Running heading (shortened title): Modeling dopamine activity
Article type: Artificial Intelligence
Title: Modeling dopamine activity by Reinforcement Learning methods: implications from two recent models
Author(s): Patrick Horgan & Fred Cummins

Affiliation & full address for correspondence, including telephone and fax number and email address:

Affiliation: Patrick Horgan Formerly, UCD School of Computer Science and Informatics, University College Dublin Belfield, Dublin 4, Ireland
(E- mail: paddy.horgan@gmail.com)

Full address for correspondence: Patrick Horgan Neuroscience and Psychiatry Unit, G.907 Stopford Building, University of Manchester, Oxford Road, Manchester, M13 9PT. UK (E- mail: paddy.horgan@gmail.com)

Telephone: + 44 161 275 7427
Fax: + 44 161 275 7429

Affiliation & full address for correspondence: Fred Cummins UCD School of Computer Science and Informatics, University College Dublin Belfield, Dublin 4, Ireland (E-mail: fred.cummins@ucd.ie)
Telephone: + 353 1 716 2902
Fax: + 353 1 269 7262

Modeling Dopamine Activity by Reinforcement Learning Methods: Implications from Two Recent Models

PATRICK HORGAN[1] & FRED CUMMINS[2]
[1]Formerly, UCD School of Computer Science and Informatics, University College Dublin Belfield, Dublin 4, Ireland (E-mail: paddy.horgan@gmail.com); [2]UCD School of Computer Science and Informatics, University College Dublin Belfield, Dublin 4, Ireland (E-mail: fred.cummins@ucd.ie)

Abstract. We compare and contrast two recent computational models of dopamine activity in the human central nervous system at the level of single cells. Both models implement reinforcement learning using the method of temporal differences. To address drawbacks with earlier models, both models employ internal models. The principal difference between the internal models lies in the degree to which they implement the properties of the environment. One employs a partially observable semi-Markov environment; the other uses a form of transition matrix in an iterative manner to generate the sum of future predictions. We show that the internal models employ fundamentally different assumptions and that the assumptions are problematic in each case. Both models lack specification regarding their biological implementation to different degrees. In addition, the model employing the partially observable semi-Markov environment seems to have redundant features. In contrast, the alternate model appears to lack generalizability.

Keywords: computational, dopamine, learning, model, reinforcement

1. Introduction

Reinforcement learning methods involving the Temporal Difference (TD) algorithm have been widely used to model the activity of dopamine neurons in animals undergoing various conditioning experiments (Wörgötter and Porr 2005). An element of reinforcement learning systems at times employed is a model of the environment - this is something that mimics the behaviour of the environment (Sutton and Barto 1998). In this article, we examine two recent reinforcement learning models involving the TD algorithm that use internal models (Daw et al. 2006; Suri 2001). Both of these models have been applied to model the activity of dopamine neurons. A comparison is made to help analyze how these models differ and understand the features of internal models in these kinds of modeling exercises.

2. Dopamine Function

Dopamine neurons seem to play an important role in movement and behaviour in animals. The importance of dopamine in the function of motor control in humans is

seen from movement disorders such as Parkinson's disease (Samii et al. 2004). The role of dopamine in behaviour is complex and is considered further below. As shown in Figure 1, dopamine neuron pathways are commonly divided into three main groups (although more complex accounts are also given such as in Haber et al. (2000)). The nigrostriatal pathway goes from the substantia nigra in the midbrain to the dorsal aspect of the striatum (Crossman and Neary 2000). The mesostriatal pathway extends from the ventral tegmental area of the midbrain to the ventral part of the striatum (Crossman and Neary 2000). A third pathway extends from the ventral tegmental area to the prefrontal area of the brain constituting the mesocortical pathway (Fuxe et al. 1974; Kelley et al. 2005). The nigrostriatal pathway seems particularly important in movement (Samii et al. 2004) but the relationship between behaviour and individual pathways is complex (Schultz 1998).

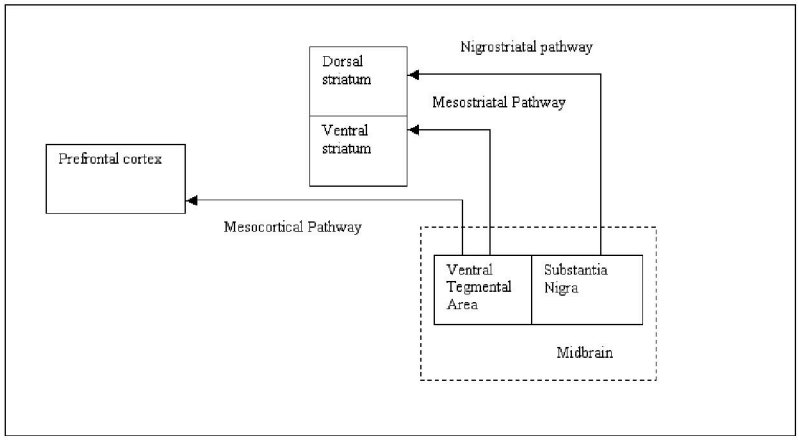

Figure 1. Outline of three main dopamine pathways in the human brain.

The role of dopamine in behaviour has been investigated in animal studies using stimulus response paradigms. Dopamine seems to be important in how reward acts as reinforcement for these associations (Wise 2006). Animals pretreated with dopamine receptor antagonists require more time to learn to lever press for food in comparison with normal animals (Wise and Schwartz 1981). Furthermore, these effects seem unrelated to the role that dopamine receptor blockers may have on motor function (Wise 2006). However, although dopamine seems to have an important role in learning these associations, recent work on genetically engineered mice indicate that dopamine is not a necessary condition for learning to occur (Cannon and Palmiter 2003).

A conceptualization of the role of dopamine in forming associations during classical conditioning experiments has been done using a reward prediction paradigm. According to this theory, a discrepancy is required between predicted and actual

reward for an animal to learn to associate a stimulus with a reward. This discrepancy is termed a reward prediction error (Waelti et al. 2001; Schultz 2000).

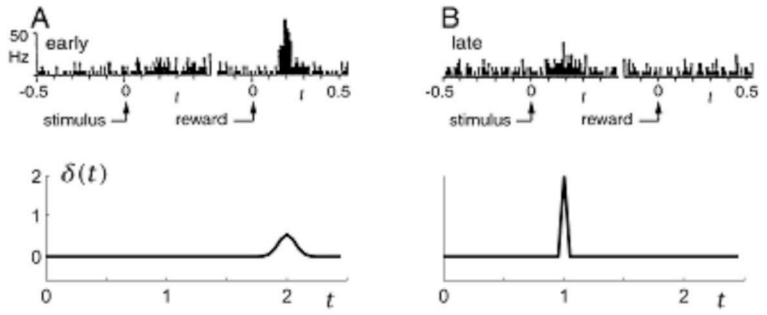

Figure 2. Histogram of the activity of a dopamine neuron (above). Representation of the corresponding Temporal Difference (δ (t)) error signals (below) - these are the signals greater than baseline in the lower diagram. (Adapted with permission from Kakade and Dayan 2002)[1]. In this figure, the Temporal Difference (TD) error term is generated using reinforcement learning algorithm methods. The TD error term roughly corresponds to the reward prediction error term described in section 2. In situation A (early learning) a dopamine neuron responds to the delivery of reward but not to the presentation of the stimulus that predicts reward. This is matched by the TD error signal (δ (t)). After an animal learns to pair a stimulus with a reward, (situation B), a dopamine cell responds to the delivery of stimulus but not to the reward. This is again similar to the TD error signal (δ (t)).

This can be seen when pairing a stimulus (sound of a bell) with a reward such as food (Schultz 1998). In early training, the reward is not predicted by the stimulus so a reward prediction error occurs at the time of reward. After training, the reward is predicted by the stimulus – the animal "knows" food will occur - hence no reward prediction signal occurs at the time of reward but instead occurs at the time of the stimulus. This reflects the fact that the stimulus acts as a reward prediction error signal for yet another earlier stimulus. In addition, as detailed by Schultz (1998), the error signal is represented by transient increased dopamine activity in the dopamine pathways. As described later in this article, it has also been proposed that the reward prediction error term corresponds to a Temporal Difference error signal generated using a reinforcement learning computational algorithm (Montague et al. 1996). An outline of this is given in Figure 2.

It should be noted that there are a number of other views about the role of dopamine. Wise (2006) argues that there are some theoretical difficulties with the experiments of Schultz such as the distinction between primary rewards and reward-predictors. For example Wise argues that most food is not primarily rewarding. Wise (2006) takes a broader view of the role of dopamine suggesting that it "stamps in" associations between a stimulus and a reward. In contrast, Berridge and Robinson (1998) argue that roughly, the function of dopamine is to transform a stimulus from something that is liked into some thing that the animal wants – that is, something that the animal will work to acquire. Berridge and Robinson (1998) address the experimental finding (Young et al. 1993) that unpleasant events may increase the

levels of dopamine. A suggestion is made that in some circumstances, instead of dopamine production being linked to desirability the reverse may occur in which a negative valence is attached to a stimulus interpreted as threatening. As can be seen from these authors, there is a level of disagreement about the specific role of dopamine. This disagreement may account in part for divergence between the internal models as discussed below.

3. Reinforcement Learning

Reinforcement Learning methods form part of a larger group of methods that examine learning in computers in general (Russell and Norvig 1995; Mitchell 1997; Sutton and Barto 1998) such as supervised learning and unsupervised learning techniques. Supervised learning is characterized by using an external "teacher" to generate needed input output relationships which the network learns. Unsupervised learning depends on self-organization based on inputs alone. Reinforcement learning has been described as intermediate between these forms of learning (Dayan and Abbott 2001).

The concepts of an agent and a state are important in the understanding of the principles of reinforcement learning. An agent here is regarded as something which learns to interact with the environment to achieve a goal (Sutton and Barto 1998). A state here is the representation that an agent gets of the environment's state (Sutton and Barto 1998). This is a flexible approach and allows the state to be determined in different ways. For example the state could be made up of direct sensations received by the agent or alternatively by the memory of past sensations. In addition to these concepts**,** Sutton and Barto (1998) describe four main components to a Reinforcement Learning system (Sutton and Barto 1998).

- A policy defines the way an agent behaves at a particular time. Roughly, this can be understood as a mapping from the perceived state of an agent to an action to be taken in those states.
- The reward function gives a definition of what are good and bad events for an agent. In a biological system, reward as a good event could be pleasure and a bad event could be pain. The agent is unable to alter the reward function but can use it as the basis for generating its policy.
- The third component is the value function: it describes what is good for an agent over a sustained period of time. Roughly speaking, this represents the cumulative reward that an agent can expect starting from a state.
- The final element is the use of a model of the environment. This is sometimes called an internal model. This occurs only in some reinforcement systems. A model is anything that an agent can use to predict how the environment will respond to its actions. So a model of an environment needs to have significant predictive capabilities.

The value function is a kind of prediction from a state. The notation $V(s_t)$ stands for the estimate of the expected return (value function) for state s at time t (this is understood to be following some policy). So, given a non-terminal state s at time t, a

method needs to be developed to estimate the value of it, i.e. estimate $V(s_t)$. One way of doing this is by waiting until the agent reaches some defined terminal state and then examining how useful s_t was in reaching the terminal state. This can be done repeatedly to get an accurate value for $V(s_t)$ A different approach is to update the value function at the following time step rather than waiting until the terminal state has been reached. At the following time step, the agent will have a value estimate for that time-step and the environment will provide the reward. This allows successive predictions to be compared. The discrepancy forms the TD error signal.

This can be summarized in the following equation (Sutton and Barto 1998).

$$V(s_t) \leftarrow V(s_t) + \alpha [r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (1)$$

In this equation $V(s_t)$ represents the value of the state at time t, $V(s_{t+1})$ represents the value of the state at time t+1 and r_{t+1} is the reward at time t+1. The expression $[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$ – the TD error signal (often termed δ_t) - is used to update the value function. α represents a constant step size parameter to allow gradual updating. γ represents a discount factor which is used to indicate that later rewards are worth less than earlier rewards. Thus the equation for δ_t can be written

$$\delta_t = [r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (2)$$

This allows equation (1) to be written as

$$V(s_t) \leftarrow V(s_t) + \alpha [\delta_t] \quad (3)$$

Roughly, these equations are implemented as follows. At the start, values are assigned in an arbitrary manner giving a baseline (V(s)) for all the states. Starting with a state at time t (s_t) an action is taken leading to the state at the next time step (s_{t+1}). The environment will provide the reward r_{t+1} at that time step. This could be equal to zero if no reward occurs. As values have been assigned to all states at the outset, so $V(s_t)$, $V(s_{t+1})$ will have been generated in addition to r_{t+1} generated by the environment. This allows $V(s_t)$ (the value of a state at time t) to be updated at this time step using equation (3) rather than having to wait until the terminal state. The process is repeated at each time step until the terminal state is reached indicating the completion of that episode. An episode here is any sort of repeated interaction with the environment ending in a terminal state. Once that state is reached there is a return to a starting state. An episode could be for example a trip through a maze. Episodes are repeated until some defined condition is fulfilled such as for example the identification of a particular route through a maze.

Although value functions generated using the TD algorithm has some predictive qualities a distinction can be drawn between them and models of the environment. A value function can be used to generate the sum of future rewards given a state. However, given a state and an action, a model of the environment allows the generation of the next state and reward. Thus a model of the environment provides a richer form of prediction in comparison to a value function. The advantage of these

models is that they can be used to extend the TD algorithm thereby improving the performance of systems in certain situations (Suri 2002).

4. Modeling of Dopamine Activity Using Reinforcement Learning Methods

As described earlier, various experiments (Schultz 1998) have shown that the activity of dopamine neurons changes while an animal learns to associate a stimulus with a reward. This form of activity has been modeled using Reinforcement Learning methods – described in (Schultz et al 1997; Daw 2003; Suri 2002; Joel et al. 2002). There are a number of variations in these models but some of the core ideas are as follows.

1. The Reinforcement Learning methods often use TD learning methods.
2. The TD error signal represents the reward prediction error signal, so the TD error is also reflected by the transient increased dopamine activity (see Figure 2).
3. The models have a means of representing the stimulus at the time of the reward.
4. The environment is Markovian: roughly this means that the current state retains all relevant information from previous states and that future states and associated relevant information such as rewards can be estimated from the current state (Sutton and Barto 1998; Montague et al. 1996).
5. Other components of the model have biological correlates.

A number of such models are discussed by (Joel et al. 2002). They note that a weakness of these models is that the biological implementation of these models often does not correlate with known anatomy and physiology.

4.1 *Model by Suri (2001)*

Further developments have been made to improve the correspondence of the TD error signal with the dopamine activity in biological experiments. Suri (2001) describes the activity pattern of dopamine neurons in animals undergoing the sensory preconditioning paradigm. In this kind of experiment, an animal undergoes a phase of training where a neutral stimulus A precedes a neutral stimulus B. In the second phase, the animal learns to associate stimulus B with a reward. In the third phase stimulus A is presented alone. The response of the animal to stimulus A in the third phase is similar to the response to the reward. This seems to indicate that given stimulus A, the reward is predicted – though A has never been explicitly paired with the reward. The animal seems to have learned that stimulus A is followed by stimulus B and the same Stimulus B then precedes a reward. Suri (2001) notes how this has been shown to be reflected by changes in dopamine concentration (Young et al. 1998).

As previous Reinforcement Learning methods using the TD error term failed to describe this activity accurately, Suri (2001) gives a description of a new model drawing from (Sutton and Pinette 1985) to achieve this task. The main difference between it and previous models is that the prediction of future events is generated in a different way. Previous models such as (Montague et al. 1996), used an estimate of future rewards by a value function such as

$$V(x(t)) = x(t).\, w(t) \tag{4}$$

Where V (x (t)) is a scalar estimate of future rewards, x (t) is a vector representing the stimulus at time t and w (t) is a vector representing the weights of a neural network. The weights of the neural network are altered by the TD error to give a more accurate value for the value function during learning.

In the model by Suri (2001) a prediction signal p(t+1) is estimated. This represents the discounted sum of future events (rewards and stimuli), a concept broader than the value function. It is estimated using equation

$$p(t+1) = x(t+1) + Wx(t+1) + W^2x(t+1) \tag{5}$$

In this formulation, a number of different events are represented in an event vector. For example, if the presence of an event is represented as 1 and the absence represented as 0 then the event vector [0 1 0] represents the presence of one event and the absence of two other events. In Suri's model, each event in an event vector is transformed into a fixed temporal pattern over some time period which is called a temporal event representation. x(t+1) is the temporal representation of the events at time t+1 and x(t+2) is the temporal representation of the events at time t+2. In this theory, W represents the weight matrix of a neural network - a form of transition matrix of a Markov process. W has the property

$$W(x(t)) \cong \gamma x(t+1)\,. \tag{6}$$

W can be seen to be a fraction (by discounting factor γ) of a transition matrix of a Markov process. As described by Sutton and Pinette (1985), the use of this kind of weight matrix is needed to allow a system to converge. In this model by Suri, given the temporal representation of events at time t+1, the weight matrix W can be used to calculate the temporal representation of the events at the next time step t+2. Using W recursively allows the calculation of the temporal representation of events after 2 time steps. Both of these calculations include the discount factor γ. In theory, the matrix W could be used to generate calculation of the temporal representation of events at even later time steps. However, temporal representations of events beyond two time steps are less important because of the discounting factor γ. In equation (5) the temporal representation of the events at time t+1 is indicated by term x(t+1), the temporal representation of the events at time t+2 is indicated by Wx(t+1) and likewise the temporal representation of the events at time t+3 is indicated by $W^2x(t+1)$. Suri

(2001) argues that the summation of these temporal representations of events as in equation (5), calculated using the matrix W, generates an estimate of p(t+1) - the sum of future events from time-step t+1. This allows an estimation of what events will happen in the future. The network learns the matrix W by using TD error techniques. By being able to predict the sum of future events through the use of W, it can be seen that W is involved in the generation of a model of the environment. In comparison, other models (as in the model of (Montague et al. 1996) mentioned above) use a matrix w(t) which generates the sum of future rewards at each time step. This is just an idea of how good or bad things will turn out from that time step; it does not have the richness of a prediction in comparison to p(t+1) generated using W.

The advantage of this model is that enhanced predictions are made. Thus the system is better able to detect when a prediction fails to occur. As a result it has an improved ability to detect prediction errors in comparison to using a value function. The prediction errors are used to update the matrix W (as above) using the TD algorithm. The prediction errors are more comprehensive than that achieved using a value function and so correlate more closely to the activity of dopamine neurons. The closer correlation to the activity of dopamine neuron activity seems to be achieved largely by the use of a model of the environment.

The biological implementation of this model is briefly outlined by Suri (2001). The prediction signal may be generated by cortical neurons. It is suggested that the formation of the associations occurs within the hippocampal area. The TD error signals - as outlined above - are reflected by dopamine cell activity. A time step of the model is related to the theta cycle of the hippocampus.

4.2 *Model by Daw et al. (2006)*

Daw et al. (2006, 2003) developed a model to address another situation where the previous models produced incorrect correlations between dopamine activity and the TD error of the Reinforcement Learning methods. This is the case (Hollerman and Schultz 1998) where animals were trained to expect a constant stimulus-reward interval which was later varied – as when a reward is given earlier than expected. For example, the model by (Montague et al. 1996) predicted that at the earlier than expected presentation, a positive error (representing increased dopamine activity) should occur. However, this model also predicted that there should be a further episode of negative error (decreased dopamine activity) at the originally expected time of delivery as shown in Figure 3. This model only partially corresponds to the activity of dopamine neurons. From experimental data it has been shown (Hollerman and Schultz 1998) that there is an increase in the dopamine activity at the earlier time but there is not a decrease at the original time of reward as can be seen in Figure 4.

Daw et al. (2006), argue that this error arises because of the way earlier models assume the environment is Markovian. Roughly, in these accounts it is assumed that a state retains all relevant information from previous states and that future states and associated relevant information such as rewards can be estimated (Sutton and Barto 1998; Montague et al. 1996). There are situations where this does not appear the case such as in trace conditioning experiments. For simplicity, view a trace conditioning experiment as learning to pair a stimulus with a reward with a time gap between the

offset of the stimulus and the onset of reward. After learning, a stimulus predicts a reward after a delay of a few seconds. Assume that the state is an animal's immediate sensory observation (Daw, 2003). Note that in this kind of experiment, what is observable immediately before the reward is much the same as that observable after the reward. The state is the same in both situations but clearly the chance of receiving a reward is not. In this case, ostensibly there are two states (as estimated from the observations) that are the same but they have different future outcomes so violating the Markov property (Daw, 2003).

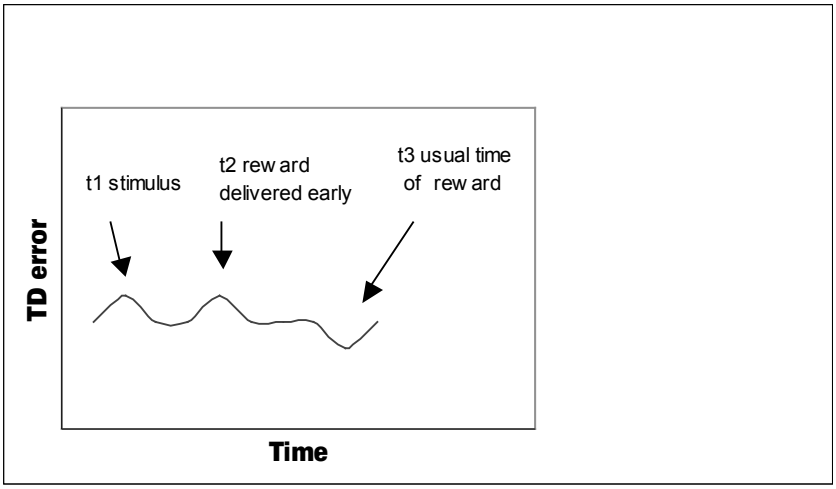

Figure 3. Figure based on the computational model in (Montague et al. 1996). Representation of TD error when reward delivered earlier than expected. Positive TD error occurs at time of stimulus (t1). Positive TD error occurs at time of earlier reward (t2). Negative TD error at time of originally expected reward (t3).

In addition, the use of a Markovian environment has certain drawbacks where there is variability in the timing between stimuli and rewards in trace conditioning experiments. Daw et al. (2006) suggest that in these circumstances it may be useful to consider the environment as having a semi-Markov property. The key difference between this model and a Markov model is that state transitions can occur at irregular time intervals in the semi-Markov model (Daw 2003). In contrast, the Markov model requires that the intervals between state transitions remain the same. Using the semi-Markov model of the environment allows a trace conditioning experiment to be represented as two states: one state corresponding to the interval between stimulus and reward (called interstimulus interval (ISI)), the other state corresponding to the interval between reward and stimulus (called intertrial interval (ITI)). According to this model an external event signals a transition from one state to another. Hence when an agent is in the ISI state and a reward occurs then a transition to ITI state occurs. This means that the situation in Figure 3 could not occur. Once the reward is delivered in these circumstances, the agent is in the ITI state and so no longer expects

a reward hence there is no error term at the usual time of reward. So the semi-Markov model allows better representation of experiments with variability in timing.

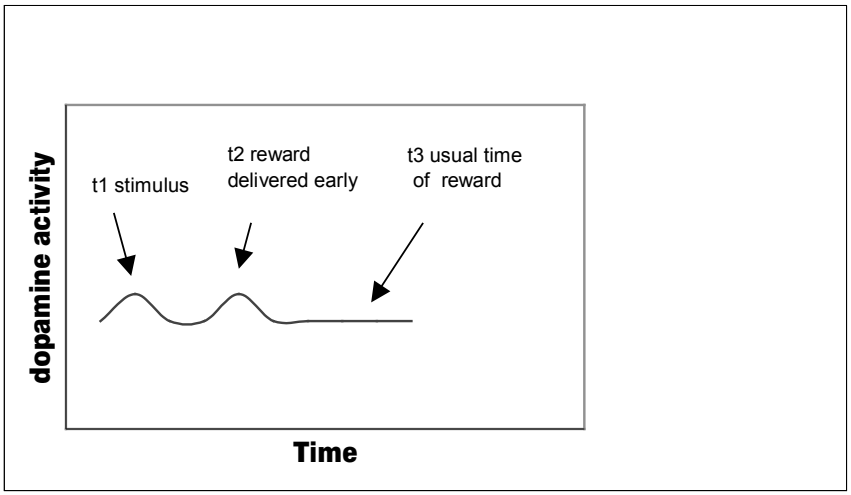

Figure 4. Figure based on the neurophysiological experiments of (Hollerman and Schultz 1998). Representation of dopamine activity when reward delivered earlier than usual. There is increased dopamine activity at time of stimulus (t1). There is increased dopamine activity at time of reward (t2) when reward delivered earlier than expected. No altered dopamine activity at time of originally expected reward (t3).

Just using a semi-Markov model presents difficulties when expected rewards fail to occur. In this case there would be no signal that a transition from ISI to ITI has occurred. So the agent would remain in the ISI state. To overcome this case as well as addressing theoretical issues involving the concept of state, as outlined earlier, Daw et al. (2006) propose an environment modeled using a partially–observable Markov model. This means that the state is not an agent's immediate sensory observation but rather that the observation received by the agent is related by a probability function to the underlying state. There are a number of inference components to this model. One of these inference components is a belief state $\beta_{s,t}$: the agent tries to figure out the probability that it left state s at a given time t given its observations to that point. For example, if the value for the belief state is high then it is likely that a transition from a state has occurred. This can be used to bias the error term used for learning. So if $\beta_{s,t}$ is low then less weight is given to the error term used for learning, whereas if $\beta_{s,t}$ is high then more weight is given to the error term. In addition, other inferences such as the dwell time in a state are computed in the model. These processes can be used to explain the situation when a reward is omitted. In this case, the agent is in an ISI state awaiting a reward. As the interval following a stimulus lengthens the state inference mechanisms gradually decide that a state transition has taken place and that now the animal is in an ITI state. If the reward does occur later this further convinces the agent that the transition has occurred and that it

is now in the ITI state. Thus the partially observable semi-Markov model enables more accurate representation when a reward is omitted than just using the semi-Markov model alone (see Figure 5).

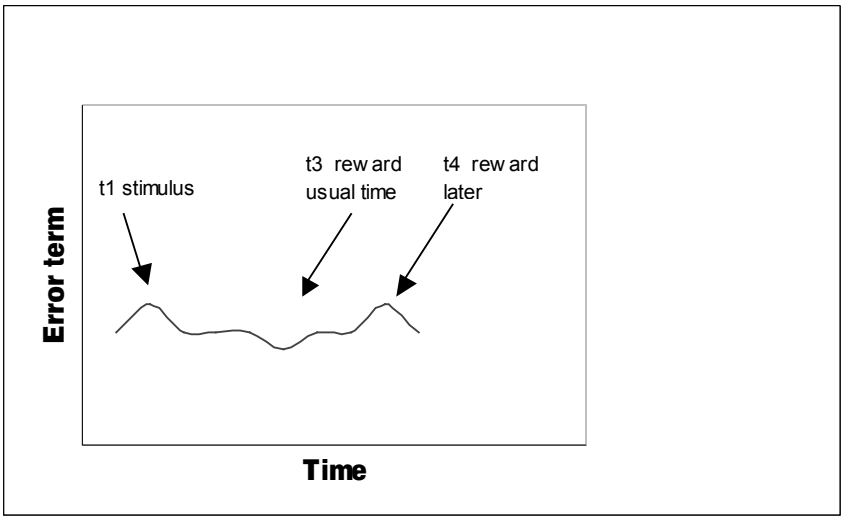

Figure 5. Figure based on the computational model of Daw et al. (2003). This is a representation of an error term when reward is omitted at the usual time and presented at a later time. A positive error term occurs at time of stimulus (t1). There is a negative error term at the time of the omitted reward (t3) and a positive error term at the time of the later arriving reward (t4). In the case where a semi-Markov model without partial observability is used, then the agent would continue to expect the occurrence of a reward indefinitely. So the agent would remain in the interstimulus interval (ISI) state and would not transition to the intertrial interval (ITI) state.

In this account it seems that the inference system for the partial observability aspect of the system makes up the internal model in the system. For example, this is the part of the system that is able to figure out whether a transition has occurred given the information to that point. This is more complex kind of prediction than the sum of future rewards for a state provided by using the TD algorithm. The advantage of this approach it can be used to model a number of different experiments. In comparison earlier models such as (Montague et al. 1996) that do not use the partially observable Markov model can only be applied to a narrower range of experiments so lacking a capacity to generalize.

It is argued by Daw et al. that their model requires a cortical perceptual system to construct the inferences as required by the partially observable Markov aspect of the model. Daw et al. assume that the animal has learned the mechanisms for making inferences. Although the model of Daw et al. predicts the activity of dopamine neurons more accurately, the biological correlates are not outlined in detail. A suggestion is made that the inferences system may be located in the sensory cortex

which is then received by the dopamine system. Dopamine activity is correlated with the TD error term in the part of the model employing this algorithm.

5. Comparison of Models by Daw et al. and Suri

The internal models of Daw et al. (2006) and Suri (2001) contrast in a number of ways. These differences are summarized in Table 1. The generation of the internal model involves fundamentally different assumptions. In the system of Daw et al., a partially observable semi-Markov environment is assumed thus inferences need to be generated and used for this model. In Suri's account a process of iteration using a weight matrix generates an estimation of the prediction of the sum of future events. The role of the internal models is different in these two systems. In the account of Daw et al., it is the inference system for the partial observability aspect of the system that seems to constitute the internal model. For the rest of this article, the view will be taken that the system of Daw et al. constitutes all the components of the model whereas the internal model corresponds to the inference component of the system. In Suri's account the weight matrix forms the key component in the generation of the model of the environment.

One of the main deficiencies in modeling using reinforcement learning methods is that the biological implementation is often not fully specified (Joel et al. 2002). This seems particularly apparent in the more complex approaches using internal models. Owing to their different implementation of internal models, this manifests itself in different ways in the two accounts. A key requirement in the generation of a model of the environment in Suri's account is the use of a weight matrix. This seems less complex to implement biologically than Daw et al.'s internal model and so a reasonable attempt is made by Suri to identify biological correlates of the components of the model. However, a difficulty with this account is that it is not clear whether there are appropriate anatomical pathways between the structures for the computation to take place as required for the internal model to function. Without this specification, it is difficult to judge whether the biological structures could operate with one another in a plausible integrated manner. So by Suri's account, although biological structures are identified to correlate with features of the model, it is unclear if the model as a whole can operate in a biologically plausible manner. For the internal model of Daw et al., inference states needs to be generated by a number of different functions. Owing to the complexity of the model it is difficult to relate it to biological structures so only a sparse account of the biological implementation can be given (Daw 2003; Daw et al. 2003; Daw et al 2006). This results in reservations about the biological feasibility of the model.

In the system (all components of the model) of Daw et al., the issue of redundancy is acknowledged by the authors. In this system, predictions are generated both by the functions required to generate inferences using the internal model and less broadly, by means of a value function generated using TD methods. In theory, predictions could be made solely by the functions used in the internal model. This seems to indicate that the TD error aspect in their system is not required. As one of the main goals of this type of modeling is to capture the manner in which dopamine neuron activity is

paralleled by variation in the TD error signal, this finding seems to undermine the goal of understanding the function of dopamine activity.

Suri's model doesn't appear to have the difficulty with redundancy as only one form of prediction is made (the sum of future events – p(t+1)). However, Suri admits that his model cannot be used for modeling of certain experiments. Thus it lacks an ability to generalize to different kinds of experiments. Bearing in mind that approaches without the use of an internal model have modeled a number of different experiments, this lack of generalizability in Suri's account may suggest his account of an internal model may not be required to model neurophysiological experiments.

Daw et al. (2005) offer a suggestion to address the redundancy issue which also may help with the problem about lack of generalizability. An argument is made that there are two distinct neural systems involved in prediction. One system employs a model-based approach and is situated in the prefrontal cortex. The other system employs a model free approach and is situated in the dorsolateral striatum. Depending on the cognitive task involved, a system of arbitration is employed to use the more appropriate model. By this account, an internal model is required only in certain tasks and that model free approaches are sufficient in other tasks. The choice between them is made in terms of demands made by different systems in terms of memory and time. In the system of Daw et al. (2006), the inference model would constitute the internal model and value function using the TD error would make up the model free approach. The account by Suri provides the framework for the internal model. So the redundancy issue is addressed because though there are two systems each is utilized differently. The lack of generalizability is addressed as it is not a requirement that all experiments need to be modeled by the approaches using the internal model.

The suggestion outlined in (Daw et al. 2005), helps resolve some of the issues about the use of the internal models. However it does not address the difficulty about the lack of biological implementation of these models. Indeed the solution in (Daw et al. 2005) requires a further system of arbitration thus increasing the complexity of the system and increasing the difficulty of biological implementation of the model.

The models differ in their account of how the models are taught. Suri outlines that the TD learning rule is implemented to teach the system a weight matrix that is used to generate a model of the environment. The prediction error in the TD algorithm is related to dopamine neuron activity in this account giving an indication of the biological structures involved in teaching the model of the environment. In comparison, Daw et al. (2006) do not outline how the model is taught. It would be of interest to see whether the inference component of their model could be taught using a TD algorithm. This could give an indication of whether a biological structure such as the dopamine neurons could be involved in teaching the model of the environment.

It seems that both of these models are involved in generating broader predictions than would be the case were the TD algorithm employed alone – this has implications about the nature of dopamine neuron activity assumed in each model. By Suri's account, the use of the model of the environment allows the prediction of the sum of future events. This includes neutral stimuli as well as rewards. This may mean that dopamine is implicated in the process of prediction in a broader sense rather than only the prediction of reward related information. This is an area under discussion as investigators differ in their view about the role of dopamine in reward. Some emphasize the role of dopamine in reward (Wise 2006) whereas other authors

question this view because of the potential role of dopamine in unpleasant stimuli such as footshock (Salamone et al. 2005; Young et al. 1993). The former view may suggest a narrower view of the role of dopamine in prediction whereas the latter accounts may suggest a broader role for dopamine in prediction. Suri's internal model seems more compatible with this latter view.

In comparison Daw et al. (2006, p.1667) suggest that according to their view the TD system receives a "refined, inferred sensory representation from cortex". This seems to suggest that the internal model (inference component of their system) is able to generate predictive information, so the TD system may have a reduced role in prediction. This may mean that the dopamine system may not need to have as important a role in prediction compared to Suri's model. Thus the model of Daw et al. could be consistent with a greater number of theoretical views of the dopamine system in comparison with Suri's account. This may also help to explain why this system seems to be more generalizable compared to the model of Suri.

The two accounts also differ regarding the proofs of their accounts. A proof of the model of Daw et al. is provided in the appendix of their paper (Daw et al. 2006). In contrast, Suri is unable to offer a convergence proof for his model. In addition, he notes for some simulations altering parameters prevented convergence during the simulation. Thus the model of Daw et al. may have a more robust theoretical framework.

Table 1. Comparison of the internal models of Daw et al. (2006) and Suri (2001).

	Daw et al.	Suri
Learning of model	not specified	TD algorithm utilized
Information generated	inference of a state	prediction of the sum of future events
Biological implementation	suggestion of how could be implemented	brief outline
Generalizability	applicable to a number of different experiments	doesn't apply to experiment when reward delivered earlier than expected
Proof for model	proof sketched	unable to provide convergence proof

6. Conclusion

The comparison of the internal models of Daw et al. and Suri has indicated how they differ in a number of features. The identification of some of the important features that occur in this kind of modeling seems to be one useful outcome of the comparison. These particular models are interesting to contrast as they take two clearly different approaches. Suri attempts to use a fairly tightly integrated model that seems biologically plausible but has the drawback that it does not generalize to a wide number of experiments. Daw et al. take a more abstract approach involving a more complex system. The internal model forms a component of the larger system. The approach allows the system to be compatible with a larger number of experiments but

it is more difficult to relate this model to biological structures. These models suggest a tension between increasing the theoretical strength of a model with the practical issue of how it is implemented. Suri's model may be more practical but it is unable to cope with some of the theoretical implications raised in the experiments. The model of Daw et al. in contrast addresses the theoretical concerns more closely and has more complexity in the system to deal with these issues. However this results in greater difficulty in applying this to neural structures. Addressing this tension would seem a critical concern in future modeling with these kinds of techniques.

Notes

1. Reprinted from Neural Networks, 15, Kakade, S. & Dayan, P., Dopamine: generalization and bonuses, 550 (2002), with permission from Elsevier. Part of diagram originally appeared in Journal of Neurophysiology, 72, Mirenowicz, J. & Schultz, W., Importance of unpredictability for reward responses in primate dopamine neurons, 1026 (1994). This is used with permission from The American Physiological Society. Another part of diagram is originally from Science, 275, Schultz ,W., Dayan, P. & Montague, P.R., A neural substrate of prediction and reward 1594 (1997), Reprinted with permission from AAAS.

References

Berridge, K. C. & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* **28:** 309-369.

Cannon, C.M. & Palmiter, R.D. (2003). Reward without dopamine. *Journal of neuroscience* **23:**10827-10831.

Courville, A. C. & Touretzky, D. S. (2001). Modeling temporal structure in classical conditioning. In Dietterich T. G. Becker, S. & Ghahramani, Z. (eds.), *Advances in neural information processing systems, 14* Cambridge, MA: MIT Press, 3-10.

Crossman, A.R. & Neary, D. (2000). *Neuroanatomy, an illustrated colour text.* 2nd edition. Edinburgh: Churchill Livingstone, 151-160.

Daw, N. D. (2003). *Reinforcement Learning models of the Dopamine System and Their Behavioural Implications*. PhD thesis, School of Computer Science, Carnegie Mellon University.

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2003). Timing and partial observability in the dopamine system. In Becker S. Thrun, S. & Obermayer, K (eds.*), Advances in neural information processing systems 15* Cambridge, MA: MIT Press, 83-90.

Daw, N.D., Courville, A.C. & Touretzky, D.S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation* **18:**1637-1677.

Daw, N.D., Niv, Y. & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* **8:** 1704-11.

Dayan, P. & Abbott, L.F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* MIT Press 279 – 330.

Fuxe, K., Hökfelt, T., Johansson, O., Jonsson, G., Lidbrink, P. & Ljungdahl, Å. (1974). The origin of the dopamine nerve terminals in limbic and frontal cortex. Evidence for meso-corticodopamine neurons. *Brain Research* **82:** 349–355.

Haber S.N., Fudge J.L. & McFarland N.R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience* **20:**2369-82.

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

Joel, D., Niv, Y. & Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks* **15:** 535-547.

Kakade, S. & Dayan, P. (2002). Dopamine: generalization and bonuses. Neural Networks, **15:** 549-559.

Kelley, A.E., Baldo, B.A., Pratt, W.E. & Will, M.J. (2005). Corticostriatal-hypothalamic circuitry and food motivation: integration of energy, action and reward. *Physiology and Behavior* **86:** 773-795.

Mirenowicz, J. & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. Journal of Neurophysiology. 72:1024-1027.

Mitchell, T.M. (1997). Machine learning. The McGraw-Hill Companies Inc.

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

Russell, S.J. & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice-Hall International Inc.

Salamone, J.D., Correa, M., Mingote, S.M. & Weber, S.M. (2005). Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine. *Current Opinion in Pharmacology* **5:**34-41.

Samii, A., Nutt, J.G. & Ransom, B.R. (2004). Parkinson's disease. *Lancet* **363:** 1783-1793.

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

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience* **1:** 199-207.

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

Suri, R. E. (2001). Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. *Experimental Brain Research* **140:** 234-240.

Suri, R. E. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Networks* **15:** 523-533.

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

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

Sutton, R. S. & Pinette, B. (1985). The learning of world models by connectionist networks. In proceedings of the *Seventh Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum, Irvine, CA 54-64.

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

Wise, R.A. & Schwartz, H.V. (1981). Pimozide attenuates acquisition of lever-pressing for food in rats. *Pharmacology Biochemistry and Behavior* **15:** 655-656.

Wise, R.A. (2006). Role of brain dopamine in food reward and reinforcement. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **361:**1149-1158.

Wörgötter, F. & Porr, B. (2005). Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. *Neural Computation* **17:**245-319.

Young, A.M., Joseph, M.H. & Gray, J.A.(1993). Latent inhibition of conditioned dopamine release in rat nucleus accumbens. *Neuroscience* **54:**5-9.

Young, A.M., Ahier, R.G., Upton, R.L., Joseph, M.H. & Gray, J.A. (1998). Increased extracellular dopamine in the nucleus accumbens of the rat during associative learning of neutral stimuli. *Neuroscience* **83:**1175-1183.