Network: Computation in Neural Systems
March 2006; 17: 61–84

Dopamine, prediction error and associative learning: A model-based account

ANDREW SMITH[1], MING LI[2], SUE BECKER[1], & SHITIJ KAPUR[2]

[1]*Department of Psychology, Neuroscience and Behaviour, McMaster University, Hamilton, Ontario, Canada,* [2]*Centre for Addiction and Mental Health Toronto, Ontario, Canada*

(Received 7 January 2005; revised 2 September 2005; accepted 7 September 2005)

Abstract
The notion of prediction error has established itself at the heart of formal models of animal learning and current hypotheses of dopamine function. Several interpretations of prediction error have been offered, including the model-free reinforcement learning method known as temporal difference learning (TD), and the important Rescorla–Wagner (RW) learning rule. Here, we present a model-based adaptation of these ideas that provides a good account of empirical data pertaining to dopamine neuron firing patterns and associative learning paradigms such as latent inhibition, Kamin blocking and overshadowing. Our departure from model-free reinforcement learning also offers: 1) a parsimonious distinction between tonic and phasic dopamine functions; 2) a potential generalization of the role of phasic dopamine from valence-dependent "reward" processing to valence-independent "salience" processing; 3) an explanation for the selectivity of certain dopamine manipulations on motivation for distal rewards; and 4) a plausible link between formal notions of prediction error and accounts of disturbances of thought in schizophrenia (in which dopamine dysfunction is strongly implicated). The model distinguishes itself from existing accounts by offering novel predictions pertaining to the firing of dopamine neurons in various untested behavioral scenarios.

Keywords: *Dopamine, prediction error, associative learning, blocking, latent inhibition, overshadowing, schizophrenia, reinforcement learning, incentive salience, motivated behavior, temporal difference algorithm, Rescorla–Wagner learning rule, psychosis*

Introduction

Dopamine, particularly within the mesolimbic sub-system, is a neuromodulator of great interest because of its central role in reward, learning and motivation, as well as its implication in diseases such as schizophrenia. The precise role of dopamine in each of these processes is still a matter of debate, and a number of partially overlapping hypotheses exist. The primary aim of this paper is to relate two interpretations of prediction error (Temporal Difference learning and Rescorla–Wagner's learning rule) within a formal framework that can link dopamine neuron firing to learning, motivation and disturbances of thought. Our theme is to extend the notion of dopamine from a mediator of "reward prediction error" to a purveyor of "unexpected significance," and to place internal modelling at the heart of a dopamine-modulated reward system. This approach allows us to consider the implications of the model for schizophrenia, in which a dopamine disturbance is suspected.

Correspondence: S. Kapur, Centre for Addiction and Mental Health, Department of Psychiatry, University of Toronto, 33 Russell Street, Ontario, M5S 2S1. Tel: 416 979 6890. Fax: 416 260 4206. E-mail: shitij_kapur@camh.net

ISSN: 0954-898X print / ISSN 1361-6536 online
DOI: 10.1080/09548980500361624

Dopamine neuron firing and prediction error

A series of influential electrophysiological experiments by Schultz and colleagues have recorded the firing of dopamine neurons in the monkey midbrain under a variety of behavioural conditions, leading to the hypothesis that dopamine neurons fire in response to unpredicted reward. Early in training, a primary rewarding stimulus such as juice elicits a phasic response, but as the animal is repeatedly exposed to the task, this response transfers to the earliest reliable predictor of that reward (Schultz 1998). Furthermore, this fundamental finding appears to be independent of whether the conditioning procedure is instrumental (the animal must work for the reward) (Mirenowicz & Schultz 1994) or Pavlovian (the reward is not contingent on the animal's actions) (Waelti et al. 2001). It has been proposed that this phasic dopamine response drives learning by signalling a prediction error that effectively labels events as "better than expected". Important parallels between this signal and the prediction error signal of the temporal difference learning algorithm (TD) have been drawn (Barto 1995; Houk et al. 1995; Montague et al. 1996; Schultz et al. 1997), linking formal and animal reinforcement learning. Figures 1 and 2 offer a highly simplified summary of the electrophysiological recordings collected by Schultz and colleagues.

The TD hypothesis proposes that the phasic dopamine response is used to update representations of stimulus → reward, and in some cases stimulus → response, associations. TD assumes that the environment itself behaves as a Markov Decision Process (MDP). An MDP consists of a number of "states" that represent external conditioned and unconditioned stimuli. For example, one state might represent the onset of a conditioned stimulus (CS), and another state might represent the onset of an unconditioned stimulus (US). Within the MDP, each state has an intrinsic reward value associated with it. These reward values are denoted by the function, r. For example, the intrinsically rewarding properties of the US might be represented by $r(\mathrm{US}) = 1$, while the intrinsically neutral properties of the CS by $r(\mathrm{CS}) = 0$. The details of the MDP are initially hidden from TD, and the fundamental goal of the algorithm is to learn to predict the future reward associated with each state through trial and error interaction with the environment. Once these future rewards are learned, they can be used to motivate appropriate behaviour (Sutton & Barto 1998). TD represents its estimates of future reward using a Value function (capitalized to distinguish its special meaning), abbreviated to V. After learning has taken place in the current example, $V(\mathrm{CS}) = 1$, even though $r(\mathrm{CS}) = 0$. This is because the US (with $r(\mathrm{US}) = 1$) consistently follows the CS. The important point is that TD does not represent the underlying cause–effect contingencies of the MDP, but just stores the future reward or Values. Consequently, when a CS is presented, TD knows how great the future reward is likely to be, but does not know what the future states are likely to be. For example, TD cannot distinguish between different types of future US, if they had a similar reward value, r, during learning of the CS–US association. Evidence suggests that, on its own, TD is insufficient for describing many animal behaviours (Dayan 2002; Dayan & Balleine 2002). For example, animals can be motivated to respond to a CS in a highly outcome specific way (Berridge & Schulkin 1989).

TD represents only one approach to formal reinforcement learning. An important alternative involves learning a more faithful version of the MDP that includes the transitions that occur between states. Model-based approaches learn to represent a link between the state representing the CS and the state representing the US. Now, when presented with the CS, this acquired link can be used to actually invoke the internal representation of the specific US that is expected to follow. Therefore, a model-based approach, when presented with a CS, can distinguish between two different types of future US, even if they were similarly rewarding on previous occasions. TD is referred to as model-free because no explicit model of the state-to-state transitions is maintained. Model-free representations are efficient in

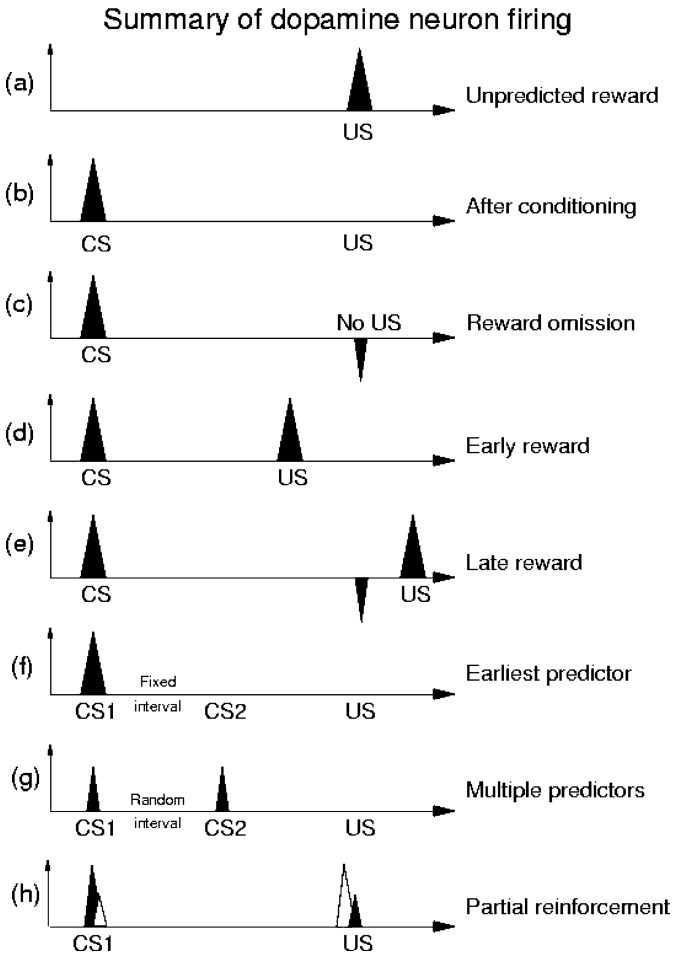

Figure 1. A simplified summary of key electrophysiological data. (a) Dopamine neurons fire in response to unpredicted reward (Romo & Schultz 1990; Schultz et al. 1992; Mirenowicz & Schultz 1994; Schultz 1997; Hollerman & Schultz 1998; Waelti et al. 2001; Fiorillo et al. 2003); (b) after training, the onset of the dopamine signal transfers to coincide with that of a predictive stimulus (sources as for (a)); (c) omission of expect reward elicits a depression below baseline firing rate (Schultz et al. 1992; Hollerman & Schultz 1998; Waelti et al. 2001); (d) if an expected reward arrives early, then a phasic response is elicited at the new time of reward, but there is no depression at the expected time of reward (Hollerman & Schultz 1998); (e) if an expected reward arrives late, then a depression is observed at the expected time of reward and a phasic response is observed at the new time of reward (Hollerman & Schultz 1998); (f) in cases where the reward is preceded by multiple (sequential) cues, the onset time of the phasic dopamine response transfers to coincide with that of the earliest predictor (Schultz et al. 1992); (g) if two sequential stimuli predict a reward, but the interval between the first and second stimulus is random, then a phasic dopamine signal persists in response to both stimuli, but not to the reward itself (Schultz et al. 1992); (h) if a conditioned stimulus predicts a reward with probability, p, then the phasic response to the CS is proportional to p, and the response to the reward itself is proportional to 1-p (Fiorillo et al. 2003). The black responses correspond to a session where $p = 0.75$, while the white responses correspond to a different session where $p = 0.25$. The sources for each figure include summary papers as well as primary experimental accounts.

environments with large numbers of states, but incomplete with respect to their ability to predict which type of US follows a CS. Figure 3 contrasts the types of internal representations learned by model-based and model-free methods.

It is a crucial and unresolved question as to whether model-free approaches such as TD are sufficient for describing the dopamine system, or whether a more faithful representation of

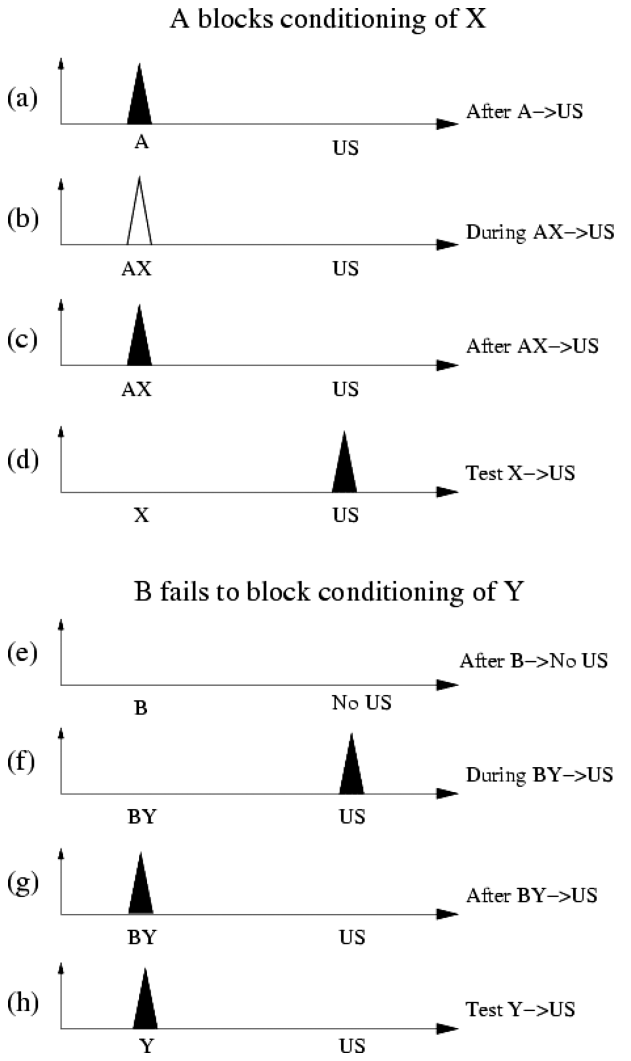

Figure 2. A simplified summary of recordings from dopamine neurons during a *Kamin blocking* task (Waelti et al. 2001). (a) A neutral stimulus, A, precedes a reward. After training, the dopamine response transfers to A, as in Figure 1b; (b) the monkey is then trained on trials in which a compound stimulus (A and X precedes the reward). There is no change in dopamine response; (c) after training on the AX → US contingency, there is still no change to the dopamine signal; (d) now, if X is tested on its own, the dopamine signal behaves as if X has not been conditioned. A lack of conditioning of X is observed in behavioural tests as well as in the dopamine neurons themselves, and this constitutes the Kamin blocking effect (Kamin 1969). (e) In a second experiment, a neutral stimulus, B, is presented but not paired with reward. No dopamine response is observed; (f) the monkey is then trained on trials in which a compound stimulus (B and Y precedes the reward). Initially the unpredicted reward elicits a dopamine response; (g) after training on the BY → US contingency, the dopamine signal transfers from the reward to the time of presentation of the compound stimulus BY; (h) now, if Y is tested on its own, the dopamine signal behaves as if Y has been successfully conditioned. In other words, stimulus B has failed to block the conditioning of stimulus Y. The unfilled triangle represents data inferred but not explicitly given.

the underlying MDP (environment) is suggested. The current paper begins by demonstrating that a model-based approach is able to account for patterns of dopamine neuron firing as easily as TD. A wider data set is then considered pertaining to traditional associative learning paradigms, motivated behaviour, tonic dopamine function, "salience"-based dopamine

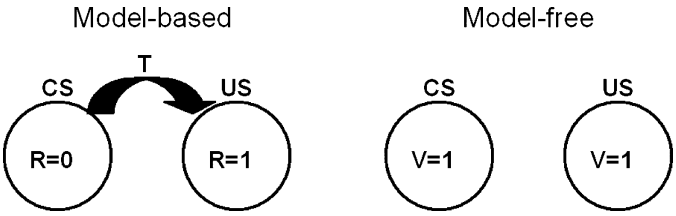

Figure 3. In model-based approaches an explicit internal model of the environment is constructed, comprising a reward function, *R*, and a transition function, *T*. In contrast, model-free techniques, such as TD, represent only a Value function, V, that estimates future reward from each state.

hypotheses, and acute schizophrenia. The advantages of using a model-based approach are explored with respect to these data.

The model

Stimulus → outcome models of the environment (also called internal models or declarative representations) have been embraced by diverse fields including artificial intelligence, formal reinforcement learning, experimental psychology and computational neuroscience (Dickinson 1980; Schmajuk et al. 2001; Sutton & Barto 1981; Sutton & Barto 1998). Our approach to modelling the data of Figures 1 and 2 is a straightforward instantiation of model-based reinforcement learning in which an estimate of the underlying MDP is explicitly represented during learning. Here, the word explicit refers to the reward function, r, and the transitions between states. The model's estimates of these components of the MDP will be denoted by R and T, respectively. No Values are learned or stored, and the cardinal feature of this model-based approach is that the all-important estimate of future reward is generated every time a CS is encountered. For example, when a CS is presented, the internal model of the environment is used to "look-ahead" to expected future outcomes (learned through experience), and the estimated rewards (R) of those future outcomes are summed to generate the estimated future reward. One advantage of generating future reward in this way is that the motivational value of the CS can be modulated based on the current motivational state of the system (i.e., hunger, salt-deprivation etc.). In contrast, TD wraps up the future reward into a pre-evaluated quantity that requires no look-ahead but that depends on the motivational state during learning. The current approach is inspired by a number of influential model-based accounts including Schmajuk (1988), Schmajuk et al. (2001), Suri (2001), Suri et al. (2001) Dayan (2002) and Dayan & Balleine (2002). For the present, we focus on the predictive relationship between CS and US and ignore action choices or the policy.

In addition to states representing the onset of external stimuli, a number of interval timing states are conventionally used, leading to the standard "tapped delay line" assumption (Montague et al. (1996) for example). This assumption has allowed formal reinforcement learning methods to simulate the interval-sensitive responses of animal behaviour and neuron firing. Continuing this convention, we assume that a learning agent has at its disposal an arbitrary number of internal states that represent every possible external stimulus at every possible time since its onset, for an arbitrary temporal resolution. One unit of time is equated with one second, for the purposes of yoking the timing to that of real experiments. We also assume that, at any time, exactly one of these internal states is appropriately activated by the environment with a value of 1, and that the active internal state will always correspond to the most recently presented external stimulus. It is convenient to define $S_{a,b}$ as the state that responds "b" seconds after the onset of stimulus "a." For example, in a standard conditioning

task in which CS and US are separated by 3s, the order of state activations might be: $S_{CS,0}$, $S_{CS,1}$, $S_{CS,2}$, $S_{US,0}$, $S_{Terminal}$, where $S_{Terminal}$ refers to a special state that signifies the end of a trial in each simulated task. After the terminal stimulus is presented, the model is dormant until the start of the next trial. These assumptions greatly simplify the discussion without, hopefully, losing any of the important details. For convenience, we will refer to a given state in one of two ways—either using the state–time index described above, or by using just a single index that refers to an abstract label. In the latter case, where stimulus and time since onset are not important, S_i simply refers to the ith state, where $i \in \{1 \ldots n\}$, with n = number of available states.

Each state, S_i, is connected to every other state, S_j, via a unique transition connection with its own weight, $0 \leq T(i, j) \leq 1$. Each transition weight, $T(i, j)$, adapts during environmental exposure to reflect the probability of internal state j following internal state i. Actual reward values are provided by the environment, and will be arbitrarily set at 1 for rewarding US, and 0 otherwise. For a given time into the current trial, t, this reward is denoted by r_t. For example, if a rewarding US happens to be present at time t, then $r_t = 1$, otherwise $r_t = 0$. Each state, S_i, maintains its own (real-valued) estimate, $R(S_i)$, of this immediate reward, which is adapted during environmental exposure. Between them, T and R model the important features of the environment, with T capturing the sequential structure of stimuli, and R capturing the rewarding impact of stimuli. R and T are the only free parameters of the model, and are initialized to 0. Figure 3 (left) shows a simple illustration.

The variable, t, will denote the time into the current trial. We define $\xi_t(S_i)$ as the real-valued activation at time t of state S_i based on the environmental input. For example if a CS is presented at time t then, $\xi_t(S_{CS,0}) = 1$, and all other internal state units have an activation of 0. If the stimulus persists then $\xi_{t+1}(S_{CS,1}) = 1$, etc. Later in the trial, a US might be presented and so $\xi_{t+5}(S_{US,0}) = 1$ for example.

At each point in each trial, a look-ahead process will be invoked to assess the expected consequences given the current state of the trial. Since the look-ahead process involves running through hypothetical outcomes inside the model, we also introduce the notion of a "look-ahead time," which will always be denoted by a new variable, v. So t defines the current point in the trial, and v denotes how many time-steps into the future (i.e., from t) the look-ahead process is currently evaluating. Therefore, the activation of hypothetical future states generated by look-ahead will be indexed by both t and v. To clarify, we define $\hat{\xi}_t^v(S_i)$ as the real-valued activation of state S_i at the vth stage of the look-ahead process, given that the look-ahead process was initiated based on the environment-driven activations of time t. For example: $\hat{\xi}_t^0(S_i) = \xi_t(S_i)$, and $\hat{\xi}_t^1(S_i)$ reflects the estimated likelihood of S_i being the next state. Similarly, $\hat{\xi}_t^2(S_i)$ reflects the estimated likelihood of S_i being the active state in two seconds time etc. The look-ahead process uses the learned transition connections (T) to simulate the next hypothetical state: $\hat{\xi}_t^1(S_i) = \sum_{j=1\ldots n} \hat{\xi}_t^0(S_j) \times T(j, i)$, and in general:

$$\hat{\xi}_t^{v+1}(S_i) = \sum_{j=1\ldots n} \hat{\xi}_t^v(S_j) \times T(j, i) \tag{1}$$

Ordinarily, only one state will be activated at a time during each step of the look-ahead process and so the calculation is trivial, simply propagating activation from one state to another as the internal model simulates expected future stimuli. As a result of the look-ahead process, $\hat{\xi}_t^v(S_i)$ provides an estimate of the probability of S_i being encountered in the actual environment v time-steps following t.

The ability to predict future reward (also called the discounted return) is central to all formal reinforcement learning techniques. When using a model-based approach, the look-ahead process can be used to generate the return, by summing all the estimated reward values

of all states encountered during look-ahead. For example, at time t, the return is estimated by:

$$\text{Return} = r_t + \sum_{v=1}^{DEPTH} \sum_{k=1\ldots n} \hat{\xi}_t^v(S_k) \times R(S_k) \tag{2}$$

where r_t is the reward elicited from the environment at time t and *DEPTH* is a finite horizon on the depth of the look-ahead process (arbitrarily fixed at 15 for all experiments reported here).

A model quantity is now introduced that will be identified with the firing of dopamine neurons. A simulated "phasic dopamine" signal is generated on the activation of a new internal state if and only if both the following criteria are satisfied:

1) **Surprise**: The currently active state was not predicted by the previously active state.
2) **Significance**: The currently active state is rewarding or predicts reward.

The simulated phasic dopamine signal will then effectively mark the current event with the tag of "unexpected Significance". A formal definition of **Significance** is given in Equation 2 and is simply the standard return—i.e., an event is significant only if it is rewarding or predicts reward. **Surprise** is simply the degree to which the current state, S_j, (at time t) is unpredicted, and can be formalised by:

$$\textbf{Surprise} = \begin{cases} \xi_t(S_j) - \hat{\xi}_{t-1}^1(S_j) & \text{If } \xi_t(S_j) > 0 \\ 0 & \textit{Otherwise} \end{cases} \tag{3}$$

Having defined **Surprise** and **Significance**, the model quantity simulating the phasic dopamine response is formalized as:

$$DA_{phasic} = Surpise \times Significance \tag{4}$$

This definition captures a very similar type of prediction error to that utilized by TD, and will be equally successful in accounting for the firing of dopamine neurons. For example, an unpredicted US will generate a response, while a fully predicted US will not. Non-rewarding events will always fail to elicit a response.

At the start of each trial, $t = 0$, and **Surprise** is automatically set to 1 since we assume that the first state of a trial can never be predicted. Every time a new state, S_j, is activated by the environment, t is incremented by 1, a reward r_t is elicited from the environment, and R is updated:

$$R(S_j) := R(S_j) + \alpha(r_t - R(S_j)) \tag{5}$$

For every new state, DA_{phasic} is also generated. In TD, DA_{phasic} is used to update the Values, but here DA_{phasic} is used to update T:

$$T(x, y) := \begin{cases} T(x, y) + \alpha \times DA_{phasic} & \text{If } \xi_{t-1}(S_x) = \xi_t(S_y) = 1 \\ T(x, y) - \alpha T(x, y) & \text{If } \xi_{t-1}(S_x) = 1 \text{ and } \xi_t(S_y) = 0 \\ T(x, y) & \textit{Otherwise} \end{cases} \tag{6}$$

for all $x \in \{1 \ldots n\}$ and $y \in \{1 \ldots n\}$, and a learning rate $\alpha < 1$ (results reported here were obtained with $\alpha = 0.2$, but other learning rates could also be used). Equation 5 simply pushes $R(S_j)$ towards the actual reward r_t. Equation 6 increases, in a Hebbian fashion, the strength of the connection between states S_x and S_y if and only if S_y follows S_x. Conversely, the connection strength is decreased if and only if S_y does not follow the previous state S_x. The

amount by which the association is strengthened is proportional to the amount of unexpected Significance in the environment (i.e., DA_{phasic}).

The equations above formally capture a simple hypothesis. First, a "dopamine" signal is generated that indicates the presence of an unexpected and significant stimulus. This signal, which for appetitive stimuli is similar to the TD prediction error signal, is then used to update the stimulus → outcome associations in an internal model.

Results

The model is now used to simulate the contingencies employed by Schultz and colleagues summarized in Figures 1 and 2. For example, Figures 1a and 1b summarize the firing of dopamine neurons in a Pavlovian experiment in which a US consistently followed a CS after a fixed interval of 4 s. The simulation is divided into 60 trials with one CS → US presentation per trial. Each trial consists of six time steps with the appropriate state being activated in each time step. The sequence of state activations is: $S_{CS,0}$, $S_{CS,1}$, $S_{CS,2}$, $S_{CS,3}$, $S_{US,0}$. The first state, $S_{CS,0}$ is always presented on the first time step of the trial, and $S_{CS,1}$ is presented on the second time step, etc. Reward is elicited only on presentation of the US, and the US is presented for just one time step. Therefore, $r_{t=5} = 1$, and $r = 0$ at all other times in a trial. The process of generating DA_{phasic}, and then updating R, and T occurs on every time step of every trial.

Figure 4a shows how the value of DA_{phasic} varies over the course of the simulation. The simulated firing of dopamine neurons within a single trial can be visualized by taking a slice perpendicular to the "Trial" axis. On trial 1, the pattern of DA_{phasic} corresponds to Figure 1a, and by trial 25, the pattern of DA_{phasic} corresponds to Figure 1b. Between trials 1 and 25, the response to the US diminishes while the response to the CS increases. A cardinal feature of this model, and also TD, is that the DA_{phasic} signal does not transfer directly from the US to the CS, but must travel back through the intervening states (i.e., $S_{CS,3}$, $S_{CS,2}$, $S_{CS,1}$). This process is manifested as a low mound that slides back from the US to the CS as the trials proceed. The height of this mound relative to the peaks at the CS and US is controlled by a number of parameters such as the number of states in the model and the learning rates. Although sustained firing of dopamine neurons has been recorded between CS and US (Fiorillo et al. 2003), this "sliding back" effect is not observed experimentally. This anomaly remains to be investigated in both model-based and model-free methods.

Figure 1d shows the experimental effect of introducing the US early after conditioning has occurred, and this effect is reproduced in the model in figure 4a at trial 30. Initially, the early response generates a signal, but the internal model is quickly adapted. Figure 5a shows the internal model constructed by trial 25. The CS is labelled "CS1," and the US is labelled "Reward." The transition connections reflect the temporal relationship between the states, and are used by the look-ahead process to estimate future reward.

Figure 4b shows the results of simulating a similar experiment, but in which the reward arrives late after a period of training. In trial 1, the pattern of DA_{phasic} corresponds to Figure 1a. By trial 30, DA_{phasic} corresponds to Figure 1b, and in trial 35 DA_{phasic} corresponds to Figure 1e. Eventually, the response to the late reward attenuates, as the internal model is adapted. The model does not account for the below baseline response (c.f. Figure 1e and see discussion).

Figure 4c shows the results of simulating a more complex experiment in which the model is first trained for around 40 trials that CS1 predicts a US. Then a second CS2 is introduced that precedes CS1 by a fixed interval, and training proceeds for another 20 trials. Finally, the interval between CS2 and CS1 is randomly varied for the last 40 trials. This reproduces the experiment performed in (Schultz et al. 1992). In both the experiment and the simulation,

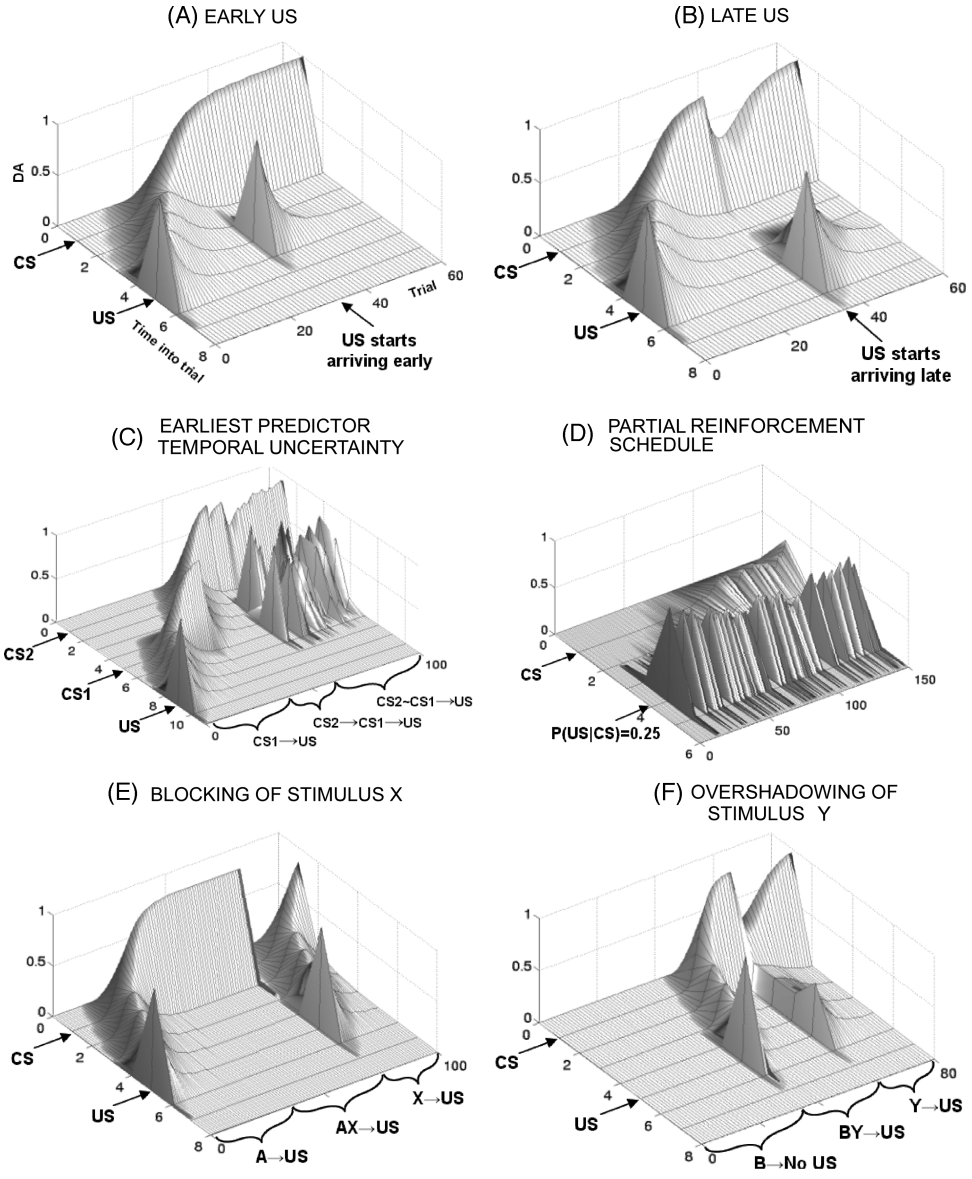

Figure 4. Each panel shows the model's performance for a subset of the data summarized in Figures 1 and 2. The simulated phasic dopamine response is plotted against trial and time into trial (see panel (a)). The labels on the trial axis show the stimuli that were presented at different stages of the session. The labels on the time-into-trial axis show the time at which those stimuli were presented within each trial. See text for panel details.

the signal of interest moves first from the US to CS1, and then from CS1 to CS2 (the earliest reliable predictor of the US). When the interval between CS2 and CS1 is varied, a signal re-appears at the time of CS1. Within the simulation, this re-appearance reflects the internal model's uncertainty regarding the time of arrival of CS1. The original experimental result is summarized in Figures 1f and 1g. Figures 5a–c show the states and the transition connections of the internal model after training on the three different contingencies of figure 4c. Figure 5d shows an example of the propagation of activation through the internal model during look-ahead following presentation of CS2 in trial 80.

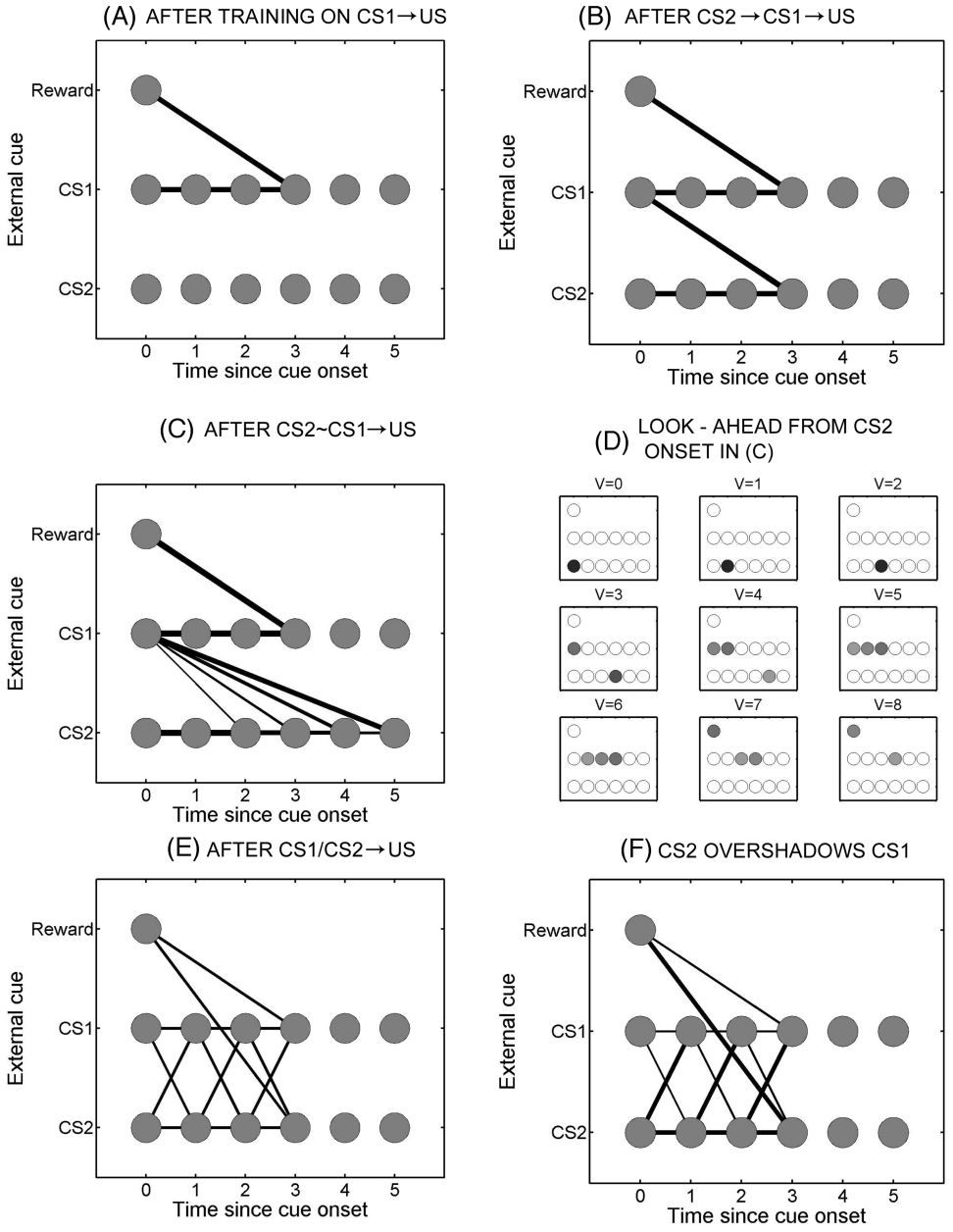

Figure 5. The internal model after training on different environmental contingencies. Each circle denotes a state of the internal model, presented according to the stimulus/time since onset represented by that state. The connecting lines show the learned strengths of the transition connections between those states. A maximum transition weight of 1 is denoted by a thick line, continuously graded in thickness down to a transition weight of 0 denoted by no line. Only the relevant internal states are shown. The learned reward values, *R*, are not shown but can be imagined inside each state. They are zero for all states except the reward state itself (i.e., the US). (a)–(c) show the internal model after the three stages of Figure 4c. (d) shows the pattern of activation on the state units of 5c during each stage of the look-ahead process following the onset of CS2. (e) and (f) show the effects of presenting a compound stimulus as CS.

Figure 4d shows the results of simulating another experiment in which the US only follows the CS on 25% of the trials. This experimental effect is summarized in Figure 1h. In both the original experiment and the model, after training, the response to the CS is approximately 0.25 and the response to the US is approximately 0.75. Although not shown, the model generalizes appropriately to all p(US|CS). For this experiment, a learning rate of $\alpha = 0.1$ was used rather than $\alpha = 0.2$, in order to allow smoother convergence of the internal model by reducing susceptibility to the random noise present in the environment.

Figure 2 summarized the effects of Kamin Blocking on the firing of dopamine neurons. Since blocking involves the simultaneous presentation of more than one stimulus, we extend the previous model definition. We now calculate DA_{phasic} by summing the existing signal for each currently active state. i.e., $DA_{phasic} = \sum Surprise(S_j) \times Significance(S_j)$ for each active state, $S_{j.}$. Here, we have indexed the two components of the phasic signal by a state. **Surprise**(S_j) is exactly as given in Equation (3), while **Significance**(S_j) is as in Equation (2) except that the look-ahead process is initiated from state S_j. The result is that the total unexpected Significance denoted by DA_{phasic} is given by the sum of the individual unexpected Significances of each simultaneously active state.

The model's simulation of the blocking of the dopamine signal is shown in Figure 4e. Initially (trials 0 to 40), the model is trained that stimulus *A* precedes the US. DA_{phasic} moves from the US to A in the usual way. Then, from trial 41–75, the model is trained that the compound stimulus, *AX*, precedes the reward. During this period DA_{phasic} responds only to the compound *AX*. From trial 75 onwards, stimulus *X* is presented on its own, followed by the US. The results show that *X* has not been conditioned during $AX \rightarrow$ US since DA_{phasic} responds to the US in trial 75, and not to *X*. The blocking of DA_{phasic} in response to *X* is consistent with the actual experimental data summarized in Figures 2a–d.

Finally, Figure 4f shows the result of simulating the experiment summarized in Figures 2e–h in which blocking is not observed. This time, initial training consists of pairing a stimulus *B* with no consequence. The second phase involves conditioning the compound BY $\rightarrow$ US. In both the experiment and the simulation the signal of interest moves from the US to the time of presentation of BY. The third and final stage involves presenting stimulus *Y* alone. In the actual experiment, *Y* is conditioned (i.e., not blocked), and a dopaminergic response occurs to *Y*. In the simulation, trial 60 shows a partial response to both *Y* and the US indicating that *Y* was partially conditioned during BY $\rightarrow$ US. This partial conditioning occurs in the model because *B* and *Y* must effectively share the prediction of the US, and represents the model's simulation of overshadowing (see next section). This overshadowing is not evident in the original electrophysiological data. Figure 5e shows the internal model that is constructed during the conditioning of the compound stimulus, BY.

Rescorla–Wagner

The previous section compares the model's performance with the firing of dopamine neurons. The current section considers the model's application to other associative learning data, focusing on behaviour. One of the most important formal models of the role of prediction error in associative learning is that of Rescorla and Wagner (1972) (RW). The model of RW, which draws on the notion of prediction error originally envisaged by Kamin (1969), provides a classic account of a suite of conditioning phenomena including Kamin Blocking (KB), Latent Inhibition (LI), and Overshadowing (OS). The relevance of KB to dopamine neuron firing is demonstrated in Figure 2. Additionally, KB and LI have been shown to be disturbed in animals and people following dopaminergic manipulations (Solomon et al.

1981; Crider et al. 1982; Moser et al. 2000), and in acute schizophrenia in which dopamine dysfunction is strongly implicated.

The RW rule states that the current change in some conditioned quantity (Q) is proportional to the associability of the CS (ϕ), the salience of the US (β), and the difference (prediction error) between the asymptotic value of the conditioned quantity (λ) and the current value of that quantity (ΣQ):

$$\Delta Q = \phi\beta\left(\lambda - \sum Q\right) \tag{7}$$

ϕ and β act as learning rates and capture observations that a more intense CS (for example), or a more rewarding US will enhance conditioning rates. Importantly, where there are multiple simultaneous CSs, all those CSs must share a fixed amount of some conditioned quantity (λ). This is achieved by summing the current value of the conditioned quantity over all CSs present (i.e., ΣQ). The effect of Equation (7) is that: 1) conditioning occurs quickly during early stages and asymptotes at λ; 2) conditioning occurs faster for intense CS and highly rewarding US; 3) all appropriate CSs effectively share the conditioned response.

Although not immediately obvious, the model-based approach discussed in the previous section is inspired by RW, where the conditioned quantity is the strength of the transition connections, T. This can be seen by substituting Equation 1 into 3, and then Equation 3 into 4 to give:

$$DA_{phasic}(i.e.\ \Delta T) = Significance \times \left(1 - \sum_{k=1\ldots n} \xi_{t-1}(S_k) \times T(k, j)\right) \tag{8}$$

where S_j is the current state, and $\xi_t(S_j)$ in Equation 3 has been replaced by 1 following our original assumption that the environment will activate the current state with a value of 1. Now, contrasting with RW, we note that the bracketed part of Equation 7 corresponds to the bracketed part of Equation 8 (i.e., **Surprise**) where λ is the asymptotic value of a transition connection (i.e., 1) and $\sum Q$ is the current degree to which the current state, S_j, was predicted by the internal model. In Equation 8, the salience of the current state (which might be a US, but need not be in the general case) is given by **Significance**. The more reward elicited by or predicted by this current state, the faster conditioning occurs. The only RW parameter not currently accounted for is the associability of the CS, ϕ. This RW-like "dopamine signal" is then used to update $T(x, y)$ in Equation 6 (top) in a manner that will help the model generalise to the range of data already captured by RW, including LI, KB and OS.

LI, KB and OS are all behavioural phenomena. A very simple additional assumption allows the model to be extended from one of neuron firing to one of behaviour. In most formal reinforcement models, estimated future reward drives machine behaviour (Sutton & Barto1998). It has already been suggested (McClure et al. 2003) that the formal quantity of estimated future reward can also be interpreted as an abstract measure of animal motivation or incentive salience following (Berridge & Robinson 1998). Although over simplifying animal behaviour, the basic principle that an animal is motivated by the future reward predicted by a stimulus is both intuitive and convenient for formal reinforcement learning models of behaviour. The approach is particularly convenient here because it allows us to interpret **Significance** as corresponding to the degree to which an animal is motivated to achieve reward, avoid punishment or suppress an ongoing response, etc., based on the expected future rewards of a CS. Under this assumption, the model is now applied to LI, KB and OS.

Latent inhibition

Latent inhibition refers to a subject's increased difficulty to form a new association between a stimulus and a reward due to prior exposure of that stimulus without consequence (Lubow & Moore 1959; Lubow 1973). The behavioural hallmark of LI is retarded conditioning of the pre-exposed stimulus, and is thought to reflect the ability of an organism to ignore irrelevant stimuli (Lubow 1989). In two pivotal studies, Solomon et al. (1981) and Weiner et al. (1981) both reported the disruption of LI (failure of pre-exposure to retard subsequent conditioning) in animals treated with the dopamine enhancing agent, amphetamine.

It is traditionally thought that non-reinforced pre-exposure of a stimulus reduces its associability with other stimuli (Rescorla & Wagner 1972; Mackintosh 1975). This can be modelled in the current context by inserting RW's associability parameter of Equation 7 at the appropriate point in the update equation for T:

$$T(x,y) := \begin{cases} T(x,y) + \alpha \times \phi(S_x) \times DA_{phasic} & \text{If } \xi_{t-1}(S_x) = \xi_t(S_y) = 1 \\ T(x,y) - \alpha T(x,y) & \text{If } \xi_{t-1}(S_x) = 1 \text{ and } \xi_t(S_y) = 0 \\ T(x,y) & Otherwise \end{cases} \quad (9)$$

where $\phi(S_i)$ is the associability parameter of state S_i (different stimuli could have different associabilities). The default associability is 1, but this value is reduced following pre-exposure, thus retarding conditioning. We have abstracted over the debated mechanisms by which the reduction in associability might actually occur.

Figure 6 (left) shows an experimental example, provided by Weiner et al. (1988), of how conditioning of the pre-exposed stimulus (labelled PE) is retarded in comparison with a non pre-exposed stimulus (labelled NPE). Figure 6 (right) demonstrates the same performance

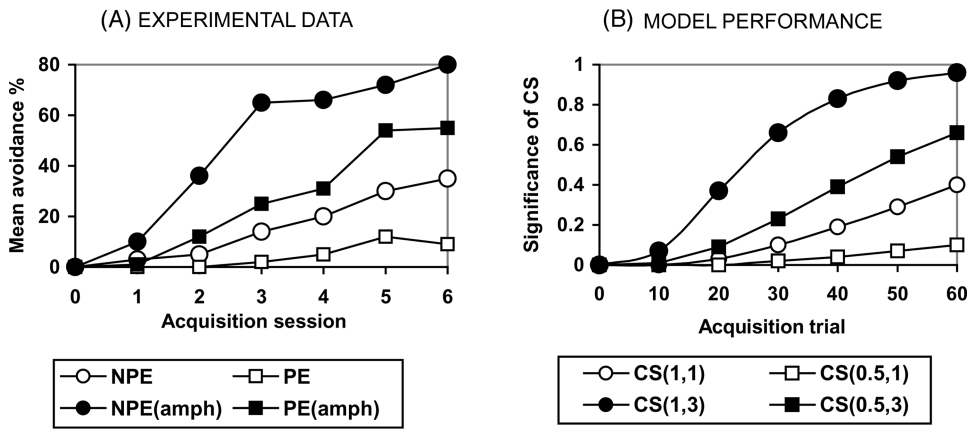

Figure 6. (left) Adapted from Weiner et al. (1988). Performance in an active conditioned avoidance task is plotted against the number of conditioning sessions (10 trials per session) in this on-baseline assessment of LI. PE = the group was pre-exposed to the CS; NPE = the group was not pre-exposed to the CS. When amphetamine was administered during conditioning then both groups condition more quickly, leading to a reversal of LI in the PE(amph) group. (right) Model simulation shows the expected future reward (**Significance** associated with the CS during conditioning). The bracketed numbers in the legend respectively indicate the values of ϕ (CS associability and π (dopamine manipulation) used for that simulation. The model posits that amphetamine ($\pi > 1$) enhances conditioning speed, pre-exposure ($\phi < 1$) retards conditioning speed, in combination these manipulations will approximately cancel, but that an asymptotic conditioned significance of 1 will always eventually be achieved given enough conditioning. A simulated CS-US interval of 2s was used, with a learning rate $\alpha = 0.025$, and R(US = 1). These parameters were hand-selected for the best quantitative match with the experimental data, but importantly the qualitative nature of the results is not dependent on these parameters.

in the model. The experimental data also show that amphetamine (an indirect dopamine agonist) enhances conditioning rate in both PE(amph) and NPE(amph) groups. In general, amphetamine disrupts LI (Weiner et al. 1988), while haloperidol (a dopamine antagonist) facilitates LI (Weiner & Feldon 1987; Weiner et al. 1987). LI disruption is indicated by amphetamine's ability to restore the conditioning rate in the PE(amph) group, to that of the un-treated NPE group. Note that acute pharmacological treatments appear to be most effective during the conditioning phase (Moser *et al.* 2000; Weiner 1990).

One obvious approach to modelling pharmacological dopamine enhancement is to multiply DA_{phasic} by a factor, $\pi > 1$. In this way, blocking the re-uptake of the neuromodulator via amphetamine is represented in the model by enhancing the impact of DA_{phasic}. Setting $\pi > 1$ (simulating dopamine enhancing treatments) will increase the conditioning rate, while $\pi < 1$ (simulating dopamine reducing treatments) will decrease the conditioning rate in both pre-exposed and non-preexposed groups (Weiner et al. 1987, 1988). Since π and ϕ are both coefficients of DA_{phasic} in the rule for updating T, simulating amphetamine in this way will act to reverse the effect of pre-exposure on conditioning speed, thus disrupting the LI effect see Figure 6 right.

Killcross et al. (1994b) demonstrated that pre-exposure and dopamine blockade mutually contribute to the expression of LI. For example, they concluded that: "DA-antagonists enhance the magnitude of an LI effect by producing a retardation in conditioning following fewer pre-exposures than are typically required." (p. 199). They also find that increasing the US intensity reverses LI, and conclude here and elsewhere (Killcross et al. 1994a) that LI results from dopamine-reinforcer interactions. If we look back at Equations 9 and 4, then we see that there are three coefficients of conditioning on presentation of the US: ϕ (CS pre-exposure), π (dopamine manipulation), and **Significance** (US magnitude, since for the US we can ignore the look-ahead process). Therefore, increasing the number of pre-exposures, blocking dopamine, and decreasing the US magnitude will all act to retard conditioning and vice-versa. Different combinations of these parameters are predicted to interact in a multiplicative manner.

Kamin blocking

KB refers to the observation that prior conditioning of a neutral stimulus (blocking CS) renders another stimulus (blocked CS) less effective in subsequent conditioning when both are presented in compound (Kamin 1968, 1969). KB has already been demonstrated within the model for DA_{phasic}. In Figure 4e, following conditioning of $A \rightarrow$US, there is no prediction error at the US during subsequent presentations of $AX \rightarrow$ US, and therefore no conditioning of the transition connection between X and the US. The **Significance** of X, if presented alone, will be 0 because all the transitions to the US come from A and no transition connections are formed from X. Under our initial assumption that **Significance** can be interpreted as any behaviour measure of conditioning, **Significance**(X) = 0 represents the classic blocking effect reported in (Kamin 1969). In Kamin's original experiment, A is a noise, X is a tone, the US is a shock, and the conditioning of X is measured by licking suppression. X fails to elicit any licking suppression.

The impact of dopamine manipulations on KB is critical for the current exposition of prediction error and the role of dopamine therein. However the literature is equivocal. For example, Crider et al. (1986) demonstrated blocking disruption in rats following chronic haloperidol treatment (leading to dopamine receptor super-sensitivity), but this treatment obscures the stage ($A \rightarrow$US or $AX \rightarrow$ US) during which disruption takes place. O'Tuathaigh et al. (2003) found KB disruption in response to acute amphetamine treatment during the compound conditioning phase, while others (Ohad et al. 1987) found that acute treatment at

either stage, but not both, disrupted KB! This is in contrast to Crider et al. (1982) who found KB disruption following chronic amphetamine treatment (i.e., in both stages). Finally, Jones et al. (1997) failed to find support for KB in people being sensitive to administration of low, acute doses of amphetamine. More detailed and consistent experimental data will determine whether dopamine manipulations are appropriately modelled as a multiplicative factor as proposed above, or whether they should be modelled as an additive factor for example.

Overshadowing

If two stimuli (CS1 and CS2) are conditioned simultaneously as a compound stimulus, then they will share their association with the US. Moreover, if one stimulus is stronger, then it will grab a greater portion of the association. This is the basic OS effect (Pavlov 1927). According to RW, the asymptotic strength of the association between CS1 and the US will be proportional to $\frac{\phi(\mathrm{CS1})}{\phi(\mathrm{CS1})+\phi(\mathrm{CS2})}$, where ϕ defines the intrinsic associability of each stimulus. Figure 5e shows the internal model after conditioning of a compound stimulus where $\phi(\mathrm{CS1}) = \phi(\mathrm{CS2})$, and Figure 5f shows the model when $\phi(\mathrm{CS2}) = 2 \times \phi(\mathrm{CS1})$. As a result, the transition connections from CS2 are twice the strength of those from CS1.

Mackintosh (1976) conditioned a compound light-noise stimulus to a shock in four groups of rats, with a different noise volume in each group. The strength of the association learned between the noise and the US, and the light and the US, was measured by the subsequent degree of licking suppression elicited by each CS presented on its own. The results are summarized in Figure 7 (left). The details of Mackintosh's experiment are easy to simulate by fixing $\phi(\mathrm{Light}) = 1$, and varying $\phi(\mathrm{Noise})$. Then, after the internal model has been constructed during conditioning, **Significance**(light) and **Significance**(noise) can be plotted against ϕ(noise) as in Figure 7 (right). The model easily simulates the experimental data. Experimentally, it is currently unclear how dopamine manipulations affect OS.

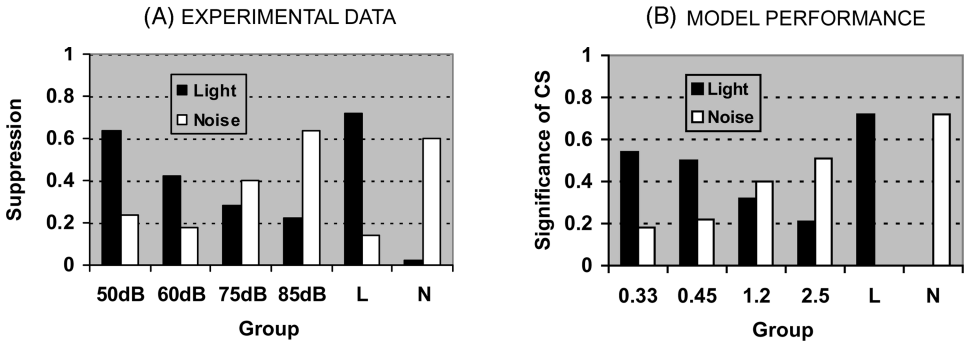

Figure 7. (left) Overshadowing data adapted from (Mackintosh 1976). Six groups of rats were used. The first four (50dB, 60dB, 75dB, 85dB) were trained that a compound stimulus (light and noise) predicted a shock. In each group, the noise was a different volume indicated by the group name. The fifth group (L) was trained that just a light predicted a shock, and the sixth group (N) was trained that just a noise predicted a shock. The graph shows the degree to which each of the two stimuli (presented alone) subsequently suppressed licking (a standard measure of the degree of learned association between CS and US). A value of 0 indicates no suppression, while a value of 1 indicates complete suppression. The basic finding is that the individual components of a compound stimulus must apparently share the degree to which they predict the US (right). Because of the RW-like approach adopted by the model, comparable performance is simulated. The ordinate shows the significance of the relevant stimulus as denoted by **Significance** (light) or **Significance** (noise). The compound groups were simulated by using R(US) = 0.72 (maximum) experimentally observed suppression, a fixed value of ϕ (light) = 1, and ϕ (noise) as indicated on the abscissa. These were selected for the best quantitative match, but the qualitative nature of the results is not dependant on these parameters. The height of each bar agrees with the prediction of RW—for example, when ϕ (noise) = 0.33, **Significance** (noise = $\frac{0.33}{1+0.33} \times R(US)$ etc.

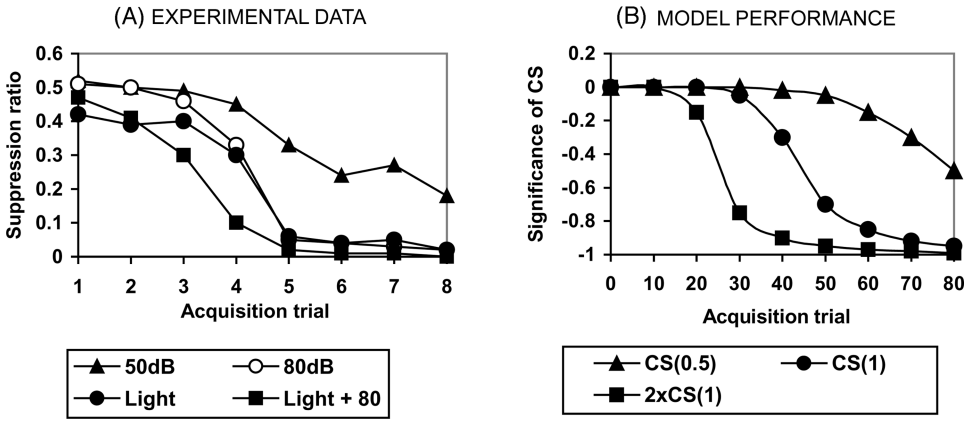

Figure 8. (Left) Adapted from (Kamin 1969). Suppression effect of four different stimuli plotted against the number of acquisition trials during conditioning to a shock. This time, a value of 0.5 indicates no suppression while a value of 0 indicates complete suppression of the ongoing behaviour (bar pressing for food). The main observations are that the less intense 50dB noise conditions more slowly than the more intense 80dB noise, the light conditions at around the same rate as the 80dB noise, and the compound stimulus conditions fastest of all. (Right) shows the model's performance under similar conditions. Again, **Significance** is plotted for the relevant CS. The circles show the rate of conditioning when a single stimulus is used with ϕ (CS) = 1. The triangles show the effect of ϕ (CS) = 0.5, simulating a less intense stimulus. The squares show the effect of using two CSs both with $\phi = 1$. Conditioning occurs more quickly because the "unexpected significance" is absorbed at twice the rate by the model. The S-shaped curve is a consequence of the time taken for the internal model to be constructed. In this simulation, R(US) = −1 so that the direction of the plots corresponds to the suppression ratio metric used by Kamin (1969).

Finally, it is generally observed that an intense stimulus will condition faster than a weaker one, and that a compound stimulus will condition more quickly than either stimulus on its own (Kamin 1969; Mackintosh 1976). Figure 8 demonstrates this both experimentally and within the model.

Incidentally, the plotted **Significance** values in Figures 6, 7 and 8 are also predictions of the DA_{phasic} response to the relevant CS. So far, the firing patterns of dopamine neurons in OS and LI paradigms have not been reported.

Discussion

Learning vs. behaviour

The model is based on the popular hypothesis that dopamine provides a prediction error signal for driving learning. However, manipulation of the mesolimbic dopamine system *after* a reward-based behaviour has been learned, also affects that behaviour. For example, Berridge and Robinson (1998) rendered rats aphagic by almost completely depleting dopamine in the nucleus accumbens and the neostriatum. These rats did not learn to become aphagic—they spontaneously stopped eating. This and other data led to rejection of the "reward learning and associative prediction" hypothesis of mesolimbic and neostriatal dopamine function. An alternative explanation places these dopamine systems at the heart of motivational or incentive salience processes (Salamone et al. 1997; Berridge & Robinson 1998; Ikemoto & Panksepp 1999). In an attempt to resolve this conflict, Parkinson et al. (2002) have suggested that the phasic dopamine signal could provide the teaching signal necessary for learning, while the background or *tonic* dopamine response (Grace 1991) could be required for the *expression* of previously acquired behaviours.

Having already addressed the role of phasic dopamine in reward-learning (in which parameters such as T and R are adapted), can the model-based approach also supply a role for tonic dopamine in motivational processes? In answering this question, we consider an interesting property of mesolimbic dopamine manipulations. Where an animal is faced with a choice between a large "distal" reward and a small "proximal" reward, ventral striatal dopamine depletion will apparently shift the preference in favour of the small "proximal" reward. For example, if a rat is trained in a T-maze that one arm yields two food pellets and the other arm yields four pellets but behind an obstruction, the animal may select the obstructed food under normal conditions. However, after dopamine is depleted in the nucleus accumbens, the rat spontaneously switches to the unobstructed arm of the maze providing less reward (Cousins et al. 1996). Other examples suggest that similar shifts towards "proximal" rewards are observed even when the motoric component of each option is equal. For example, impulsivity studies reveal that when rats are faced with the choice of two levers—one yielding a large reward after a long delay and the other yielding a small reward after a short delay—their preference can be shifted towards the immediate reward by systemic dopamine depletion (Wade et al. 2000) or by ventral striatal lesions (Cardinal et al. 2001). Conversely, their preference can be shifted towards the delayed alternative by a dopamine enhancing agent such as amphetamine (Richards et al.1999; Cardinal et al. 2000). The selectivity of dopamine manipulations on distal outcomes is also observed in conditioned avoidance paradigms (Maffii 1959).

The question now becomes: can a "non-learning" role for tonic dopamine be proposed within the model that explains not just the effect of dopamine blockade on motivation, but also the selectivity of such manipulations on distal outcomes? We have already discussed how the formal quantity of future reward can be interpreted as incentive salience. Within the model, future reward is generated by the look-ahead process, and so this is where tonic dopamine is introduced. A parsimonious solution is to introduce a new term in Equation (1) such that tonic dopamine modulates the ability of the look-ahead process to generate future reward. This is achieved by multiplying the previously acquired transition strengths by a new term, DA_{tonic}:

$$\hat{\xi}_t^{v+1}(S_i) = \sum_{j=1\ldots n} \hat{\xi}_t^{v}(S_j) \times \{T(j,i) \times DA_{tonic}\} \tag{10}$$

Under normal conditions, $DA_{tonic} = 1$ and the look-ahead process proceeds in a frictionless manner. However, under treatments reducing tonic dopamine ($DA_{tonic} < 1$) the activity of the states decays on each new cycle of the look-ahead process. Note that DA_{tonic} only modulates the efficacy of the transition connections, and does not directly change either T or R. Modification of the transitions is still mediated by DA_{phasic}. Since the generation of future reward is dependent on the activity of the states during look-ahead (Equation (2)), temporary disruption of DA_{tonic} will temporarily reduce the future reward or incentive salience of a CS. The advantage of Equation 10 is that not only will reducing DA_{tonic} reduce incentive salience, but reducing DA_{tonic} will have a particularly pronounced effect on incentive salience for distal rewards. This is because the dampening effect of DA_{tonic} accumulates over successive iterations of the look-ahead process. In this way DA_{tonic} is proposed to play the role of an online "discount factor" (see formal reinforcement learning accounts, e.g. (Sutton and Barto 1998)).

Using an internal model to account for the distance-dependent effects of manipulation of the ventral striatal D2-receptor was conceived before the current model of phasic dopamine. In an early account (Smith et al. 2005), which reviews a range of "proximal vs. distal" dopamine data, the role of phasic dopamine in reward learning and associative prediction was ignored, and it was assumed that the appropriate internal model was already constructed. In the current model, reward learning and incentive salience have been formally brought

together in a dual model of phasic and tonic dopamine function. *In summary, we propose that phasic dopamine is responsible for the formation of, and tonic dopamine for the subsequent utilization of, the transition connections of an internal model whose role is to represent salient contingencies for the purposes of motivating behaviour.*

Neural substrates: speculations

Based on experiments involving manipulation of either the ventral striatum, the D2-receptor subtype, or both, we have previously argued that the ventral striatum (and the D2 receptor) is a promising candidate for the dopaminergic modulation of the transition connections (Smith et al. 2005). A range of electrophysiological, behavioural and imaging studies implicate ventral striatal activity not just in learning about rewards (Pennartz 1996; Schultz 1998; Wilkinson et al. 1998; Wise 2004), but also in the anticipation of appetitive and aversive outcomes (Schultz et al. 2000; Jensen et al. 2003; O'Doherty et al. 2003, 2004; Seymour et al. 2004). In short, the model predicts that tonic dopamine blockade in the ventral striatum reduces motivation for future rewards, and reduces motivation for distal future rewards more than motivation for proximal future rewards.

The reward values associated with each state, $R(S_i)$, and the estimate of future reward (**Significance**) are plausibly represented in the OFC and basolateral amygdala (BLA) which have both emerged as key brain regions in associating rewards with stimuli (Rolls 2000; Tremblay & Schultz 2000). It has been suggested that the OFC in particular is crucially involved in the motivational control of goal-directed behaviour (Schultz et al. 2000), learning about rewards (Dias et al. 1996; Rolls 2000), and evaluating alternatives (Bechara et al. 1998; Schultz et al. 2000; Arana et al. 2003) via a common neural currency (Montague & Berns 2002). In summary, it is suggested that the states and associated reward values are accessed in the OFC/BLA, and the "next look-ahead state" is accessed via the transition connections in the striatum. In this way, iterations of the look-ahead process are suggested to involve the cycle of activity around the OFC → Striatal → OFC loop.

Impulsivity and "delay-discounting" studies are particularly relevant to this hypothesis. For example, physical or dopaminergic manipulation of the OFC → Striatal → OFC loop would be expected to have a greater impact on distal rewards because more cycles around the loop are required to implement the look-ahead. In a highly relevant investigation of the interaction between amygdala lesions and ventral striatal dopamine in rats, (Cador et al. 1989) conclude that the BLA plays a qualitative role in stimulus → reward associations (possibly $R(S_i)$ in our model), while ventral striatal dopamine modulates the magnitude of this role (possibly DA_{tonic} modulation of T in our model). Studies correlating dopaminergic (Richards et al. 1999; Wade et al. 2000; deWit et al. 2002), ventral striatal (Cardinal et al. 2001) and OFC (Bechara et al. 1998; Mobini et al. 2002) manipulations with changes in impulsivity and delay discounting are all consistent with the proposed neural substrate of the model. In summary, the OFC → Striatal → OFC loop, as discussed in Robbins and Rogers (2000) for example, is well organized to implement a dopamine-modulated, model-based, motivation system of the kind discussed here.

Hippocampal lesions have been shown to disrupt blocking (Solomon 1977), possibly by causing the US to remain surprising even after conditioning of the blocking stimulus (Rickert et al. 1978). In terms of the current model, this suggests that the hippocampal area could be implicated in the generation of the **Surprise** component of DA_{phasic}, and indeed Gray (1982) has proposed that the hippocampus acts as a comparator of predicted and actual events. Hippocampal lesions have been shown to disrupt mismatch detection in rats (Honey et al. 1998), and in an fMRI study, Ploghaus et al. (2000) found that hippocampal activity

correlated with the occurrence of unexpected events. However, as far as the current model is concerned, all these potential neural substrates remain highly speculative.

Predictions

At the heart of both model-free and model-based reinforcement learning techniques is the estimation of future reward and the generation of prediction error. However, one important feature of the model-based approach is that future reward is dynamically recalculated on every stimulus presentation, whereas in model-free learning (e.g., TD) this value is pre-calculated during prior exposure to that stimulus. This leads to some unique predictions from the current model-based account. First, if an animal expects a reward of type A following an appropriate CS, but instead receives a reward of type B that is approximately equal in reward-magnitude, then only the model-based account predicts a phasic response to the reward itself. Under these circumstances, the type of US is surprising in the model-based approach, whereas in the model-free approach the only value of interest is the quantity of reward, and this is not surprising. Second, according to the current model, blocking dopamine D2-receptors (possibly in the ventral striatum) is expected to attenuate dopamine neuron firing in response to a CS. This is because DA_{phasic} is dependent on the **Significance** of the CS, which is calculated by the look-ahead process, which is itself dependent on DA_{tonic}.

Although dopamine neurons appear to preferentially respond to appetitive stimuli (Mirenowicz & Schultz 1996), voltammetry and microdialysis data (see Joseph et al. (2003) for a review), have been used to argue that dopamine is released in response to a wide range of salient events (Horvitz 2000, 2002), where salience is characterized by the presence of not just appetitive, but also aversive properties. Although latest evidence supports the hypothesis that dopamine neuron firing is restricted to rewarding events (Ungless 2004), dopamine neuron firing and dopamine release may in fact be doubly dissociable (Grace 1991; Garris et al. 1999; Kilpatrick et al. 2000), suggesting additional complexities that remain poorly understood. From a behavioural point of view it is interesting that dopamine blockade leads to disruption of responses to aversive as well as appetitive CS (Courvoisier 1956)—a fact that has been used to test potential clinical efficacy of antipsychotic drugs for many years. In the current model, one potentially useful consequence of separating reward from outcome is that the definition of **Significance** is easily extended to include aversive stimuli. Following the usual convention of representing aversive events by $R(\text{US}) < 0$, Equation (2) can be changed to sum the absolute values of $R(\text{US})$ when generating **Significance**. Under these circumstances, DA_{phasic} responds to a wider range of events that include stimuli that are either immediately rewarding or punishing, or predictive of rewarding or punishing outcomes—i.e., any stimulus that is important to a useful internal model of the environment. This leads to a third prediction: if a CS precedes both an aversive and an appetitive outcome, then the model-free approach predicts a cancellation effect, while the model-based approach predicts an enhanced phasic response to that CS.

Limitations

The discussion has been boiled down to basic associative components and any discussion of how animals might select different actions (also termed the *policy* in formal reinforcement learning accounts) has been avoided. Future work is required to consider the more general case where the internal model represents action choices as well as stimuli and rewards. In this respect, model-free methods are further advanced (Montague et al. 2004). Also, the below baseline firing of dopamine neurons (Figure 1c and 1e) has not been modelled, and

further work is required to address a possible role for the sub-baseline response in extinction of the transition connections for example. Third, the current account provides, at best, only a highly abstract model of real neural processes. For example, the method must perform actual vs. predicted outcome comparisons concurrently with the look-ahead process, and therefore cannot be considered as "real-time." Fourth, the gradual "sliding back" of the prediction error from the US to the CS, as seen in Figure 4, is not observed experimentally. This failure needs to be addressed in both model-based and model-free approaches to modelling dopamine neuron firing with reinforcement learning methods. Finally, other dopamine hypotheses have been proposed that eschew the notion of prediction error altogether. For example, Redgrave et al. (1999) propose an alternative "switching" hypothesis that accounts for the short time scale response of dopamine neurons to novel stimuli—a response that apparently precedes full identification of those stimuli and their rewarding properties.

Psychosis as a failure of prediction error

Multiple lines of evidence suggest that psychosis, within the context of schizophrenia, results from a dysregulation of mesolimbic dopamine function (Weinberger 1987; Grace 1991; 2000; Moore et al. 1999). For example, repeated administration of dopamine enhancing drugs such as amphetamine can induce psychosis in otherwise healthy people (Bell 1973; Connell 1958), and the correlation between a drug's antipsychotic efficacy and the ability of that drug to block the dopamine D2-receptor is striking (Seeman & Lee 1975; Kapur & Mamo 2003). However, the links between mesolimbic dopamine dysfunction and the clinical symptoms of psychosis (such as delusions) are not well understood. In terms of associative learning, KB and LI may be disturbed in acute schizophrenia (Baruch et al. 1988; Gray et al. 1995; Jones et al. 1992; Jones et al. 1997; Moran et al. 2003; Vaitl and Lipp 1997)—although see also (Lubow et al. 1987; Swerdlow et al. 1996; Williams et al. 1998) for counter evidence—and LI has enjoyed some acceptance as a model of the processing deficits associated with the disorder. Therefore, formally linking prediction-error models of dopamine neuron firing with classical models of associative learning is seen as an important step towards a better understanding of the pharmacology and psychology of acute schizophrenia.

Maher and Ross (1984) suggested that delusions, a hallmark of thought disturbance in schizophrenia, are aberrant associations that the patient conceives in an attempt to explain unexpected observations: "... delusions represent explanations of anomalous experiences and the processes whereby the delusional belief is formed are similar in all essential respects to those that operate in the formation of normal beliefs ..." (p. 404). The model-based approach proposed here provides a formal interpretation of such a process, in which an aberrant phasic dopamine response inappropriately labels internal and external events with the tag of surprising or unexplained significance. The normal response to this dopamine signal is then invoked—that is to construct an internal model that will reduce the phasic response next time. The aberrant internal model is effectively constructed as an explanation for the mislabelled event, and forms the basis of a delusional belief. Antipsychotic drugs (all of which currently block dopamine) may act to protect against the formation of these aberrant internal associations by attenuating the impact of the phasic response, and could also dampen the motivational efficacy of existing associations via attenuation of the tonic response (as in Equation 10).

Conclusion

Here, in conjunction with Smith et al. (2005), we have attempted to integrate electrophysiological, behavioural and pharmacological data within a computational framework that is

founded on existing formal notions of prediction error. We conclude that model-based reinforcement learning has significant potential for addressing a wide range of experimental observations, and also for linking neuromodulator dysfunction with disturbances of thought in human disorders such as schizophrenia.

Acknowledgements

This work was primarily supported by an OMHF Special Initiative grant and a NET grant from the Canadian Institutes of Health Research. SK is supported by a Canada Research Chair in Schizophrenia and Therapeutic Neuroscience.

References

Arana FS, Parkinson JA, Hinton E, Holland AJ, Owen AM, Roberts AC. 2003. Dissociable contributions of the human amygdala and orbitofrontal cortex to incentive motivation and goal selection. J Neuroscience 23:9632–9638.

Barto AG. 1995. Adaptive critics and the basal ganglia. In: Beiser J, editor. Models of information processing in the basal ganglia. Cambridge, MA: MIT Press). p. 215–232.

Baruch I, Hemsley DR, Gray JA. 1988. Differential performance of acute and chronic schizophrenics in a latent inhibition task. J Nerv Ment Dis176:598–606.

Bechara A, Damasio H, Tranel D, Anderson SW. 1998. Dissociation of working memory from decision making within the human prefrontal cortex. J Neuroscience18:428–437.

Bell DS. 1973. The experimental reproduction of amphetamine psychosis. Arch General Psych 29:35–40.

Berridge KC, Robinson TE. 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res Rev 28:309–369.

Berridge KC, Schulkin J. 1989. Palatability shift of a salt-associated incentive during sodium depletion. Quarterly J Exp Psych 41B:121–138.

Cador M, Robbins TW, Everitt BJ. 1989. Involvement of the amygdala in stimulus-reward associations: Interaction with the ventral striatum. Neuroscience 30:77–86.

Cardinal RN, Pennicott DR, Sugathapala CL, Robbins TW, Everitt BJ. 2001. Impulsive choice induced in rats by lesions of the nucleus accumbens core. Science 292:2499–2501.

Cardinal RN, Robbins TW, Everitt BJ. 2000. The effects of d-amphetamine, chlordiazepoxide, alpha-flupenthixol and behavioural manipulations on choice of signalled and unsignalled delayed reinforcement in rats. Psychopharmacology 152:362–375.

Connell PH. 1958. Amphetamine psychosis. London: Chapman and Hall.

Courvoisier S. 1956. Pharmacodynamic basis for the use of chlorpromazine in psychiatry. Quarterly Rev Psychiatry Neurol 17:25–37.

Cousins MS, Atherton A, Turner L, Salamone JD. 1996. Nucleus accumbens dopamine depletions alter relative response allocation in a T-maze cost/benefit task. Behavioural Brain Res 74:189–197.

Crider A, Blockel L, Solomon PR. 1986. A selective attention deficit in the rat following induced dopamine receptor supersensitivity. Behavioural Neuroscience 100:315–319.

Crider A, Solomon PR, McMahon MA. 1982. Disruption of selective attention in the rat following chronic d-amphetamine administration: relationship to schizophrenic attention disorder. Biol Psychiatry 17:351–361.

Dayan P. 2002. Motivated reinforcement learning. In: Ghahramani T, editor. Advances in neural information processing system, Cambridge, MA: MIT Press.

Dayan P, Balleine BW. 2002. Reward, motivation and reinforcement learning. Neuron 36:285–298.

deWit H, Enggasser JL, Richards JB. 2002. Acute administration of d-amphetamine decreases impulsivity in healthy volunteers. Neuropsychopharmacology 27:813–825.

Dias R, Robbins TW, Roberts AC. 1996. Disssociation in prefrontal cortex of affective and attentional shifts. Nature 380:69–72.

Dickinson A. 1980. Contemporary animal learning theory. Cambridge, MA: Cambridge University Press.

Fiorillo CD, Tobler PN, Schultz W. 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299:1898–1902.

Garris PA, Kilpatrick M, Bunin MA, Michael D, Walker QD, Wightman RM. 1999. Dissociation of dopamine release in the nucleus accumbens from striatal self-stimulation. Nature 398:67–69.

Grace AA. 1991. Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. Neuroscience 41:1–24.

Grace AA. 2000. Gating of information flow within the limbic system and the pathophysiology of schizophrenia. Brain Res Rev 31:330–341.

Gray JA. 1982. The neuropsychology of anxiety: an enquiry into the functions of the septo-hippocampal system. Oxford: Clarendon Press.

Gray JA, Joseph MH, Hemsley DR, Young AM, Warburton EC, Boulenguez P, Grigoryan GA, Peters SL, Rawlins JN, Taib CT, et al. 1995. The role of mesolimbic dopaminergic and retrohippocampal afferents to the nucleus accumbens in latent inhibition: implications for schizophrenia. Behav Brain Res 71:19–31.

Hollerman J, Schultz W. 1998. Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neuroscience 1:304–309.

Honey RC, Watt A, Good M. 1998. Hippocampal lesions disrupt an associative mismatch process. J Neuroscience 18:2226–2230.

Horvitz JC. 2000. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience 96:651–656.

Horvitz JC. 2002. Dopamine gating of glutamatergic sensorimotor and incentive motivational input signals to the striatum. Beh Brain Res 137:65–74.

Houk JC, Adams JL, Barto AG. 1995. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Beiser J, editor. Models of information processing in the basal ganglia. Cambridge, MA: MIT Press. pp. 249–270.

Ikemoto S, Panksepp J. 1999. The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. Brain Res Rev 31:6–41.

Jensen J, McIntosh AR, Crawley AP, Mikulis DJ, Remington G, Kapur S. 2003. Direct activation of the ventral striatum in anticipation of aversive stimuli. Neuron 40:1251–1257.

Jones SH, Gray JA, Hemsley DR. 1992. Loss of the Kamin blocking effect in acute but not chronic schizophrenics. Biol Psychiatry 32:739–755.

Jones SH, Hemsley D, Ball S, Serra A. 1997. Disruption of the Kamin blocking effect in schizophrenia and in normal subjects following amphetamine. Beh Brain Res 88:103–114.

Joseph MH, Datla K, Young AMJ. 2003. The interpretation of the measurement of nucleus accumbens dopamine by vivo dialysis: the kick, the craving or the cognition. Neuroscience Biobehavioral Rev :527–541.

Kamin LJ. 1968. "Attention–like" processes in classical conditioning. In: Jones M.R, editor. Miami symposium on the prediction of behavior: aversive stimulation. University of Miami Press.

Kamin LJ. 1969. Predictability, surprise, attention and conditioning. In: Punishment and aversive behavior. New York: Appleton-Century-Crofts. pp. 279–296.

Kapur S, Mamo D. 2003. Half a century of antipsychotics and still a central role for dopamine D2 receptors. Prog Neuropsychopharmacology Biol Psychiatry 27:1081–1090.

Killcross AS, Dickinson A, Robbins TW. 1994a. Amphetamine-induced disruptions of latent inhibition are reinforcer mediated: implications for animal models of schizophrenic attentional dysfunction. Psychopharmacology 115:185–195.

Killcross AS, Dickinson A, Robbins TW. 1994b. Effects of the neuroleptic alpha-flupenthixol on latent inhibition in aversively- and appetitively-motivated paradigms: evidence for dopamine-reinforcer interactions. Psychopharmacology 115:196–205.

Kilpatrick MR, Rooney MB, Michael DJ, Wightman RM. 2000. Extracellular dopamine dynamics in rat caudate-putamen during experimenter-delivered and intracranial self-stimulation. Neuroscience 96:697–706.

Lubow RE. 1973. Latent inhibition. Psychol Bull 79:398–407.

Lubow RE. 1989. Latent inhibition and conditioned attention theory. Cambridge: Cambridge University Press.

Lubow RE, Moore AU. 1959. Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. J Comp Physiol Psychol 52:415–419.

Lubow RE, Weiner I, Schlossberg A, Baruch I. 1987. Latent Inhibition and schizophrenia. Bull Psychonomic Soc 25:464–467.

Mackintosh NJ. 1975. A theory of attention: variations in the associability of stimuli with reinforcement. Psychological Rev 82:276–298.

Mackintosh NJ. 1976. Overshadowing and stimulus intensity. Animal Learning Behav 4:186–192.

Maffii G. 1959. The secondary conditioned response of rats and effects of some psychopharmacological agents. J Pharmacy Pharmacology 11:129–139.

Maher B, Ross JS. 1984. Delusions. In: Sutker H, editor. Comprehensive handbook of psychopathology, New York: Plenum Press. pp. 383–409.

McClure SM, Daw N, Montague PR. 2003. A computational substrate for incentive salience. Trends in Neuroscience 26:423–428.

Mirenowicz J, Schultz W. 1994. Importance of unpredictability for reward responses in primate dopamine neurons. J Neurophysiology 72:1024–1027.

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

Mobini S, Body S, Ho MY, Bradshaw CM, Szabadi E, Deakin JFW, Anderson IM. 2002. Effects of lesions of the orbitofrontal cortex on sensitivity to delayed and probabilistic reinforcement. Psychopharmacology 160:290–298.

Montague PR, Berns GS. 2002. Neural economics and the biological substrates of valuation. Neuron 36:265–284.

Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neuroscience 16:1936–1947.

Montague PR, Hyman SE, Cohen JD. 2004. Computational roles for dopamine in behavioural control. Nature 431:760–767.

Moore H, West AR, Grace AA. 1999. The regulation of forebrain dopamine transmission: relevance to the pathophysiology and psychopathology of schizophrenia. Biol Psychiatry 46:40–55.

Moran PM, Al-Uzri MM, Watson J, Reveley MA. 2003. Reduced Kamin blocking in non-paranoid schizophrenia: associations with schizotypy. J Psychiatric Res 37:155–163.

Moser PC, Hitchcock JM., Lister S, Moran PM. 2000. The pharmacology of latent inhibition as an animal model of schizophrenia. Brain Res Rev 33:275–307.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. 2003. Temporal difference models and reward-related learning in the human brain. Neuron 28:329–337.

O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. 2004. Dissociable roles of the ventral striatum and dorsal striatum in instrumental conditioning. Science 304:452–454.

Ohad D, Lubow RE, Weiner I, Feldon J. 1987. The effects of amphetamine on blocking. Psychobiology 15:137–143.

O'Tuathaigh CMP, Salum C, Young AMJ, Pickering AD, Joseph MH, Moran PM. 2003. The effect of amphetamine on Kamin blocking and overshadowing. Behav Pharmacology 14:315–322.

Parkinson JA, Dalley JW, Cardinal RN, Bamford A, Fehnert B, Lachenal G, Rudarakanchana N, Halkerston KM, Robbins TW, Everitt BJ. 2002. Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. Behav Brain Res 137:149–163.

Pavlov I. 1927. Conditioned reflexes. Oxford: Oxford University Press.

Pennartz CMA. 1996. The ascending neuromodulatory systems in learning by reinforcement: comparing computational conjectures with experimental findings. Brain Res Rev 21:219–245.

Ploghaus A, Tracey I, Clare S, Gati JS, Rawlins JNP, Matthews PM. 2000. Learning about pain: the neural substrate of the prediction error for aversive events. PNAS 97:9281–9286.

Redgrave P, Prescott TJ, Gurney K. 1999. Is the short-latency dopamine response too short to signal reward error? Trends Neurosciences 22:146–151.

Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Prokasy A, ed. Classical conditioning II: Current research and theory. New York: Appleton-Century Company. pp. 64–99.

Richards JB, Sabol KE, Wit HD. 1999. Effects of methamphetamine on the adjusting amount of procedure, a model of impulsive behavior in rats. Psychopharmacology 146:432–439.

Rickert EJ, Bennett TL, Lane P, French J. 1978. Hippocampectomy and the attenuation of blocking. Behav Biol 22:147–160.

Robbins TW, Rogers RD. 2000. Functioning of frontostriatal anatomical "loops" in mechanisms of cognitive control. In: S. Driver, editor. Proceedings of Attention and Performance XVIII, pp. 475–509.

Rolls ET. 2000. The orbitofrontal cortex and reward. Cerebral Cortex10:284–294.

Romo R, Schultz W. 1990. Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. J Neurophysiology 63:592–606.

Salamone JD, Cousins MS, Snyder BJ. 1997. Behavioural functions of nucleus accumbens dopamine: empirical and conceptual problems with the anhedonia hypothesis. Neuroscience Biobehavioural Rev 21:341–359.

Schmajuk NA. 1988. The hippocampus and the classically conditioned nictitating membrane response: A real-time attentional-associative model. Psychobiology 16:20–35.

Schmajuk NA, Cox L, Gray JA. 2001. Nucleus accumbens, entorhinal cortex and latent inhibition: A neural network model. Behav Brain Res 118:123–141.

Schultz W. 1997. Dopamine neurons and their role in reward mechanisms. Curr Opin Neurobiol 7:191–197.

Schultz W. 1998. Predictive reward signal of dopamine neurons. J Neurophysiology 80:1–27.

Schultz W, Apicella P, Ljungberg T. 1992. Response of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neuroscience 13:900–913.

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

Schultz W, Tremblay L, Hollerman JR. 2000. Reward processing in primate orbitofrontal cortex and basal ganglia. Cerebral Cortex 10:272–283.

Seeman P, Lee T. 1975. Antipsychotic drugs: direct correlation bewteen clinical potency and presynaptic action on dopamine neurons. Science 188:1217–1219.

Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS. 2004. Temporal difference models describe higher order learning in humans. Nature 429:664–667.

Smith A, Becker S, Kapur S. 2005. A computational model of the selective role of the ventral striatal D2-receptor in the expression of previously acquired behaviours. Neural Computation 17:361–395.

Solomon PR. 1977. Role of the hippocampus in blocking and conditioned inhibition of the rabbit's nictitating membrane response. J Comp Physiol Psych 91:407–417.

Solomon PR, Crider A, Winkelman JW, Turi A, Kamer RM, Kaplan LJ. 1981. Disrupted latent inhibition in the rat with chronic amphetamine or haloperidol-induced supersensitivity: relationship to schizophrenic attention disorder. Biol Psychiatry 16:519–537.

Suri RE. 2001. Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. Exp Brain Res 140:234–240.

Suri RE, Bargas J, Arbib MA. 2001. Modeling functions of striatal dopamine modulation in learning and planning. Neuroscience 103:65–85.

Sutton RS, Barto AG. 1981. An adaptive network that constructs and uses an internal model of its world. Cognition Brain Theory 4:217–246.

Sutton RS, Barto AG. 1998. Reinforcement learning. Cambridge: MIT Press.

Swerdlow NR, Braff DL, Hartston H, Perry W, Geyer MA. 1996. Latent inhibition in schizophrenia. Schizophrenia Res 20:91–103.

Tremblay L, Schultz W. 2000. Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. J Neurophysiology 83:1864–1876.

Ungless MA. 2004. Dopamine: the salient issue. Trends Neuroscience 27:702–706.

Vaitl D, Lipp OV. 1997. Latent inhibition and autonomic responses: a psychophysiological approach. Behav Brain Res 88:85–93.

Wade TR, Wit HD, Richards JB. 2000. Effects of dopaminergic drugs on delayed reward as a measure of impulsive behavior in rats. Psychopharmacology 150:90–101.

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

Weinberger DR. 1987. Implications of normal brain development for the pathogenesis of schizophrenia. Arch Gen Psychiatry 44:660–669.

Weiner I. 1990. Neural substrates of latent inhibition: the switching model. Psychological Bull 108:442–461.

Weiner I, Feldon J. 1987. Facilitation of latent inhibition by haloperidol. Psychopharmacology 91:248–253.

Weiner I, Feldon J, Katz Y. 1987. Facilitation of the expression but not the acquisition of latent inhibition by haloperidol in rats. Pharmacology Biochem Behav 26:241–246.

Weiner I, Lubow RE, Feldon J. 1981. Chronic amphetamine and latent inhibition. Behav Brain Res 2:285–286.

Weiner I, Lubow RE, Feldon J. 1988. Disruption of latent inhibition by acute administration of low doses of amphetamine. Pharmacology Biochem Behav 30:871–878.

Wilkinson LS, Humby T, Killcross AS, Torres EM, Everitt BJ, Robbins TW. 1998. Dissociations in dopamine release in medial prefrontal cortex and ventral striatum during the acquisition and extinction of classical aversive conditioning in the rat. Eur J Neuroscience 10:1019–1026.

Williams JH, Wellman NA, Geaney DP, Cowen PJ, Feldon J, Rawlins JN. 1998. Reduced latent inhibition in people with schizophrenia: an effect of psychosis or of its treatment. Brit J Psychiatry 172:243–249.

Wise RA. 2004. Dopamine, learning and motivation. Nature Rev Neuroscience 5:1–12.