. Left: In the TD model of Montague et al., early reward delivery (top) causes positive followed by negative TD error, reward delivery at the expected time (middle) produces no error, and late reward delivery (bottom) causes negative followed by positive error. Right: Our semi-Markov TD model performs similarly, except that early reward delivery produces no negative error, in accord with the data.

In another experiment [13], animals were trained on a variable delay of one to three seconds between stimulus and reward. Dopamine responses to the rewards persisted throughout extensive training, with stronger responses to earlier rewards.

As we will discuss in the next section, both of these results require some sophistication of a TD timing mechanism.

4 Results

Previous TD models of dopamine capture variation in the timing of reward by manipulating the probability of reward in each of a number of states representing different time intervals. But though these states should be coupled, they are all treated independently. For example, when early reward is delivered to the model of Montague et al. [2], no reward is expected during the state corresponding to that delay, so positive TD error results (Figure 2, top left). But the arrival of reward in the early state has no effect on the expectation of reward in the later state where reward usually arrives; the nonoccurence of reward *there* triggers negative TD error, predicting a pause in dopamine cell firing. No such pause is seen experimentally [12]. A previous attempt to correct this inconsistency [11] assumed that the receipt of a reward reset the representational system, clearing all predictions and thereby suppressing the pause in dopamine cell activity. Such an ad hoc device would not generalize appropriately; for instance, animals can learn to predict multiple rewards in sequence, an ability that

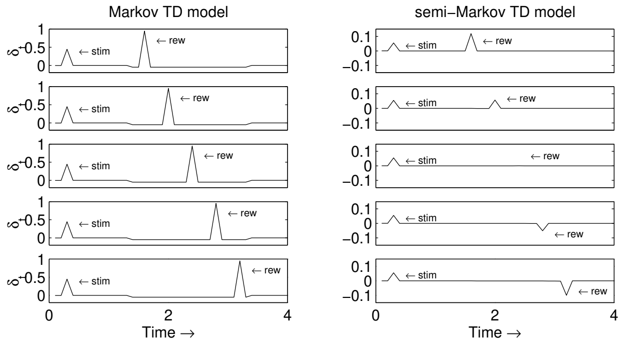

Figure 3: Modeled dopamine signals when reward timing varies uniformly. Left: In the TD model of Montague et al., the TD error does not vary with the stimulus-reward interval. Right: In our semi-Markov TD model, the modeled dopamine signal decreases as the stimulus-reward interval increases.

would be eliminated by the hypothesized reset.

The data are explained more elegantly under a model that properly treats variation in the timing of events. Under this account, the system infers that the early reward is the same reward that had been expected to arrive later, and thus does not expect it to arrive again. We demonstrate this idea using the two-state semi-Markov model shown in Figure 1 and the TD algorithm of equation 7. The ISI dwell duration is modeled by a lognormal distribution with the appropriate mean. (The distribution is sharply peaked, but allows for some uncertainty in the reward timing.) This model shows positive error to an early reward (Figure 2, top right); this is not because the reward is wholly unexpected but because it occurs *earlier* than usual and is thus worth *more* than usual due to the discounting by dwell duration in equation 7. Having received the reward, the model immediately enters the ITI state, and there is thus no error at the time when reward is normally delivered.

Both models behave similarly in the case of late reward. In the semi-Markov model, initial negative error occurs because as the interval without a reward becomes unusually long, the model infers that an unsignaled state transition from the ISI state to the ITI state has taken place without the expected reward. The negative error is smeared out in time since this inference occurs gradually, as probability mass leaks from one state into the other. The subsequent reward in the ITI state is then surprising, producing positive error.

Figure 3 compares the models when the stimulus-reward interval varies uniformly over an interval. For this, we replace the inference model's lognormal ISI distribution with a uniform one. Traces are shown for a number of trials, ordered by delay. In the Montague et al. [2] model, identical positive error is seen for all rewards, since reward is equally likely at all delays. Due to discounting, the error in the semi-Markov model depends on the timing of reward relative to the mean delay. For earlier-than-average reward delivery, positive error is seen, more for shorter ISIs. For later-than-average delivery, negative error occurs at the time of the reward. (In contrast to the late-reward condition of figure 2, phasic negative error is not seen at the *average* time of reward delivery because the inference model expects reward time to vary uniformly.) The increase in positive error with decrease in ISI is consistent with experimental results [13], though no reports have yet noted *pauses* in dopamine activity at the time of later-than-average rewards, as we predict here.

5 Discussion

We propose a combination of model-based and model-free reinforcement learning techniques, to model the dopamine system as a TD learner using an inferred representation of the hidden state of the world. We are mainly concerned here with inference about the timing of events; we use a semi-Markov model to capture variance in this timing. The main advantage of treating variance so explicitly is that it gives a clear, normative picture of how the TD state representation should evolve in these situations, which is easily contrasted with previous models. The model of Montague et al. [2] fails when reward timing varies because reward delivery does not affect its state representation at all; it consequently misses dependencies between states such as early reward arrival reducing the chance that the same reward is coming later. Suri and Schultz [11] address this by using a simple rule to adapt the state when reward arrives, but whether such a device approximates what a full model would infer depends on the situation.

Our combination of model-based and model-free techniques raises a question: Given a complete world model, why not solve it directly (e.g. using value iteration) rather than doing TD? One answer is that animals must learn world models online in nonstationary situations. We envision our system could learn its world models with online versions of hidden (semi-) Markov model learning, as have been used to model animal conditioning behavior [6]. In this setting, it might be infeasible to re-solve the model for $\widehat{V}$ at every model update, and instead sensible to learn $\widehat{V}$ incrementally in parallel using TD. TD might alternatively be necessitated by inadequacies in animals' state estimation systems. While we have assumed animals perform statistically valid state inference using a complete world model, more rudimentary inference using a simpler observation model could suffice. Such a system could accord with the thesis of this paper that dopamine activity reveals evidence of state inference, but would still require sample-based TD to learn to predict values.

A future direction is to connect this work with behavioral data on animal timing. The number of conditioning trials it takes animals to learn an association is invariant to dilations or contractions of the speed of events [14]. Such timescale invariance is difficult to capture using the discretely clocked temporal dynamics of a Markov TD model, but follows naturally from the event-driven transitions of the semi-Markov model presented here. Variance in animals' interval judgments follows a similar scalar property [15]; such estimation noise would be a fairly straightforward addition to the semi-Markov model.

A surprising prediction of this model is that later-than-average events can trigger a pause in dopamine neuron firing. This effect can occur not only to rewards (as in figure 3) but also to reward-predictive stimuli whose delivery time varies. (The traces depicted in figures 2 and 3 were taken after shorter-than-average intertrial intervals so the stimulus evoked positive error. Negative error would follow longer intertrial intervals.) These results suggest that dopamine responses to a reward predictive stimulus can eventually be trained away *on average*, even when the intertrial interval is randomized. Individual stimulus presentations could still evoke dopamine bursts or pauses, depending on presentation time relative to expectation, but in the aggregate they should nearly balance out. Dopamine responses to reward predictive stimuli do eventually disappear in overtrained animals [5], but the experiment contained only slight variability in the intertrial interval, making it difficult to judge the model's predictions.

Another experimental question raised by this research is whether a vector-valued TD error signal could account for observed variation in responses between dopamine neurons, with different neurons carrying error for different states the animal believes it might be in. Our vector and scalar models differ as to whether each state's value is updated in the direction of its own mismatch with expected reward, or whether all states' values are updated in the direction of the aggregate mismatch. It should be possible to experimentally distinguish these situations, by searching for dopamine activity in situations where positive and negative error for different states cancel out to produce zero aggregate error.

Acknowledgments

This work was supported by National Science Foundation grants IIS-9978403 and DGE-9987588. Aaron Courville was funded in part by a Canadian NSERC PGS B fellowship.

References

[1] J. C. Houk, J. L. Adams, and A. G. Barto, "A model of how the basal ganglia generate and use neural signals that predict reinforcement," in *Models of Information Processing in the Basal Ganglia* (J. C. Houk, J. L. Davis, and D. G. Beiser, eds.), pp. 249–270, Cambridge, Mass.: MIT Press, 1995.

[2] P. R. Montague, P. Dayan, and T. J. Sejnowski, "A framework for mesencephalic dopamine systems based on predictive Hebbian learning," *Journal of Neuroscience*, vol. 16, pp. 1936–1947, 1996.

[3] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, pp. 1593–1599, 1997.

[4] R. S. Sutton, "Learning to predict by the method of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, 1988.

[5] W. Schultz, "Predictive reward signal of dopamine neurons," *Journal of Neurophysiology*, vol. 80, pp. 1–27, 1998.

[6] A. C. Courville and D. S. Touretzky, "Modeling temporal structure in classical conditioning," in *Advances in Neural Information Processing Systems 14* (T. Dietterich, S. Becker, and Z. Ghahramani, eds.), (Cambridge, MA), MIT Press, 2001. (in press).

[7] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, pp. 99–134, 1998.

[8] S. J. Bradtke and M. O. Duff, "Reinforcement learning methods for continuous-time Markov Decision Problems," in *Advances in Neural Information Processing Systems 7* (G. Tesauro, D. S. Touretzky, and T. K. Leen, eds.), (Cambridge, MA), pp. 393–400, MIT Press, 1995.

[9] L. Chrisman, "Reinforcement learning with perceptual aliasing: The perceptual distinctions approach," in *National Conference on Artificial Intelligence*, pp. 183–188, 1992.

[10] Y. Guedon and C. Cocozza-Thivent, "Explicit state occupancy modelling by hidden semi-Markov models: Application of Derin's scheme," *Computer Speech and Language*, vol. 4, pp. 167–192, 1990.

[11] R. E. Suri and W. Schultz, "A neural network with dopamine-like reinforcement signal that learns a spatial delayed response task," *Neuroscience*, vol. 91, pp. 871–890, 1999.

[12] J. R. Hollerman and W. Schultz, "Dopamine neurons report an error in the temporal prediction of reward during learning," *Nature Neuroscience*, vol. 1, pp. 304–309, 1998.

[13] C. D. Fiorillo and W. Schultz, "The reward responses of dopamine neurons persist when prediction of reward is probabilistic with respect to time or occurrence," in *Society for Neuroscience Abstracts*, vol. 27, p. 827.5, 2001.

[14] C. R. Gallistel and J. Gibbon, "Time, rate and conditioning," *Psychological Review*, vol. 107, no. 2, pp. 289–344, 2000.

[15] J. Gibbon, "Scalar expectancy theory and Weber's law in animal timing," *Psychological Review*, vol. 84, pp. 279–325, 1977.