Psychological Review
2007, Vol. 114, No. 3, 784–805

Copyright 2007 by the American Psychological Association
0033-295X/07/$12.00 DOI: 10.1037/0033-295X.114.3.784

Reconciling Reinforcement Learning Models With Behavioral Extinction and Renewal: Implications for Addiction, Relapse, and Problem Gambling

A. David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson
University of Minnesota

Because learned associations are quickly renewed following extinction, the extinction process must include processes other than unlearning. However, reinforcement learning models, such as the temporal difference reinforcement learning (TDRL) model, treat extinction as an unlearning of associated value and are thus unable to capture renewal. TDRL models are based on the hypothesis that dopamine carries a reward prediction error signal; these models predict reward by driving that reward error to zero. The authors construct a TDRL model that can accommodate extinction and renewal through two simple processes: (a) a TDRL process that learns the value of situation–action pairs and (b) a situation recognition process that categorizes the observed cues into situations. This model has implications for dysfunctional states, including relapse after addiction and problem gambling.

Keywords: temporal difference reinforcement learning (TDRL), dopamine, reinstantiation, problem gambling

Temporal difference reinforcement learning (TDRL) algorithms have gained popularity in behavioral neuroscience to explain conditioning tasks. These models learn to select actions and to make decisions. Generally, these models do this by predicting the expected *valu*e (expected future reward) of taking an action from a given recognized situation (termed a *state* of the world). If the agent (the animal or simulation) knows the value of the consequences of its actions, it can take an action to maximize that value. Estimated value is updated through a value-prediction error term, δ, defined as the difference between expected and observed changes in value (Sutton & Barto, 1998). Positive δ indicates that the value observed is better than expected, and the estimated value of the sequence of actions and observations leading up to the event should be increased. Negative δ indicates that the value observed is worse than expected, and the estimated value of the sequence of actions and observations should be decreased (see Figure 1). The strongest support for TDRL models lies in the similarity of the dopamine signal to the value-prediction error term δ (Barto, 1995; Bayer & Glimcher, 2005; Montague, Dayan, Person, & Sejnowski, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, 2002; Waelti, Dickinson, & Schultz, 2001). TDRL models have recently found additional support from functional magnetic resonance imaging (fMRI) experiments that have examined changes in value under careful economic controls (McClure, Berns, & Montague, 2003; O'Doherty, 2004; Paulus, Feinstein, Tapert, & Liu, 2004; Tanaka et al., 2004).

Because associations are so easily reinstated after extinction (Bouton, 2004; Pavlov, 1927; Rescorla, 2004), extinction cannot entail unlearning of the original association (Bouton, 2004; Pavlov, 1927). However, standard associative models do not differentiate learning from unlearning (e.g., Rescorla & Wagner, 1972). TDRL models are generalizations of associative models (Sutton & Barto, 1981) and also do not differentiate learning from unlearning (Kakade & Dayan, 2002; Suri, 2002): A missing reward produces $\delta < 0$, which produces a decrease in value (expectation of reward), which produces a decrease in action selection. Although these models do successfully capture the slow decrease of responding that characterizes the extinction process (Kakade & Dayan, 2002; Rescorla & Wagner, 1972; Suri, 2002), they are unable to capture the quick "relearning" that is renewal.

The direct increase and decrease of estimated value as a function of the observed value-prediction error term δ will converge under appropriate learning conditions on an accurate estimate of the true value (Sutton & Barto, 1998), given the assumption of a completely described, stationary, stable world. However, that assumption clearly does not hold for real animals interacting with the real world. The real world is not completely described—which variables are available, which are important, which are unimportant, and which are hidden must be derived by the agent. In an extinction experiment, there is an explicit hidden variable unbeknownst to the animal (i.e., the experimenter has changed the reward contingency). It is our contention that the existence of hidden variables is ubiquitous in an animal's interaction with the real

A. David Redish, Department of Neuroscience, University of Minnesota; Steve Jensen, Graduate Program in Computer Science, University of Minnesota; Adam Johnson, Graduate Program in Neuroscience and Center for Cognitive Sciences, University of Minnesota; Zeb Kurth-Nelson, Graduate Program in Neuroscience, University of Minnesota.

This work was supported by a Sloan Fellowship (ADR) and a TTURC Career Development Award (NCI/NIDA P50 DA01333). We thank Erin Larson, Jonathan Gewirtz, Yael Niv, and Betsy Murray for comments on a draft of this article. We thank Jadin Jackson, Mark Thomas, Daniel Smith, Karim Nader, Josh Gordon, Greg Quirk, Martin Paulus, Stan Floresco, Geoff Schoenbaum, and Steve Grant for helpful discussions. The simulations reported here were run using Matlab R2006b and are publicly available from http://web.ahc.umn.edu/~redish/TDRLXT-Simulations.zip

Correspondence concerning this article should be addressed to A. David Redish, Department of Neuroscience, University of Minnesota, 6-145 Jackson Hall, 321 Church Street SE, Minneapolis, MN 55455. E-mail: redish@ahc.umn.edu

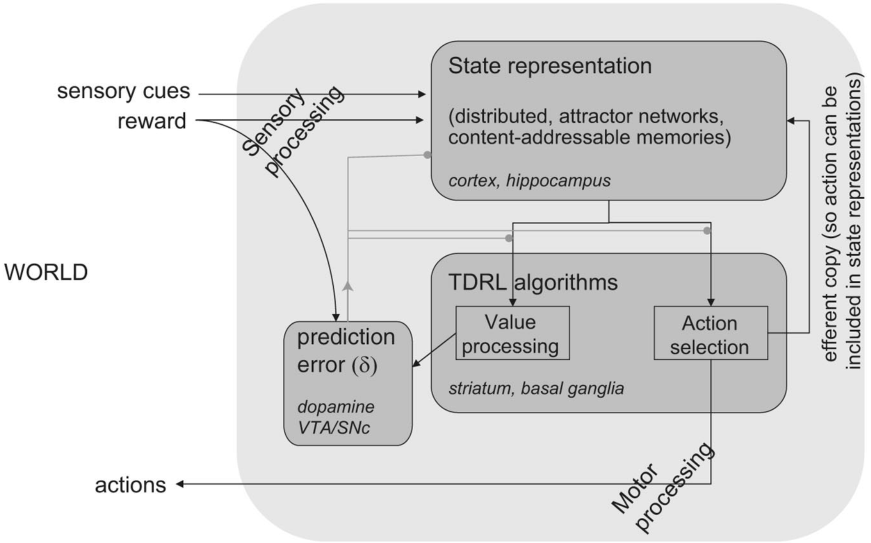

Figure 1. Reinforcement learning. In temporal difference reinforcement learning (TDRL), an agent receives sensory cues, including specialized sensory cues (identified as rewards), processes those cues, and acts upon them. From the cues, the agent must represent a (possibly distributed) state hypothesis. Value hypotheses and action selections can be made from those state hypotheses. The difference between expected and observed values is calculated as the value-prediction error δ term, which is fed back to inform the state, value, and action processing. VTA = ventral tegmental area; SNc = substantia nigra, pars compacta. A color version of this figure is available on the Web at http://dx.doi.org/10.1037/[articleDOI].supp

world and that animals have evolved mechanisms to handle this nonstationarity by categorization processes that enable the identification of (and reaction to) changes in reward contingency without unlearning.

TDRL models rely on the notion of *state* (Daw, Courville, & Touretzky, 2006; Sutton & Barto, 1998). In behavioral terms, a state is the recognition that the agent's current situation shares properties with previous (similar) situations. State is thus a categorization of the agent's current situation as a member of a class of similar situations. In practice, a state is a representative collection of salient observations that might include notable events, environmental configurations, actions, et cetera. States can include both spatial and temporal extents. Each unique state is associated with a value representing the time-discounted future reward that a behaving animal would expect when starting from that state. Any implementation of TDRL in the animal brain would have to include a representation of the state itself, a representation of the time the animal had been in the state, and an expected value of the state (Daw et al., 2006). From this information, the animal could predict the expected reward and make appropriate actions accordingly.

Theory: Implications of Two Processes

TDRL thus depends on two processes: an evaluation function that determines the value of taking an action given that the agent is in a certain situation (or state) and a situation recognition process that categorizes the observable cues into "situations" from which to reason.

A typical conditioning experiment consists of two phases: the acquisition phase and the extinction phase. The acquisition phase entails the development of an association and is learned through the increase in the value estimate associated with observation of the conditioned stimulus. We propose that the extinction phase entails the development of a new (parallel) state space that can then contain a different value estimate.

Thus, we hypothesized that tonically low δ produces a "splitting" of the representation of the state, such that new states are created that can be differentiated from the original state (*s*). Because these new states are different from *s*, actions taken can have different consequences associated with the new states compared with those of state *s*. Similarly, the same actions taken from *s* and the new states can have different estimated values.

Because the splitting of the state space is dependent on low δ, it only occurs as a consequence of a lack of expected reward. Thus, if we compare two experiments in which an animal is faced with two contexts, one in which reward is provided in both contexts and another in which reward is provided in only the first context, the model splits only the state in the situation in which reward is not provided in the second context. As shown in Figure 4, this produces context-dependent extinction and renewal in the second simulation.

Any mechanism that produces development of a new state in response to repeatedly low δ would produce the appropriate extinction with renewal. We have built a model based on increased attention to cues in response to low δ for simulation purposes, but it is important to note that any model that produces state changes in response to the undelivered, expected reward would produce similar results.

Model

All simulations were performed with the same agent model.[1] Manipulations were made to the cues provided (particularly the contextual cues) and to the probability of reward receipt. Each situation was identified by a set of cues (a context cue [A or B], binary state-identification cues [e.g., 0, 1, 0], a magnitude-of-reward-delivered cue, and a time-since-last-reward cue). The time-since-last-reward cue was reset to zero on entry into a new environment or condition. All cues were treated identically by the agent. A small amount of noise (1%) was added to each cue on each time step.

The agent itself consisted of two components: a state-classification component (implemented as a radial-basis competitive learning model with expansion; Bishop, 1995; Duda et al., 2001; Grossberg, 1976; Hertz, Krogh, & Palmer, 1991) and a temporal difference learning component (implemented as a *Q*-learning model by using *Q*[*s, a*]; Sutton & Barto, 1998; Szepesvári & Littman, 1999).

The state-classification model used for the agent was based on standard competitive learning algorithms (Grossberg, 1976; Hertz et al., 1991) that were based on radial-basis functions (Bishop, 1995; Hertz et al., 1991) with classifier expansion (Duda et al., 2001; Hertz et al., 1991). At each time step, the actual situation in the world provided the agent with a multidimensional cue,

$$c(t) = (\text{context}, s_{\mathrm{W}}, \mathrm{R}, \text{time since last reward}) + \nu, \qquad (1)$$

where s_{W} was the binary state-identification cue (e.g., light on, sound off, house light on), R was the magnitude of reward delivered in that time step, and ν was the zero-mean Gaussian noise with standard deviation, σ_{CN}. Also, ν was a vector of dimension $n_c = \text{length}\ (s_{\mathrm{W}}) + 3$. For numerical stability, context and state identifying binary cues were either 0 or 100; time since last reward was an integer identifying the number of time steps since the agent last received reward. The activation $h(s_{\mathrm{A}})$ of each potential agent state s_{A} was calculated through the following three steps:

$$Zc(t)) = (w_A(c(t) - \mu_i)) \qquad (2)$$

$$D^2(c(t)) = Z(c(t))' \textstyle\sum_i^{-1} Z(c(t)) \qquad (3)$$

$$P(c(t)|s_i) \propto \frac{1}{\sqrt{2\pi^{n_c}|\Sigma_i|}} \exp\left(-\frac{1}{2}D^2(c(t))\right). \qquad (4)$$

$Z(c(t))$ measures the difference between the current set of cues $c(t)$ and the mean for the prototype for state i, μ_i. This distance was weighted by the cue weight w_{A}. (In this formulation, $c(t)$, μ_i, and w_{A} are all tuples of size n_c; see Equation 1.) $D^2(c(t))$ transformed this by the covariance matrix for state i, Σ_i. This measure is the Mahalanobis distance, after the dimensions have been stretched or compressed through the weighting function w_k. $P(c(t)|s_i)$ transforms $D^2(c(t))$ into a Gaussian distribution; $P(c(t)|s_i)$ is normalized so that under normal conditions (when $\forall_k w_k = 1$). $P(c(t)|s_i)$ measures the probability that the current set of cues $c(t)$ could have been drawn from a multivariate normal of n_c dimensions centered at μ_i with covariance matrix Σ_i (Duda et al., 2001).

If any state had a stronger activation than threshold (ϑ_s), then the state with maximal activation was identified as the current agent state. If no state had a stronger activation than threshold (ϑ_s), a new state was created with center $\mu = c(t)$ and spherical covariance matrix with variance $\sigma_0 = 25$.

Once more than 100 observations were classified as part of a state, the parameters for each state were updated at each time step. For each cue, μ_i was updated to the mean of all observations classified as state s_i, and Σ_i was updated to the covariance matrix of all observations classified as state s_i. Thus a state with highly consistent observations would tighten its variance to match those observations, whereas a state with very variable observations would expand its variance to cover a large range of cues. Once more than 100 observations had been observed in the world, the attention parameter w_k (cue weight for cue k) was also updated on each time step on the basis of the information that cue provided to the state space:

$$w_k = 0.5 + 0.5\ \tanh(I_M(C_k,S) - 0.5)/\xi_{cw}) \qquad (5)$$

$$I_M(C_k,S) = H(C_k) - \sum_i H(C_k^i) \qquad (6)$$

$$H(C_k) = \sum_t p(c_k)\log_2 p(c_k(t)) \qquad (7)$$

$$H(C_k^i) = \sum_{t \in s_i} p(c_k)\log_2 p(c_k(t)). \qquad (8)$$

$I_M(C_k, S)$ was the mutual information between the cues and the states, defined as the increase in entropy in the observed cue distribution. $H(C_k)$ was the total entropy of all observations in the cue space, and $H(C_k^i)$ was the entropy over those observations that were categorized as being in state s_i.

Thus cues that provided no information to the state space were ignored. ξ_{cw} was a parameter that controlled the slope of the sigmoid. Cue weight w_k is a form of attention and was modified by tonic levels of δ. (See Equation 13.)

The TDRL model used for the agent was based on standard *Q*-learning value-prediction methods (Sutton & Barto, 1998). For each state,[2] an expected value[3] of taking action *a* was stored, *V*(*s, a*). Newly created states had all expected values set to zero. Expected value was updated using standard one-step temporal difference algorithms (Daw et al., 2006; Dayan & Abbott, 2001; Sutton & Barto, 1998).

On each time step, an action was selected on the basis of a standard softmax selection process (Sutton & Barto, 1998; Dayan & Abbott, 2001),

$$P(\text{select action } a|s(t)) = \exp(\beta V(s(t),a))/\sum_a \exp(\beta V(s(t),a)), \qquad (9)$$

[1] The simulations reported here were performed using Matlab R2006b and are publicly available from http://web.ahc.umn.edu/~redish/TDRLXT-Simulations.zip

[2] Remember, states are internal to the agent—they are categorizations of cues provided by the world. They reflect regularities in the world, but they do not necessarily have direct correspondence to experimenter-introduced "world states."

[3] Typical formulations of value as a function of state–action pairs are termed *Q-learning* to differentiate them from values based only on states. We prefer to use the term *V*(*s, a*) to reflect that these are "values."

where $s(t)$ was the agent's current state; and β was a parameter that balanced exploration and exploitation (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Doya, 2002). High β drove the agent to be more likely to select the high-value action, whereas low β made the agent more likely to select actions more randomly.

On every time step, δ was calculated as the difference between the expected value of the previous state–action pair and the observed value,

$$\delta(t) = \gamma(\max_{a}(V(s(t),a)) + R(t)) - V(s(t-1),a), \quad (10)$$

where $V(s, a)$ was the stored value of taking action a from state s, and γ was a discounting parameter set to 0.25.[4] In the TDRL component, δ adjusted the value estimate of the agent's hypothesized state via the standard temporal difference reinforcement learning rule,

$$V(s(t-1),a) \leftarrow V(s(t-1),a) + \eta\delta(t), \quad (11)$$

where η was a learning rate parameter, set to 0.05.

$\bar{\delta}$ was defined as an exponentially decaying running average of recent δ signals that were < 0:

$$\bar{\delta}(t) = \xi_0\bar{\delta}(t-1) + \xi_1\lceil\delta\rceil, \quad (12)$$

where $\lceil\delta\rceil$ indicates rectification at 0, such that $\lceil 1\rceil = 0$ but $\lceil -1\rceil = -1$. ξ_0 and ξ_1 were parameters controlling the speed at which $\bar{\delta}$ could change.

We explored allowing $\bar{\delta}$ to control key parameters in the equations above, including the distribution of attention to cues (Equation 5), the acceptable width of the radial-basis function ϑ_i (Equation 4), and the exploration/exploitation parameter β (Equation 9). However, we found that $\bar{\delta}$ controlling the attention to cues (Equation 5) was the most stable and was sufficient to produce all the necessary results. This is therefore the version we report here:

$$\text{effective } w_k = (-\tanh(\bar{\delta}/\xi_{DB})) + (1 + \tanh(\bar{\delta}/\xi_{DB}))w_k. \quad (13)$$

Theoretically, $\bar{\delta}$ could range from 0 to $-\infty$, approaching $-\infty$ as the agent started missing rewards. As $\bar{\delta} \rightarrow -\infty$, $\tanh(\bar{\delta}/\xi_{DB}) \rightarrow -1$, the effective $w_k \rightarrow 1$. As $\bar{\delta} \rightarrow 0$, $\tanh(\bar{\delta}/\xi_{DB}) \rightarrow 0$, the effective $w_k \rightarrow w_k$, which was normally <1. Thus, when the agent missed expected rewards, it began to pay closer attention to the cues. This made the agent more likely to create a new state hypothesis as described above. ξ_{DB} was a standard squashing parameter, controlling the slope at which changes in $\bar{\delta}$ affected w_k. The parameters used in the simulations are summarized in Table 1.

Simulation 1: Acquisition of a Response

The first simulation (acquisition) simply tested the simulation's ability to acquire a response. It simulates a simple FR1 experiment. The world consisted of two situations and 10 actions (see Figure 2). If the agent took Action 1 when the world was in Situation 0 (S0), it received reward (R) with probability (P), and the world transitioned to Situation 1 (S1). Under any other condition, the world returned to Situation S0. Actions 2–10 simulate other things the agent can do in the environment (sleep, run around, groom, etc.). Simulated agents ($N = 50$) were run under conditions of ($R = 1$, $P(R) = 1$), indicating a reward of 1 with probability of delivery of the reward of 1, and ($R = 0$, $P(R) = 0$) for 250 time steps. We measured the number of reward attempts by measuring the proportion of times in which the agent selected Action 1 from S0.

Table 1
Parameters Used in Simulations

Parameter	Variable	Value
	Learning	
η	Learning rate	0.05
γ	Discount factor	0.25/ts
	States	
σ_0	Initial covariance of RBF	25
ϑ_p	Threshold for new state	10^{-8}
σ_{CN}	Cue noise	1
ξ_{CW}	Cue weight sigmoid factor	3
	Value	
ξ_0	$\bar{\delta}$ history	0.9999
ξ_1	$\bar{\delta}$ scale factor	1.50
ξ	$\bar{\delta}$ sigmoid factor	1
	Actions	
β	Action sigmoid factor	5

Note. RBF = radical basis function.

Simulation 2: Extinction With Renewal

This simulation tested the basic extinction with renewal result. In the same world from Simulation 1, each agent experienced 250 time steps in an acquisition condition in Context A (Action 1 in state S0 led to reward $R = 1$, $P(R) = 1$). Then, agents were potentially moved to an extinction condition. Some agents were moved to a new context (Context B), and some agents remained within the same context (Context A). The two contexts were identical except for the context cue. Some agents received reward ($R = 1$, $P(R) = 1$) in the potential extinction condition, and other agents did not ($P(R) = 0$). Agents remained in the potential extinction condition for 250 time steps. Finally, all agents were returned to Context A for 250 time steps, in which they all had the opportunity to respond for reward again (renewal: $R = 1$,$P(R) = 1$). Agents ($N = 50$) were run under each condition (reward in Context A during potential extinction condition, no reward in Context A during potential extinction condition, reward in Context B, no reward in Context B).

Simulation 3: Nonrewarded Cued Renewal

This simulation tested the effect of returning the agent to a familiar environment without reward. In the same world from Simulation 1, each agent experienced 250 time steps in an acquisition condition in Context A (Action 1 in state S0 led to reward $R = 1$, $P(R) = 1$). The association was then extinguished for 250 time steps in Context B, just as in Simulation 2. The agents were

[4] We used a discount factor of 0.25 because the world model did not include a long intertrial interval (ITI) separating each trial (see Figure 2). Adding such an ITI state does not change the results, but it increases the complexity of the simulations unnecessarily. Discount factors as high as 0.90 provide qualitatively similar results to those presented here.

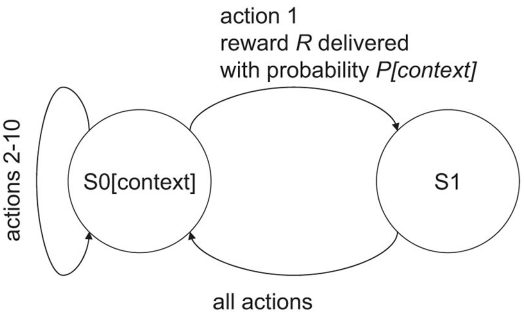

Figure 2. The world used for the simulations.

then returned to Context A, but no reward was provided. Agents ($N = 50$) were run, and the proportion of times in which the agent selected Action 1 in S0 was measured.

Simulation 4: The Partial Reinforcement Extinction Effect (PREE)

This simulation tested the effect of partial reinforcement on extinction. In the same world as in Simulation 1, agents experienced 750 time steps in an acquisition paradigm, followed by 250 time steps in an extinction paradigm. Four partial reinforcement paradigms were tested ($P(R) \in (1.0, 0.75, 0.50, 0.25)$). The reward received on each successful attempt (taking Action 1 in S0) was titrated so that the total average reward expected from full responding was held constant. Thus, the agents experiencing $P(R) = 1.00$ received one unit of reward each time the world delivered reward, whereas agents experiencing $P(R) = 0.25$ received four units of reward each time the world delivered reward. Extinction occurred in the same context with the same cues as acquisition. Agents ($N = 50$) were run under each condition. Proportion of extinction was measured as the mean number of attempts per unit of time in the first 100 steps of the extinction condition divided by the mean number of attempts per unit of time in the last 50 steps of the training condition.

Simulation 5: PREE With Intervening Continuous Reinforcement

This simulation replicated the results of Jenkins (1962) and Theios (1962), in which a continuous reinforcement condition was interposed between the partial reinforcement condition and the extinction condition. Agents experienced 750 time steps in a partially reinforced acquisition paradigm (identical to that in Simulation 4), followed by 150 time steps in a full reinforcement paradigm, $R = 1$, $P(R) = 1$, followed by 250 time steps in an extinction paradigm (as in Simulation 4). Extinction was measured (as in Simulation 4) as the ratio of attempts after extinction to the number of attempts at the end of the partial training condition (i.e., the continuous reinforcement condition was not included in the measurement).

Simulation 6: Problem Gambling

In order to examine the impact of state categorization on problem gambling, agents were allowed to experience the basic single-choice world (Figure 2). Agents were first allowed 250 steps in the world with $R = 0.6$, $P(R) = 0.1$. In order to simulate the cost of playing, taking Action 1 under any condition entailed a cost, assessed as a negative reward (cost $= -0.50$). This meant that if the agent regularly took Action 1 in S0, it would lose an average of 0.44 per try. Agents were then put into a world with a $R =$ payout, $P(R) = 0.50$ for 100 time steps, where payout was 5, 10, 25, or 100, depending on the trial (different payouts were tried for different agents, but each agent experienced only a single payout size. This was done to simplify simulations and does not change the results). Afterward, agents were then allowed 250 steps under the original conditions (i.e., $R = 0.6$, $P(R) = 0.10$). Because the number of wins changed randomly between agents, we used regression statistics to examine the effects of number of wins, the total payout, and the variance of winning on the agents' likelihood to continue playing. Agents ($N = 50$) were run under each condition.

Results

Acquisition

When faced with no reward, agents selected all actions randomly. Individual differences between agent responses arose from noise in the cues provided and in the action-selection process (see the *Method* section). Because there were 10 actions available, random chance produced a 10% chance of selecting Action 1 in S0. In contrast, when receiving reward on each attempt, the agents learned to increase the probability of attempting reward (see Figure 3). Because the β parameter continued to drive some exploration as well as some exploitation, maximal responding approached but did not reach 100%. The slope of the increase depended directly on the learning rate η and less directly on other parameters. The maximum response depended primarily on the exploration–exploitation parameter β. Agents tended to categorize the two situations (S0, S1; see Figure 2) into two internal states. However, some agents used three or four states to represent the environment. In these cases, agents split the categorization of either S0 or S1 into multiple states.

Extinction With Renewal

Once the agent has learned to select the appropriate action in the appropriate situation, it does not need to relearn this association if

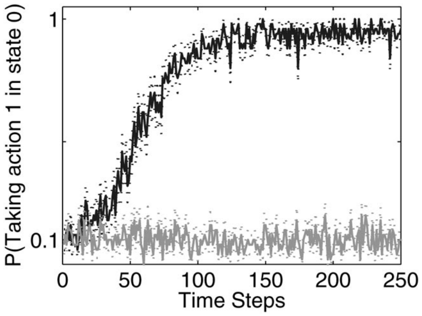

Figure 3. The agent learns to make responses that lead to rewards. Black line: response leads to reward; gray line: response does not lead to reward. A color version of this figure is available on the Web at http://dx.doi.org/10.1037/[articleDOI].supp

it is placed in a new context. Figure 4 shows the responses made by agents when faced with either reward or no reward in the new context. Agents that continued to get reward in the new context continued to respond at high rates, whereas agents that did not receive reward in the new context dropped their response rates back to random (10%). However, both groups responded at high rates when returned to the original context. Note the very fast reacquisition of the response in the renewal condition for the extinction groups.

This fast reacquisition occurred because the extinction group did not forget the association learned in the acquisition condition. Instead, they created new internal representations of the situation (i.e., they "split" their states). This allowed the extinction group to revert back to the original state representation when returned to the original context (renewal). This state-splitting process can be seen in Figure 4 (right panels): Agents that continued to receive reward used fewer states to represent the two situations (S0, S1) than did agents that received extinction training.

The slope of the rate of extinction depended on the extent of the difference between the two acquisition and extinction contexts. For example, in the top panels of Figure 4, the context was not changed between conditions. However, in the lower panels, the two contexts were quite different, and extinction happened quickly. In either case, as long as the situation identifying cues did not change between the two contexts, there was no acquisition slope for agents that continued to receive reward. And in all cases, renewal of responding happened very quickly (compare renewal conditions with acquisition conditions in Figure 4).

Cued Renewal

When an agent is trained in one context, extinguished in another, and then returned to the first context, spontaneous renewal of responding is often seen (Bouton, 2002, 2004; Bouton, Westbrook, Corcoran, & Maren, 2006). This can also be seen in cued associations; animals are likely to "spontaneously recover" responding to a cue they have not been exposed to for a long time (Pavlov, 1927; Rescorla, 2004; Robbins, 1990). This cued renewal of responding is thought to underlie cued relapse to addiction (Childress, Ehrman, Rohsenow, Robbins, & O' Brien, 1992; Childress et al., 1993; Childress, McLellan, Ehrman, & O'Brien, 1988; O'Brien, Childress, McLellan, & Ehrman, 1992). In our model, extinction primarily proceeds by re-representation of a situation as a new state. Thus, extinction does not remove the association previously stored. Presenting a cue that suggests to the agent that it is in the original situation (rather than in the new situation) will renew responding, even if no reward is presented. In Figure 5, the agents renewed responding when returned to Context A. Because

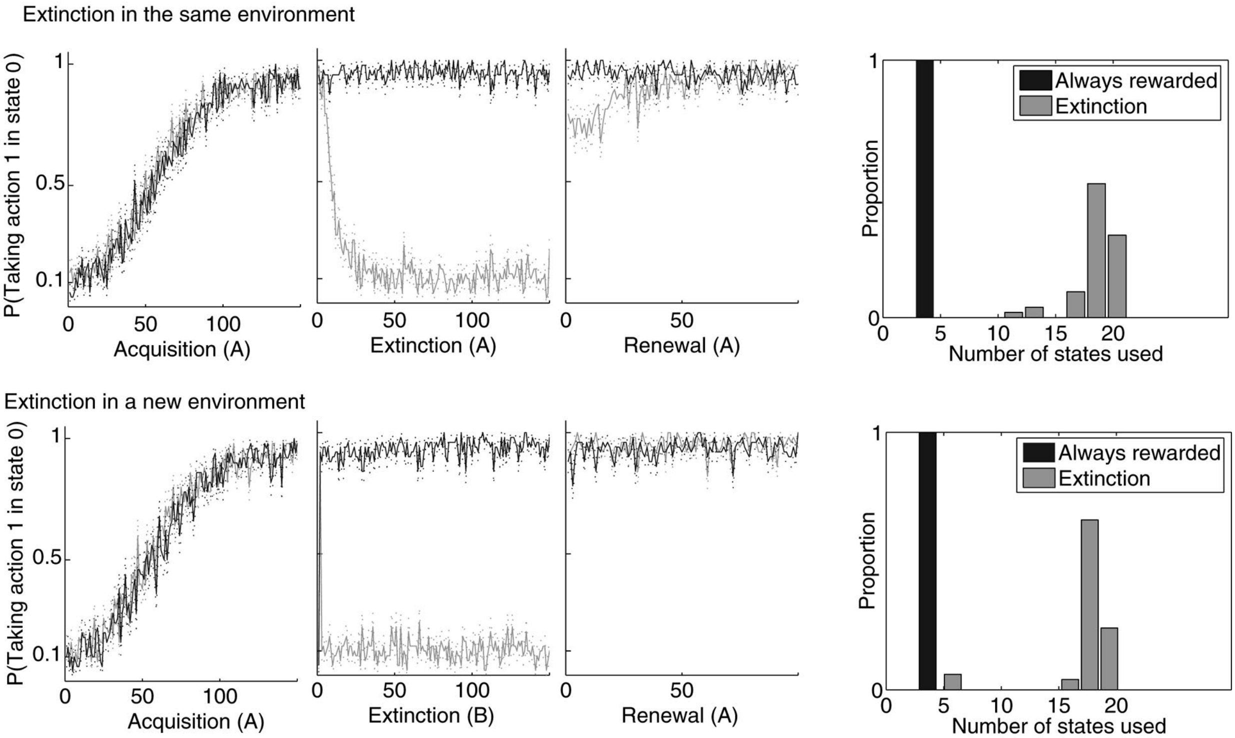

Figure 4. Extinction with renewal. (Left panels) Acquisition: The agent learns to make responses that lead to rewards (i.e., taking Action 1 in Situation 0 [S0]). Extinction: Some agents no longer receive reward for responding (gray line), whereas other agents continue to receive reward for responding (black line). Renewal: Agents were all then returned to getting reward for taking Action 1 in S0. (Right panels): Number of states used. In the top panels, the agent remained in the same context (A) through all three conditions. In the lower panels, the agent was moved from Context A to Context B between conditions. A color version of this figure is available on the Web at http://dx.doi.org/10.1037/[articleDOI].supp

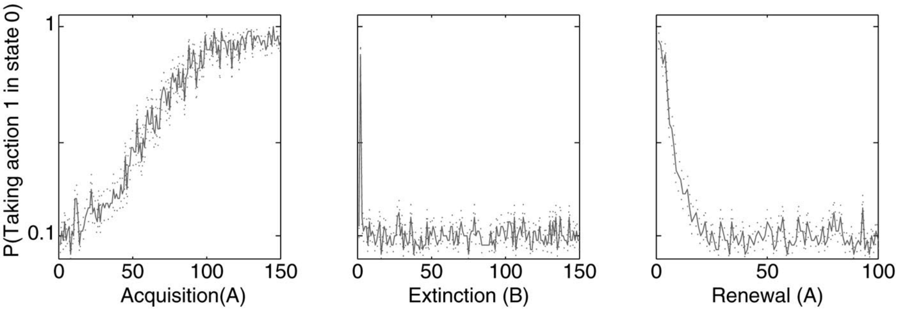

Figure 5. Cued renewal of responding. Acquisition: The agent learns to make responses that lead to rewards (i.e., taking Action 1 in Situation 0[S0]). Extinction: Agents are placed in a new context and no longer receive reward for responding (gray line). Renewal: Agents were all then returned to the original context (A). Because in the renewal condition agents were not rewarded for actions taken in Context A, they reextinguished their responding. Because the renewal condition occurred in the same context under conditions more similar to the acquisition context than to the extinction context, extinction was slower during renewal than during the original extinction. A color version of this figure is available on the Web at http://dx.doi.org/10.1037/[articleDOI].supp

no reward was given in the renewal condition, extinction then proceeded normally in Context A.

PREE

When an agent is trained with partial reinforcement, the agent is slower to extinguish its response (Capaldi, 1957; Domjan, 1998). Our simulations also show this effect (see Figure 6). In our model, slowed extinction after partial reinforcement arises because the agent has learned that nonrewarded conditions can occur during partial reinforcement. Thus, when the agent does not receive reward during the extinction condition, it takes longer to determine that the extinction condition is different and will require a different situation representation (a “splitting” of the state). This is very compatible with Capaldi and colleagues’ suggestion that extinction arises from a discrimination between acquisition and extinction conditions (Capaldi, 1957, 1958; Capaldi & Birmingham, 1998; Capaldi & Lynch, 1968).

The Role of Discrimination in PREE

Early theories explained the PREE as a consequence of the ability of the animal to discriminate between the acquisition and extinction conditions (Domjan, 1998). These theories were tested by interposing a continuous reinforcement condition (in which the animal was always rewarded for responding) in between the partially reinforced acquisition condition and the extinction condition. Under the simplest discrimination hypothesis, this interposed continuous reinforcement should always enable the discrimination of the extinction condition and should remove the PREE (Domjan, 1998). However, interposing fully reinforced conditions did not remove the PREE (Jenkins, 1962; Theios, 1962).

In our model, the PREE does arise from a stronger ability to discriminate between fully reinforced and extinction conditions than between partially reinforced and extinction conditions. In our simulations, interposing an intervening fully reinforced condition does not remove the PREE. (See Figure 7.) As above, this is because the partially reinforcing condition includes unrewarded responses. The interposing fully reinforcing condition does not remove the memory of the partially reinforced condition, which allows the animal to discriminate between the current observations (in the extinction condition) from previous experience (i.e., the acquisition condition). Thus the agent continues to show a PREE.

Discussion

The theory put forward in this article explains acquisition and extinction as two interacting learning processes: a storage of new associations, driven by positive δ signals (signaling reward, positive value larger than expected, leading to acquisition) and a splitting of the state space, driven by low tonic $\bar{\delta}$ signals (signaling disappointment, a lack of delivered expected positive value, leading to extinction). This theory suggests that the decision as to whether to act after extinction is not a decision-process question—*Should I act or not?*—but rather a cognitive question—*Which situation am I in?*

This new theory captures the time course of acquisition via standard TDRL associative learning processes. The time course of extinction, however, arises from probabilities of recall. Whether the representation is always either in a specific state and the slow decay time course of extinction occurs via averaging over trials (as suggested, e.g., by Gallistel, Fairhurst, & Balsam, 2004) or whether the believed state of the world is somehow mixed proportionally between states producing probabilistic action selection (as suggested, e.g., by Daw et al., 2006) is still an open question. In any case, in our model, renewal occurs via a sudden return of the representation of the world to the original state (s) from which the agent expected to receive reward. The time course of renewal would then depend on the probability of entering the state in which the agent expects reward over other potential states.

Relation to Other Models of Extinction and Renewal

Current theories of extinction in both psychology and neuroscience (Bouton, 2002, 2004; Bouton et al., 2006; Delameter, 2004; Milad & Quirk, 2002; Myers & Davis, 2002; Quirk et al., 2006; Rescorla, 2004) are based on the addition of new variables, often identified as contextual (Bouton, 2002, 2004; Bouton et al., 2006) or inhibitory (Delameter, 2004; Pavlov, 1927). In any content-addressable memory, the cues provide inputs from which

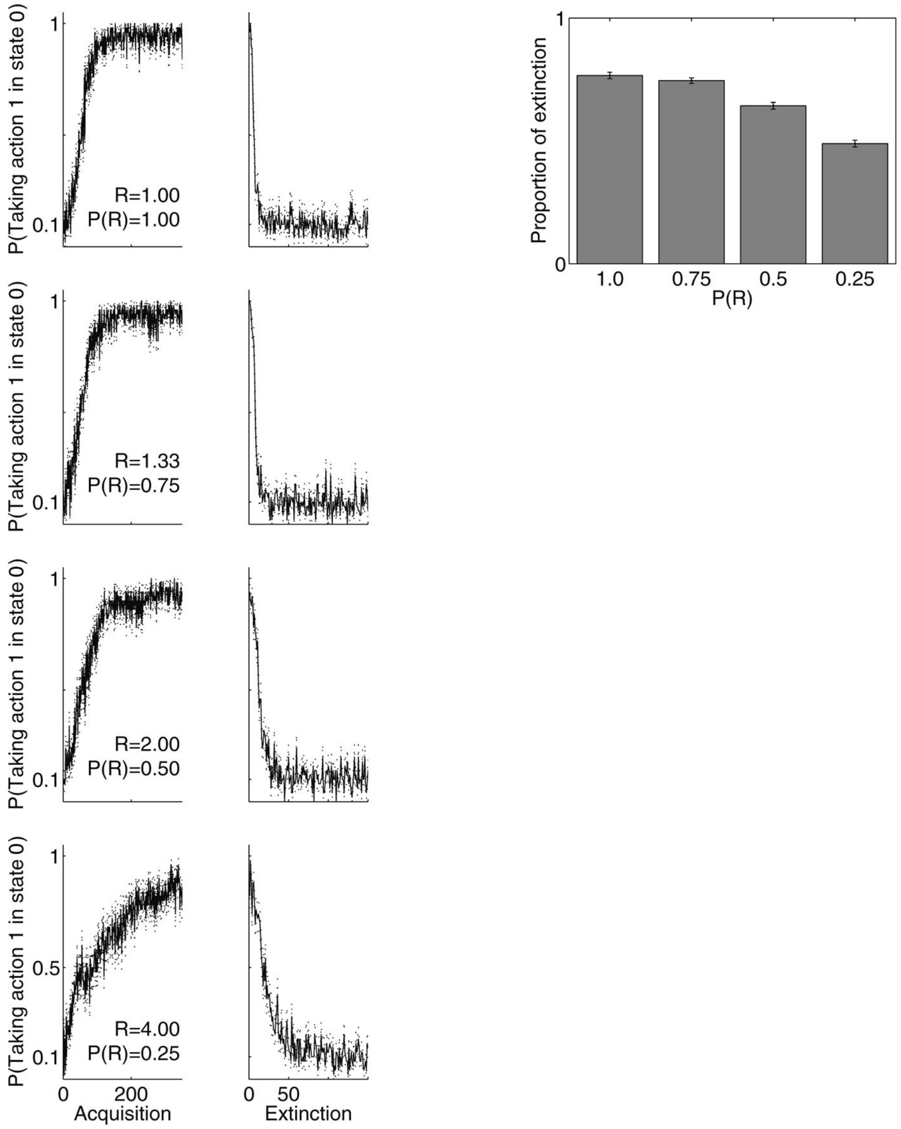

Figure 6. The partial reinforcement extinction effect. P = probability; R = reward. Error bars show standard error of the mean measured over simulations (n = 50). A color version of this figure is available on the Web at http://dx.doi.org/10.1037/[articleDOI].supp

a unique state can be recalled. Our theory suggests that the "inhibitory cues" are simply the new cues that serve to disambiguate the original state s from other states. This may explain why "inhibitory cues" are always modulatory and are often contextual (Bouton, 2004; Delameter, 2004)—they do not serve to instantiate a new state, but rather to differentiate the new state from the old state. Acquisition in these theories, however, is based on simple conditioned stimulus–unconditioned stimulus associations and do not have the explanatory power that the temporal difference model does.

Computational models in which acquisition is based on temporal difference learning rules handle extinction by unlearning (Kakade & Dayan, 2002; O'Reilly & Munakata, 2000; Pan, Schmidt, Wickens, & Hyland, 2005). Although all three of these models capture both acquisition and extinction, only the O'Reilly and Munakata (2000) model captures a renewal that is faster than acquisition. It does so by using a δ rule with threshold units. The prediction unit decreases its activity in response to missing rewards only until it drops below activation. Because the input weight matrix remains just below threshold, a single reward can retrigger responding (bringing it back above threshold). However, because the prediction unit waiting below threshold needs a positive δ signal to drive it back above threshold, the O'Reilly and Munakata model will not show a response to nonrewarded associated cues and cannot explain contextual or spontaneous recovery of responding (Bouton, 2002, 2004; Rescorla, 2004; Robbins, 1990). Because extinction occurs by unlearning of the input weights to the prediction unit, extinction speed in the O'Reilly and Munakata model is inversely proportional to the partial reinforcement rate. Contrary to the animal behavior literature (Capaldi,

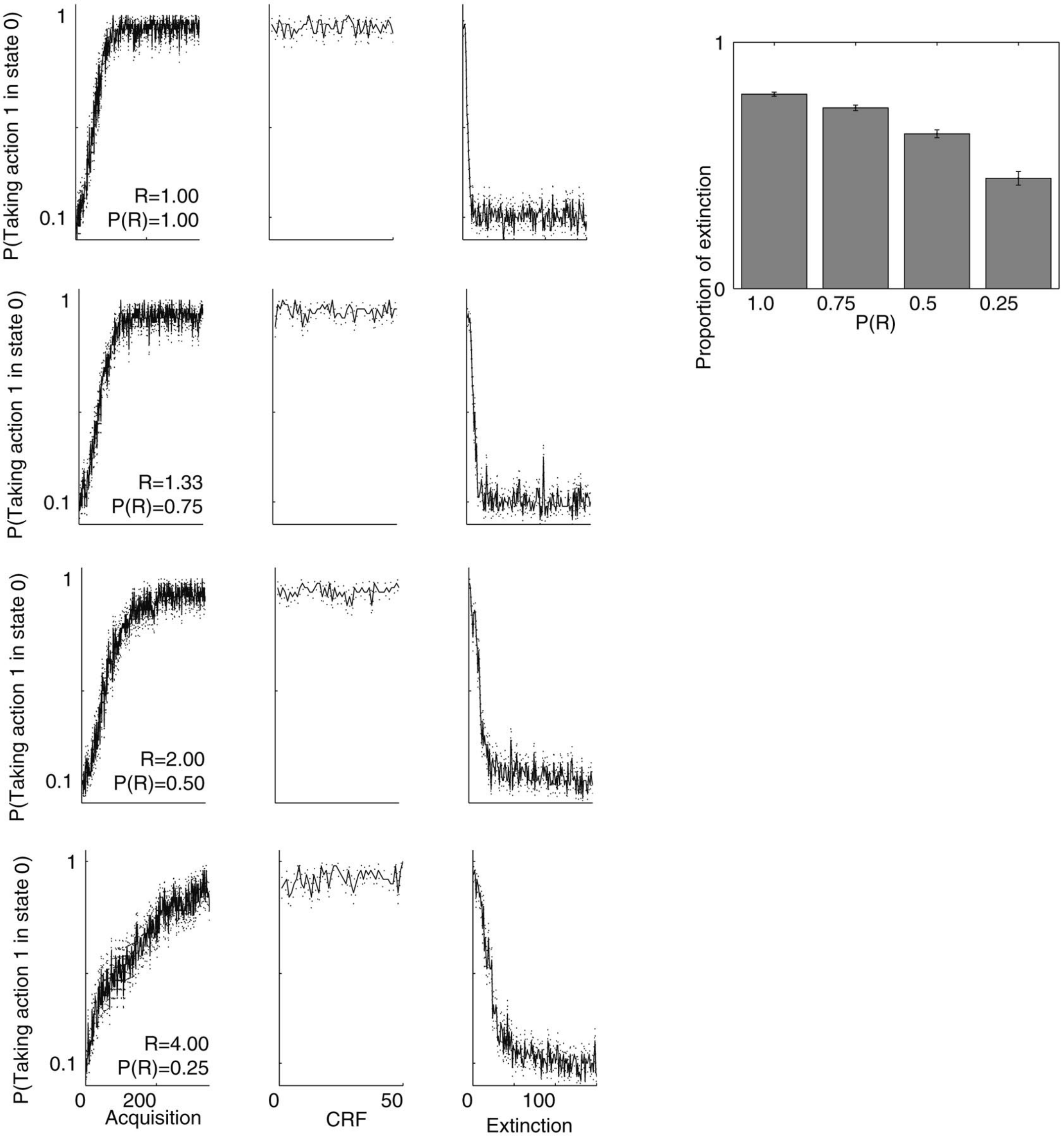

Figure 7. The partial reinforcement extinction effect continues to occur even with an intervening fully reinforced condition (CRF). P = probability; R = reward. Errors bars show standard error of the mean measured over simulations (n = 50). A color version of this figure is available on the Web at http://dx.doi.org/10.1037/[articleDOI].supp

1957; Domjan, 1998), in the O'Reilly and Munakata model, agents that receive a reinforcement at a low probability rate extinguish as fast or faster (but definitely not slower) than do animals that receive a reinforcement at high probability. This means that the O'Reilly and Munakata model cannot explain the PREE (Domjan, 1998; Pearce & Bouton, 2001). In particular, the O'Reilly and Munakata model has no long-term memory of past reinforcement schedules and so cannot carry effects across intervening reinforcement schedules (Jenkins, 1962; Theios, 1962). In contrast, our model proposes that extinction follows as a consequence of changes in the state-space representation, which provides explanations for nonrewarded cued renewal (Simulation 3), for the PREE (Simulation 4), and for the continued effect of partial reinforcement across intervening reinforcement schedules (Simulation 5).

Our state-space expansion hypothesis reconciles current psychological theories of extinction with current TDRL learning theories. Our model can be seen as a mathematical instantiation of those current psychological theories, particularly those of Capaldi (1957, 1958), who suggested that the key to extinction is the ability to discriminate between the acquisition and extinction situations. It is important to note that we have been able to model experiments that have shown that the partial reinforcement effect is not removed by an intervening continuous reinforcement schedule (Domjan, 1998; Jenkins, 1962; Theios, 1962). In our model, the agent's prior experience with partial reinforcement can be recalled during the extinction training, which reduces the discriminability of the extinction training and slows down extinction (see Figure 7). These ideas are also consistent with those of Bouton (2002, 2004), who suggested that extinction arises from secondary cues that gate new associations. Cues that discriminate between the two situations are those that differentiate the original state *s* from other states. These cues effectively gate the associations.

This theory provides immediate explanations for the context sensitivity (Bouton, 2004) and cue dependence (Bouton, 2004; Pavlov, 1927) of extinction, both of which provide cues to disambiguate state. Reinstatement by provision of unsignaled rewards (Pavlov, 1927; Rescorla, 2004; Robbins, 1990) can also be envisioned as a cued recall of the original state over other states. The presence of recent/available rewards is one of the most salient differences between states (Capaldi, 1957, 1958; Capaldi & Birmingham, 1998; Capaldi & Lynch, 1968).

What Defines a State?

Taking the correct action in a situation requires solving a classification problem—the agent must identify the set of previous situations that are similar enough to the current situation in order to make predictions about value and consequences of its actions. Once the agent has made this classification, it can make decisions about actions. Without this classification, agents will show random responses.

This classification is fundamentally a memory process and requires a model of memory encoding and retrieval. The agent has to both identify the important cues and use them to retrieve the appropriate situation categorization. Cues, particularly in the real world, are ubiquitous. Some of these cues are reliable and can provide consistent definitions of situations. Other cues are unimportant, unreliable, or inconsistent. The mechanisms that control reliability are complex and beyond the scope of this article. Clearly, an agent needs to differentiate reliable from unreliable cues. However, whether a cue becomes fundamental to the definition of a situation is likely to depend on factors beyond mere reliability, such as novelty, salience, and other factors (Collett, 1987; Gallistel, 1990; Knierim et al., 1995; Pearce & Bouton, 2001). Cues can also modulate the interpretation of other cues (Domjan, 1998). The extent to which cues serve as separate elements or as configurations is still vigorously debated (Delameter, 2004; Domjan, 1998; Pearce, 1994; Pearce & Bouton, 2001; Rescorla, 2003).

How agents use cues to recognize internal states or "situations" is a very large open question, and is beyond the scope of this article. Fundamentally, however, the final action must be unitary—either an agent takes an action or it does not. That final action reflects a categorization. Whether that unitary action also reflects a unitary state is an open question (Daw, 2003; Daw et al., in press, 2006; Doya et al., 2002). The results in this article follow from the hypothesis that this situation classification problem is fundamentally a memory process and that it requires a balance between encoding and retrieval.

Context

In our model, the state categorization is based on both the quickly changing cues (e.g., s_W) as well as slowly changing contextual cues. Because our model uses the mutual information between the cues and the states to determine the attention to the cue (cue weight, w_k; Equation 5), cues that do not change are ignored. Thus, context is ignored in Simulation 1 because it does not change at all. Even when context does change in Simulation 2, attention to it is paid only when states start to split. The definition of context is highly complex and depends on issues of spatial and temporal continuity and scale (Cassaday & Rawlins, 1997; Fuhs, VanRhoads, Casale, McNaughton, & Touretzky, 2005; Hirsh, 1974; Hirsh, Leber, & Gillman, 1978; Nadel, 1994, 1995; Nadel & Willner, 1980; Nadel et al., 1985; O'Keefe & Nadel, 1978; see Redish, 1999, for a review), but this model provides one potential explanation for the difference between contextual and standard conditioning stimuli—contextual cues change slowly and provide little information to the immediate decision making that must occur within the context. Similarly, this provides an explanation for the role of contextual cues in conditioning (Bouton, 2002, 2004)—contextual cues only come into play when agents need to differentiate actions in response to similar stimuli occurring in different contexts.

When Are Representations Split? When Are Values Unlearned?

In the model proposed in this article, $\delta < 0$ can produce two effects: It can drive the TDRL component to reduce a learned (associated) value back toward 0 (i.e., unlearning), or it can drive the state-classification component to reinterpret the set of cues as a new state (i.e., state splitting). Both of these processes are available to our model. These two processes, however, can be dissociated by their different effects on renewal. If a value is unlearned, then renewal is the same as relearning and will occur with the same time course as the original learning. In contrast, if the state classification of the cue set has changed, then the original state is still available; a process that drives the agent to use the original state classification will produce essentially instant renewal. In the simulations presented above, which process will occur is entirely dependent on the state-classification process. If the classification of the cue set changes, the original value is left intact. If the classification of the cue set does not change (i.e., the agent continues to use the original state), then the original value will be unlearned. In our simulations, the state-classification process appears faster than the unlearning process, therefore the simulated agents tend to extinguish by state reclassification rather than by unlearning.

That being said, it is likely that under some conditions, agents may prefer not to reclassify situations and will show unlearning rather than state splitting. In a recent article, Myers, Ressler, and Davis (2006) tested the contextual reinstatement of fear in rats that experienced extinction trials 10 min, 1 hr, 24 hr, or 72 hr after acquisition.[5] They found that animals in the 10-min and 1-hr groups did not show subsequent renewal (suggesting the animals had in fact forgotten the association), whereas the animals who had 24 hr or 72 hr between acquisition and extinction showed strong subsequent contextual renewal as well as a higher probability of

[5] Although this experiment addresses extinction of aversive associations, similar processes drive extinction of aversive associations.

spontaneous renewal (suggesting the animals had not forgotten the original association).

Do all memories require 24 hr to solidify such that when the animals are reexposed to the same situation, extinction proceeds by state reclassification rather than by unlearning? We find this highly unlikely. Certain extremely salient events can produce unforgettable associations, even though they occur only once. For example, people who have received shocks from implanted cardiodefibrilators often make surprising associations that can produce severe fear and anxiety (Bourke, Turkington, Thomas, McComb, & Tynan, 1997; Godemann et al., 2001; Hamner, Hunt, Gee, Garrell, & Monroe, 1999). In our current model, we have not included an explicit solidification parameter; all agents and all simulations used the same classification mechanism. State reclassification depended on $\bar{\delta}$ and tended to occur before unlearning could take place. Thus our simulated agents reclassified situations rather than unlearned associations. It is possible that this time course difference seen by Myers et al. (2006) reflects consolidation (Gais & Born, 2006; Smith, 1995; Teng & Squire, 1999), long-term potentiation (LTP) processes (Huang & Kandel, 1995; Lynch, 1998; Lynch, Rex, & Gall, in press; Sajikumar & Frey, 2004), or some other process providing stabilization of representation in the state-classification system. Extensive theories have suggested that changes in memory storage can occur over very long time scales (Cohen & Eichenbaum, 1993; Nadel & Moscovitch, 1997; Squire, 1987), but this literature is beyond the scope of this article. Memory stability would likely be related to the strength of the association, as evidenced by Wang, Marin, and Nader (2005). We suggest that the balance between extinction by reclassification and extinction by unlearning depends on the specific parameters of the association, including the time elapsed since the association was made; the salience, surprise, and magnitude of the events; as well as the internal parameters such as attention, stress, and sleep-consolidation processes. In any case, our theory implies that the key to differentiating between state-reconsolidation and unlearning processes is exactly the classic question: *Can the association be renewed*?

This viewpoint on extinction also provides an entry into the reconsolidation phenomenon: Both extinction and reconsolidation require memory retrieval (Ouyang & Thomas, 2005), and both require protein synthesis (Berman & Dudai, 2001; Eisenberg, Kobilo, Berman, & Dudai, 2003; Nader, Schafe, & LeDoux, 2000; Suzuki et al., 2004; Vianna, Szapiro, McGaugh, Medina, & Izquierdo, 2001). Recently acquired memories are susceptible to manipulations during retrieval (Berman & Dudai, 2001; Myers et al., 2006; Nader et al., 2000; Suzuki et al., 2004); however, for well-established memories, it is the extinction process that is susceptible to manipulations (Eisenberg et al., 2003). We suggest that these effects may arise from the difference between poorly established and well-established memories. Poorly established memories are susceptible to unlearning (thus an unreinforced recall combined with protein synthesis inhibitors reduces the stored association). However, extinguishing behaviors arising from well-established memories requires the storage of a new state. Thus, extinction of strongly stored memories requires new learning, and the extinction rather than the original memory is susceptible to protein synthesis inhibitors.

What Is the Effect of Extinction on Other Available Actions?

The model as presented here first categorizes the available cues into a state or situation and then decides upon the most appropriate action given that categorization. This means that in our model, extinguishing one action in response to a stimulus will have the effect of extinguishing all actions in response to the same stimulus. To our knowledge, the effect of extinction of one action in response to a cue on the taking of other actions in response to the same cue is still an open question. It is important to note that the agent's state is a categorization derived from the immediately available cues. Thus the agent's state categorization includes both the immediate conditioning stimuli (tone, light, etc.) as well as the contextual cues (environment, slowly changing cues, etc.). Just because the model splits the state derived from one stimulus in a context (e.g., tone), this does not mean that it must split the state derived from another stimulus in the same context (e.g., light).

An additional possibility is that states are never represented separately from actions: That is, representations of states are really state–action pairs. Recent evidence from midbrain dopamine recordings in monkeys suggests that phasic dopamine signals encode not just state–action pairs, but actually, the entire state–action–reward–state–action set (Morris et al., 2006; Niv et al., 2006a). If all situations are encoded as state–action pairs, then state splitting due to disappointment would (by definition) only apply to the action with disappointing consequences. States in our model as implemented are based entirely on the available cue set and do not include actions in the state definition, but our theory is agnostic as to whether states are based solely on cues or whether they include an expected (or recent) action component.

Anatomical Instantiations

The TDRL component. Most models of TDRL's anatomical instantiation are based on models of the basal ganglia in general and of the striatum in particular (Barto, 1995; Daw et al., 2006; Doya, 2000a, 2000c; Foster, Morris, & Dayan, 2000; Gurney, Prescott, & Redgrave, 2001a, 2001b; Houk, Davis, & Beiser, 1995; Johnson & Redish, 2005; Montague et al., 1996; Redgrave, Prescott, & Gurney, 1999; Samejima, Ueda, Doya, & Kimura, 2005; Suri & Schultz, 1999). Striatal neurons have been found to represent key parameters of the temporal difference reinforcement learning algorithm (e.g., situation–action associations; T. D. Barnes et al., 2005; Carelli & West, 1991; Daw, 2003; Gardiner & Kitai, 1992; Hikosaka et al., 1999; Hikosaka, Nakamura, & Nakahara, 2006; Jog, Kubota, Connolly, Hillegaart, & Graybiel, 1999; Kermadi & Joseph, 1995; Kermadi, Jurquet, Arzi, & Joseph, 1993; Matsumoto, Hanakawa, Maki, Graybiel, & Kimura, 1999; Miyachi, Hikosaka, Miyashita, Kárádi, & Rand, 1997; Schmitzer-Torbert & Redish, 2004; Tremblay, Hollerman, & Schultz, 1998), reward delivery (Daw, 2003; Schmitzer-Torbert & Redish, 2004; White & Hiroi, 1998), and value signals (Daw, 2003; Kawagoe, Takikawa, & Hidosaka, 2004; Nakahara, Itoh, Kawagoe, Takikawa, & Hikosaka, 2004). The fMRI data from humans playing sequential games show similar correlates to value, δ, and other parameters of these models in striatum (McClure et al., 2003; McClure, Laibson, Loewenstein, & Cohen, 2004; O'Doherty et al., 2004; O'Doherty, 2004; O'Doherty, Buchanan, Seymour, &

Dolan, 2006; Seymour et al., 2004; Tanaka et al., 2004). These data suggest that the TDRL component likely resides primarily within the basal ganglia.

The δ *signal.* Extensive research has implicated a role for dopamine in reward learning (Schultz, 2002; Wise, 2004). In particular that phasic bursts of dopamine signal differences between expected and observed changes in value (Bayer & Glimcher, 2005; Ljungberg et al., 1992; Schultz, 2002; Waelti et al., 2001). Unexpected rewards or reward-predicting stimuli produce phasic bursts of activity in dopaminergic neurons, whereas undelivered rewards or stimuli that predict that an expected reward will not be delivered produce pauses in the baseline firing of dopaminergic neurons (Bayer & Glimcher, 2005; Ljungberg et al., 1992; Schultz, 2002; Waelti et al., 2001). For both positive (value higher than expected, indicated by increases in dopamine-cell firing rate) and negative (value lower than expected, indicated by length of pause in firing) value-prediction errors, Bayer and Glimcher (2005) found that the dopamine signal was related to the magnitude of the error. Fast-scan voltammetry has confirmed that these changes are reflected (although transformed; Montague et al., 2004; Ungless, 2004) in extracellular dopamine levels (Stuber, Wightman, & Carelli, 2005). These data suggest that phasic bursts of dopamine are likely to carry the δ signal (Barto, 1995; Daw et al., 2006; Montague et al., 1995, 1996).

The situation-categorization component. Correct action within a situation requires the recognition that the situation is similar to previous situations (Daw et al., 2006; Sutton & Barto, 1998). This is, fundamentally, a memory-retrieval and categorization problem. Extensive work in categorization problems have implicated the cerebral cortex in such processes (e.g., Fuster, 1997; Kéri, 2003; Kohonen, 1980; Logothetis & Sheinberg, 1996).

Our theory suggests that extinction arises from changes in this categorization function. Although changes in sensory cortical representations have been seen with learning (Myers & Davis, 2002; Weinberger, 1998) and with changes in dopamine levels (Bao, Chan, & Merzenich, 2001), we find it unlikely that such changes will occur at the level of initial sensory processing. Much more likely would be to see changes in representations in structures that subserve more flexible representations, such as prefrontal cortex and hippocampus (Cohen & Eichenbaum, 1993; Fuster, 1997; O'Keefe & Nadel, 1978; Redish, 1999; Robbins, 2005). Lesion, stimulation, and recording evidence suggest a direct role for the medial prefrontal cortex in the addition of new signals to drive extinction (Lebron, Milad, & Quirk, 2004; Milad & Quirk, 2002; Quirk et al., 2006; Sotres-Bayon, Cain, & LeDoux, 2006). A role for the hippocampus in reversal learning has been known since the 1970s (Hirsh, 1974; Hirsh et al., 1978; Isaacson, 1974; Nadel & Willner, 1980; O'Keefe & Nadel, 1978). Hippocampal lesions interfere with the contextual dependence of extinction and remove the ability of context to renew responding (Bouton et al., 2006). The hippocampus provides a direct projection to medial prefrontal cortex (Ferino, Thierry, & Glowinski, 1987; Jay, Burette, & LaRoche, 1995; Jay & Witter, 2004), which may allow it to provide contextual signals to the prefrontal representation.

As noted above, current theories of extinction have suggested that extinction arises from the addition of new variables, identified as "inhibitory" (Delameter, 2004; Pavlov, 1927) or "contextual" (Bouton, 2002, 2004). The data examining prefrontal effects on amygdala associations suggest a role of prefrontal cortex in the addition of new "inhibitory" signals (Milad, Vidal-Gonzalez, & Quirk, 2004). In contrast, hippocampal activity and hippocampal integrity have both been implicated in the representation and processing of contextual cues (O'Keefe & Nadel, 1978; Redish, 1999). One possibility is that animals have two mechanisms through which the state classification can be changed: a prefrontal mechanism providing the addition of new variables to the classification problem (set shifting: Robbins, 2005; dimension augmentation: Grossberg, 1976) and a hippocampal mechanism providing a change in the systemic representation of the underlying context (remapping: C. A. Barnes, Suster, Shen, & McNaughton, 1997; Bostock, Muller, & Kubie, 1991; Redish, 1999; Sharp, Blair, Etkin, & Tzanetos, 1995; Wills, Lever, Cacucci, Burgess, & O'Keefe, 2005). In any case, we suggest that the flexibility of representations in the prefrontal cortex and hippocampus provides the animal with an ability to change the state classification function, which provides the animal with the ability to associate similar situations with new states with which new values can be associated.

We did not explicitly model frontal cortex or hippocampus in our model. Instead, the implementation of state in the model was based on a very abstract competitive-learning model of state spaces. However, many models of categorization learning have been based on a similar prototype-centered process much like that used here (Ashby & Maddox, 2005; Hertz et al., 1991; Kéri, 2003; Lakoff, 1990). This model can be conceptually translated into standard distributed (neural) models of cortex and hippocampus (Arbib, 1995; Durstewitz, Kelc, & Gunturkun, 1999; Durstewitz, Seamans, & Sejnowski, 2000; Redish, 1999; Rumelhart & McClelland, 1986; Seamans, Gorelova, Durstewitz, & Yang, 2001; Seamans & Yang, 2004).

We assume that the state hypothesis is represented through distributed encoding across an autoassociative network. The specific details of the autoassociative network are irrelevant to the general hypothesis put forward here, but we note that there is a large family of autoassociative networks with the necessary properties (Hertz et al., 1991). The key properties are (a) that cells with similar representations should be coupled with excitatory connections, which provides completion for an incomplete input (Hertz et al., 1991; Hopfield, 1982; Kohonen, 1980), (b) global, or near-global, inhibition, which provides for competition between possible representations (Grossberg, 1976; Hertz et al., 1991; Wilson & Cowan, 1973), and (c) excitatory inputs on the main cells with associative learning (e.g., LTP) across the inputs, which provides for storage and recognition of input cues (Hertz et al., 1991; Hopfield, 1982). Such networks fall into three major families: *cell assemblies*, in which small, but separate groups of cells support each other (Hebb, 1949; Marr, 1971; McNaughton & Nadel, 1990); *attractor networks*, in which a continuous representation is identified along a dimension, and the recurrent excitatory coupling follows a unimodal kernel such that cells with similar preferred values are preferentially coupled (Kohonen, 1980; Laing & Chow, 2001; Redish, 1999; Wilson & Cowan, 1973); *general autoassociators*, such as the Hopfield (1982), Willshaw (Willshaw, Bruneman, & Longuet-Higgins, 1969), or Kohonen (1980) networks, in which all cells are coupled and learning occurs across the recurrent connections. In these networks, incomplete and nearby patterns will settle to one of a subset of final patterns (Hertz et al., 1991). The stable, final states carry remembered information and can serve as representations of the world that can be used to drive value and action processing in TDRL. The set of similar patterns that settle to the same final state define a *basin of attraction*. The resistance of this final state to noise defines the depth of the basin and can be said to form a sort of inertia in the system.

We propose that once the agent has acquired a learned association, the network has formed a deep basin of attraction. When the agent repeatedly receives $\delta < 0$, some process occurs that effectively moves the currently observed cues $c(t)$ outside the basin of attraction for the memory of the situation s. This can occur either by shrinking the basin of attraction or by increasing the distance between the currently observed cues and the prototype. (The model as implemented above uses the second process.) This allows a second, slightly changed input to retain its own representation (e.g., s') and not to fall into the original basin (s). As the agent learns to correctly accommodate the new state (s), the error signal returns to 0, and the system will now have two separable basins of attraction, differentiating the two states s and s'. Further differentiation between the states will occur via the competitive learning that arises from the recurrent inhibitory connections.

Identifying changes between states. Changes between states can be directly identified neurophysiologically through the measurement of changes in tuning curves. A tuning curve is a categorization: The set of stimuli for which the cell fires is a categorization of the world. For example, a place field is a categorization of situations: Changes in place fields without changes in explicit stimuli encode a change in the categorization of the situation occurring at that location (remapping; Redish, 1999; Touretzky & Redish, 1996). This remapping can be seen after important events in an animal's experience (Moita, Rosis, Zhou, LeDoux, & Blair, 2004; Sharp et al., 1995). Dopamine D1/D5 (ant)agonists change the likelihood for place cells to remap with situational changes (Kentros, Agnihotri, Streater, Hawkins, & Kandel, 2004).

Deficits in initiation and storage of new states will also produce observable effects in behavior. An inability to recall situations (an overwillingness to create a new state) will appear as an inability to learn (i.e., responses will be random). An inability to recognize a change in situations (an underwillingness to create a new state) will appear as an inability to change responses or to extinguish those responses (i.e., as response perseveration). Experimentally, D1/D5 agonists produce response perseveration (Floresco & Phillips, 2001; Seamans & Yang, 2004; Zahrt, Taylor, Mathew, & Arnsten, 1997), whereas D1/D5 antagonists lead to random responding (Seamans, Floresco, & Phillips, 1998; Seamans & Yang, 2004).

The $\bar{\delta}$ signal. In our theory, $\bar{\delta}$ signals *disappointment*—the lack of expected reward. It is a signal that one's expectations are incorrect. We propose that agents handle this signal by assuming that the situation has changed, and a new situation categorization must be identified. This has the key benefit of keeping the previous situation category around should it become useful again.

In our model, $\bar{\delta}$ is a running estimate of $\delta < 0$. It has the effect of changing the situation-categorization component. We can thus predict that the $\bar{\delta}$ signal will control the stability of the situation-categorization component's state space. As $\bar{\delta}$ decreases (because the agent is not receiving the expected reward), the agent should become more willing to categorize the situation as new (i.e., the agent should become more willing to split the state space). If $\bar{\delta}$ is high (because the agent is receiving the expected reward), the agent should be less willing to change its situation categorization (i.e., the agent should be unwilling to categorize the situation as new). As noted above, any mechanism that increases a split in the state space can be controlled by $\bar{\delta}$. We have explored changing the exploration/exploitation parameter β, changing the acceptable width of the radial-basis function ϑ_s and changing the attention to cues w_k.

Humans and animals can show different balances between exploration and exploitation (Daw, O'Doherty, et al., 2006; Doya, 2000b). Animals faced with extinction conditions often increase their responding to the stimulus and also can show increased exploration (Domjan, 1998; Ferster & Skinner, 1957). Although increases in β with changes in $\bar{\delta}$ did not disrupt our results, they also were unnecessary to the results shown in this article. Therefore, we did not include a dependence of β on $\bar{\delta}$ in our simulations. Identifying the dependence of β on $\bar{\delta}$ will require more detailed and specific neural models and analyses.

The most obvious way to force a splitting of state would be to decrease the acceptable width of the radial-basis function ϑ_s) with decreases in $\bar{\delta}$. Thus, as the agent found itself to be losing expected reward, it would become less willing to categorize an observed set of cues as part of a known memory. However, in practice, we found this to be a highly unstable mechanism to drive state-space splitting in our simulations. Very small changes in ϑ_s made agents split states into hundreds of states and made it difficult to learn multiple environments. It is possible that with more accurate neural simulations (which are based, e.g., directly on attractor network dynamics; Durstewitz et al., 1999, 2000; Tanaka, 2006), these sorts of changes will be stable and could be a mechanism driving the influence of $\bar{\delta}$ on state spaces.

In practice, we found that the most stable mechanism to drive state splitting in our model was to increase the agent's attention to cues by increasing the w_k parameter with decreases in $\bar{\delta}$ .This made the agent more sensitive to any real changes in cues (e.g., a change in context) but also more sensitive to nonreal changes in cues (e.g., changes caused by noise).

What signals $\bar{\delta}$? Dopamine neurons pause in firing with unexpected decreases in value (Bayer & Glimcher, 2005; Ljungberg et al., 1992; Ungless, Magill, & Bolam, 2004). Thus tonic levels of dopamine could carry the necessary information—when dopamine levels tend to be low, the agent is not getting the expected value from its actions and should reconsider its expectations. However, the model that we have implemented in this article is based on changes in attention, which may be more similar to the effects of changes in norepinephrine or acetylcholine levels (Aston-Jones, Chiang, & Alexinsky, 1991; Aston-Jones & Cohen, 2005; Hasselmo, 1993; Hasselmo & Bower, 1993; Yu & Dayan, 2005). The simplest anatomical instantiation of $\bar{\delta}$ would be to have a direct consequence of tonic dopamine levels on state-space stability. Just such a model has been proposed by Seamans and Yang (2004). Both the cortex and the hippocampus have large numbers of dopamine receptors. Dopamine (ant)agonists affect representational (in)stability in both the frontal cortex (Zahrt et al., 1997) and the hippocampus (Kentros et al., 2004).

In the hippocampus, dopamine D1/D5 antagonists decrease the stability of place fields in mice across days, whereas dopamine D1/D5 agonists increase the stability of place fields (Kentros et al., 2004). As noted above, changes in the set of cells firing at a specific location are indicative of changes in the situation categorization within that environment. This would be evidenced as changes in place field stability.

In the frontal cortex, both D1 agonists and D1 antagonists impair memory function (the classic "inverted-U curve" of performance; Zahrt et al., 1997). D1 agonists produce response perseveration (Floresco & Phillips, 2001; Seamans & Yang, 2004; Zahrt et al., 1997), whereas D1 antagonists produce random responding (Seamans et al., 1998; Seamans & Yang, 2004). As noted above, response perseveration will occur when an animal is unwilling to consider new hypotheses (i.e., when it is unwilling to expand the state space). Random responses will occur when an animal was unwilling to consider that the situation is similar to previous situations (i.e., overwilling to expand the state space).

Dopamine works through diverse mechanisms to modulate excitation and inhibition within cortical networks (Arnsten, Cai, Murphy, & Goldman-Rakic, 1994; Durstewitz et al., 1999, 2000; Floresco & Phillips, 2001; Murphy, Arnsten, Goldman-Rakic, & Roth, 1996; Seamans et al., 1998, 2001; Seamans & Yang, 2004; Zahrt et al., 1997). In a recent review, Seamans and Yang (2004) concluded that slow dopamine signals to the cortex provide a "tone" that controls the depth of the cortical attractor. With high dopamine, basins of attraction become deep, and representations have more inertia and tend to be difficult to dislodge. With low dopamine, basins of attraction become shallow, and representations have less inertia and tend to shift to new representations more easily. This is also analogous to the "gain" hypothesis put forward by Servan-Schreiber, Printz, and Cohen (1990): With high dopamine, active cells tend to stay active, and inactive cells tend to remain inactive, thus making it more difficult to change the representation. This suggestion is also akin to that of Goto and Grace (2005), who suggested (in their supplemental material) that low tonic dopamine arising from undelivered rewards "may underlie the ability to switch to a new strategy for achieving a goal" (p. 2).

Goto and Grace's (2005) data also directly support our low tonic dopamine hypothesis in that decreases in tonic levels of dopamine increased the effect of prefrontal cortical stimulation on nucleus accumbens. It is interesting to note that they found no effect of changes in tonic dopamine on the hippocampal effect on nucleus accumbens. Instead, they found that increases in phasic dopamine signals increased hippocampal effects on the accumbens, whereas decreases had no significant effect. This may be because the prefrontal cortex is only recruited into the process during extinction (Milad & Quirk, 2002; Milad et al., 2004), whereas the hippocampus always provides input to accumbens functionality. D1/D5 agonists stabilized hippocampal representations (Kentros et al., 2004) and increased the effect of hippocampal stimulation on the accumbens (Goto & Grace, 2005), whereas D1/D5 antagonists destabilized hippocampal representations (Kentros et al., 2004). It may be that low tonic dopamine drives the hippocampus to find a new contextual representation that is then used to cue decision making once rewards are found again. These questions require more detailed experiments directly examining the effect of extinction on prefrontal and hippocampal representations and the effect of extinction on prefrontal and hippocampal signaling into the accumbens.

However, few studies have looked at direct measures of tonic dopamine. These studies have found tonic dopamine signals more related to uncertainty, with increases in tonic dopamine more related to decreases in certainty (Fiorillo, Tobler, & Schultz, 2003, 2005), and changes in representation (Stefani & Moghaddam, 2006), than to our $\bar{\delta}$ signal. Other theories have suggested a role for tonic dopamine as a baseline signal (such that phasic increases are measured against it; Daw, 2003; Grace, 1991, 1995; O'Reilly & Frank, 2006), as a measure of the uncertainty in the delivery of reward (Fiorillo et al., 2003), as a measure of response vigor and motivation (Niv et al., 2006a, 2006b), as an enabler of learning (such that without tonic dopamine, learning becomes fixed; Gutkin, Dehaene, & Changeux, 2006), and as having a role in the general issue of attention (Young, Moran, & Joseph, 2005).

An interesting possibility is that dopamine may actually carry three signals, a burst signal carrying $\delta > 0$, a pause signal carrying $\delta < 0$, and a tonic signal carrying another alternative signal (average reward, response vigor, attention, etc.). In our model, the key parameter would likely be the pauses, signaling value worse than expected, leading to behavioral extinction through a changing of the situation-categorization function.

Other neuromodulators. Of course, dopamine is not the only neuromodulator to play a role in learning. Many neuromodulators, including (but not limited to) acetylcholine, serotonin, dopamine, and norepinephrine, have been implicated in changes in learning (Doya, 2000b, 2002; Hasselmo, 2005; Hasselmo & Bower, 1993; Yu & Dayan, 2005).

Yu and Dayan (2005) examined the effects of uncertainty on learning models. They noted differences between expected and unexpected uncertainty. In our model, the threshold ϑ_s for categorizing a cue observation $c(t)$ with a state s and the cue-attention parameter w_k reflect expected uncertainty. In our model, $\bar{\delta}$ reflects unexpected uncertainty in that it reflects a lack of expected reward. It might be possible to obtain similar results with a direct measure of uncertainty. Yu and Dayan placed the signal for expected uncertainty in acetylcholine signaling and unexpected uncertainty in norepinephrine signaling. Yu and Dayan argued for a cholinergic role in measuring cue validity and thus attention to a cue (analogous to the role $\bar{\delta}$ plays in our w_k term). In fact, Yu and Dayan (2002) explicitly suggested a role for acetylcholine in controlling the width of a radial-basis field in a hidden Markov model of learning.

A likely candidate for the $\bar{\delta}$ signal is also norepinephrine. Both the frontal cortex and hippocampus receive noradrenergic inputs (Bouret & Sara, 2005; Jones & Moore, 1977; Kalaria et al., 1989). Noradrenergic signaling has long been associated with novelty and stimulus significance (Aston-Jones & Cohen, 2005; Bouret & Sara, 2005). Aston-Jones and colleagues (Aston-Jones et al., 1991; Aston-Jones & Cohen, 2005; Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999) suggested a role for norepinephrine in attention, particularly in the realm of incorrect trials. Bouret and Sara (2005) proposed a role for norepinephrine in reset mechanisms, particularly in the context of reversal tasks.

Finally, it is possible that there is no external single signal carrying $\bar{\delta}$. Intracellular processes could integrate the dopaminergic input and thus change the stability of cellular learning in response to repeated pauses in dopamine firing. Such mechanisms are beyond the scope of this article, but we note their possibility for completeness.

Reinforcement and Aversion

Stimuli that change responding can be categorized into four separate processes: reinforcement (positive value larger than expected—leading to increased responding), disappointment (lack of expected positive value—leading to extinction), aversion (negative value larger than expected—leading to escape), and relief (lack of expected negative value—leading to extinction). This article addresses the first two and suggests that they arise from different neurobiological mechanisms and have different neurobiological consequences. Reinforcement and aversion may arise from opponent processes. Extinction of negative associations (e.g., aversion and relief) have also been well studied (e.g., cue→shock; Lebron et al., 2004; Myers & Davis, 2002; Quirk, 2002; Quirk et al., 2006) and show very similar properties to extinction of positive associations (e.g., cue→food; Bouton, 2004; Capaldi & Birmingham, 1998; Rescorla, 2004). These similarities imply that aversive extinction is likely to also occur by a changing of the agent's representation of the world (e.g., state splitting), consistent with observations that the original association is not forgotten under aversive (fear) extinction.[6]

If aversion/relief work in a parallel manner to reinforcement/disappointment, one would predict the existence of a parallel system with phasic firing increases in response to aversive stimuli and pauses in firing in response to relief. This hypothesis implies that the single reinforcement signal δ cannot subserve both reinforcement and aversion learning. This suggests that another signal (ζ) should exist for aversion learning, with firing patterns that represent whether the event is worse than expected ($\zeta > 0$), not as bad as expected ($\zeta < 0$), or as bad as expected ($\zeta = 0$). Following our suggested mechanisms for reinforcement/disappointment (δ, $\bar{\delta}$), we hypothesized the existence of ζ and $\bar{\zeta}$ terms representing aversion and relief.

Although no current signal has been identified with this ζ term, potential candidates are likely to be monoamines like dopamine. Obvious candidates are thus serotonin (Daw, 2003; Daw et al., 2002, Daw, O'Doherty, et al., 2006) and norepinephrine. Although neither current recordings of norepinephrine signals (Aston-Jones, Rajkowski, Kubiak, & Alexinsky, 1994; Clayton, Rajkowski, Cohen, & Aston-Jones, 2004) nor current theories of norepinephrine function (Aston-Jones & Cohen, 2005; Bouret & Sara, 2005) are compatible with the ζ term, both recordings and theories are based on data from the locus coeruleus. Whether norepinephrine signals from other noradrenergic structures (e.g., A1, A2) carry aversive signals is still unknown. However, norepinephrine is known to be necessary for conditioned taste aversion (Miranda, LaLumiere, Buen, Bermudez-Rattoni, & McGaugh, 2003), and norepinephrine from the A1 and A2 noradrenergic cell groups projecting to the bed nucleus of the stria terminalis is critically involved in opiate-withdrawal aversion (Aston-Jones, Delfs, Druhan, & Zhu, 1999; Delfs, Zhu, Druhan, & Aston-Jones, 2000). Norepinephrine signals directly inhibit dopamine firing, producing distinct pauses (Paladini & Williams, 2004).

Negative stimuli, such as tail-pinch and shock, produce reliable pauses in dopamine signaling (Mirenowicz & Schultz, 1996; Ungless et al., 2004). An implication of the $\delta < 0$-implies-state-splitting hypothesis is thus that aversive learning could produce state splitting. As noted above, aversion could also be learned by either the association of an appropriate action (such as an escape), reducing ζ back to 0. Moita et al. (2004) have seen evidence of this. Moita et al. recorded hippocampal neurons from rats exposed to shock. In the case in which the shock was preceded by a stimulus (thus allowing the prediction of the oncoming shock), the place cells remained stable throughout the two conditions. But in the case in which the shock was not preceded by a stimulus (thus removing the possibility of prediction), the conditioning changed the place fields of a subset of cells. In the second (context) case, we would expect that the dopamine pauses continued to build up, which could have produced state splitting, evidenced by the change in representation of the environment by the place cell population.

It is important to note that the available actions in response to reinforcement and aversion are not necessarily parallel. With positive reinforcement, the agent has to find the action that led to reinforcement. With aversion, the agent has to find any action except the one that led to aversion. Escape paradigms may be more closely related to positive reinforcement paradigms in that they entail the finding of a single action. This may predict that uncorrelated rewards would not lead to the same sort of state splitting that Moita et al. (2004) found to occur with uncorrelated punishments. In cases of uncorrelated rewards, the agent can find any (or all actions) to identify with it. This can lead to superstitious behaviors (Domjan, 1998; Ferster & Skinner, 1957). However, in the cases of uncorrelated punishments, the lack of predictable structure leads to numerous states from which actions continue to predict shock. This may have some relevance to the phenomenon of learned helplessness (Domjan, 1998; Seligman, 1972).

Addiction and Relapse

Addictive drugs have been hypothesized to drive maladaptive decision making through pharmacological interactions with neurophysiological mechanisms evolved for normal learning systems (Everitt, Dickinson, & Robbins, 2001; Hyman, 2005; Kelley, 2004; Lowinson, Ruiz, Millman, & Langrod, 1997; Redish, 2004). This means that models of extinction should have implications for our understanding of addiction.

Although the self-administration (acquisition), extinction, reinstatement sequence has long been used as a model of addiction and relapse, the use of this model is still controversial (Kalivas & Volkow, 2005; Katz & Higgins, 2003; Shaham et al., 2003). The potential suddenness of relapse (Gawin, 1991) does suggest that the drug-seeking association is not forgotten. This suggests that the cessation of drug seeking may require the same processes that underlie extinction (a changing of the state space), even if the behavioral sequences leading to the changes are different. We suggest that, as with extinction, the key to recovery from addiction is a changing of the state-space representation, which allows different consequences (i.e., natural rewards predominating over drug rewards) and actions (i.e., not drug seeking) to be associated

[6] An intriguing possibility that follows from this parallel mechanism hypothesis is that posttraumatic stress disorder (PTSD) may be the negative counterpart to problem gambling. Whereas problem gambling arises from a strong reinforcing event that cannot be forgotten, PTSD may arise from a strong aversive event that cannot be forgotten. Studying similarities between PTSD patients and problem gamblers may lead to a deeper understanding of how the extinction process breaks down.

with the new state. The mechanism by which the categorization of situation changes might be different (i.e., it might depend on $\bar{\delta} < 0$ for animals experiencing extinction training, but might depend on top-down executive functions in humans attempting to break an addiction), but the consequences are the same: Similar contexts and cues are associated with alternate representations.

Relapse, then, occurs when the neural representation falls back into the old state, returning to the original representation, which leads to an overvalued addictive path to drug use (Redish, 2004). As with extinction processes, this implies that relapse will be particularly sensitive to context and other cues that can drive the representation back to the original representation. Consistent with these predictions, drug craving and relapse are strongly influenced by drug-associated cues (Childress et al., 1988, 1992, 1993) and by context (O'Brien et al., 1992). This learning-theory explanation of relapse is independent of whether the association produces positive desire for drugs (Jaffe, 1992; O'Brien et al., 1992; Redish, 2004; Wise, 2004) or negative symptoms that need to be relieved (Jaffe, 1992; Koob & Le Moal, 2001; O'Brien et al., 1992). In either case, relapse occurs when the representation returns to the original state s and makes the pathway to drug use available again.

Two other factors that increase the probability of relapse are reexposure to the drug (Shaham et al., 2003) and stressors (Sinha, Catapano, & O'Malley, 1999). Reexposure to the drug will have detectable internal effects (Kamien, Bickel, Hughes, Higgins, & Smith, 1993; Meyer & Mirin, 1979; Tarter, Ammerman, & Ott, 1998), the set of which provide strong cues to differentiate state. Stress may carry cue signals that relate to drug seeking (driven by the postdrug withdrawal, which often drives immediate drug seeking; Gawin, 1991; Koob & Le Moal, 2001). It is interesting that stress does not increase the likelihood of reinstantiation of food-seeking behavior after extinction of lever pressing for food and only reinstates drug seeking in the drug-taking environment (Shalev, Highfield, Yap, & Shaham, 2000).

Whether or not the cessation of drug seeking is well modeled by extinction, we suggest that non-drug-seeking addicts are not seeking drugs for the same reason that animals are not behaving after extinction training: They have represented the world by different states that have different associated consequences and values. Relapse, then, occurs through the same mechanism as reinstantiation of extinguished behaviors—a return to the original representation s, which leads to a return to drug seeking.

Our tonic dopamine hypothesis suggests that recovered addicts will be particularly susceptible to relapse if they have low dopamine levels. In the tonic dopamine hypothesis presented above, high dopamine makes the current state stable, whereas low dopamine makes the current state unstable. In this case, the new non-drug-seeking state is the current state, thus low dopamine would increase instability, making the agent more likely to fall out of the current state s' and more likely to relapse to the original state s.

Problem Gambling

The different consequences of unexpected gains ($\delta > 0$, producing acquisition) and unexpected losses ($\delta < 0$, leading to decreasing $\bar{\delta}$, producing state splitting) lead to important consequences of variable reward processes, such as gambling: A sequence of unexpected wins can produce anomalous expectations that will not be unlearned by subsequent losses.

Many compulsive gamblers start with a very big strike or a statistically unlikely sequence of wins and then are faced with many small losses (Custer, 1984; Wagenaar, 1988). A big unexpected strike would produce a very large positive δ and would create a general association with the actions leading to the reward. Presumably, the large strike would be a very salient event. Because subsequent losses lead to $\bar{\delta} < 0$, which leads to state splitting, subsequent losses will not remove the expectation of winning, but will instead produce alternate, special-case states identified with the losses.

We have modeled this using our basic world (Figure 2). Agents were first provided with 250 time steps in which the costs of playing outweighed the rewards. Then agents experienced 100 time steps in which rewards dramatically outweighed the costs. Finally, agents experienced 250 time steps in the original condition (costs outweighed rewards). Cues remained identical under the three conditions. We found that the larger the payout, the more likely the agent was to split states in response to the subsequent losses. (See Figure 8.) In our simulations, this occurred because the larger the payout, the more likely $\bar{\delta}$ was to drive state-space splitting before the payout could be unlearned. Once the states split, the agent was trapped in a relapse–extinction sequence in which they responded at high rates, quit, and then responded again. In particular, even a single small win could produce a dramatic relapse to responding, which was not seen before the winning streak (Figure 8).

In order to simulate the cost of playing, a small cost was applied to each time the agent took Action 1 in any situation. After a large payout, agents continued to play, even at substantial cost. The most significant predictor of the size of the postwin losses was the total amount won during the winning streak (stepwise linear regression, coefficient non-zero, $P < 10^{-10}$, reduced root-mean-square error [RMSE] from 13.30 to 9.90). Other factors that played a predictive role included the variance in the number of wins (Step 2, coefficient non-zero, $P < 10^{-8}$, reduced RMSE to 9.50), and the number of wins (Step 3, coefficient non-zero, $P < 10^{-7}$, reduced RMSE to 8.70). Also considered, but not significant, were constant factors ($p > .99$) and the variance in the payout ($p > .39$). Of course, the key to the value-association process is not the actual dollar amount of the payout, but the internal representation within the agent. Thus, agents who experience rewards more strongly than do others may be more susceptible to problem gambling. Our model suggests that anyone can become a problem gambler but that the specifics will depend on the sequence of wins and losses experienced by the agent and the agent's reactions to those wins and losses.

The identification of losses with a new state s' that must be differentiated from an original state s may potentially explain the effect of "hindsight bias" (Custer, 1984; Langer & Roth, 1975; Wagenaar, 1988), in which gamblers analyze their losses and explain them away by noting what went wrong and why they should have known they would lose. Similarly, this theory explains the "illusion of control," in which gamblers believe they can control statistical situations (Custer, 1984; Elster, 1999; Wagenaar, 1988), as a misclassification of a single statistical situation into multiple differentiable situations (i.e., as a problem with the situation-categorization component).

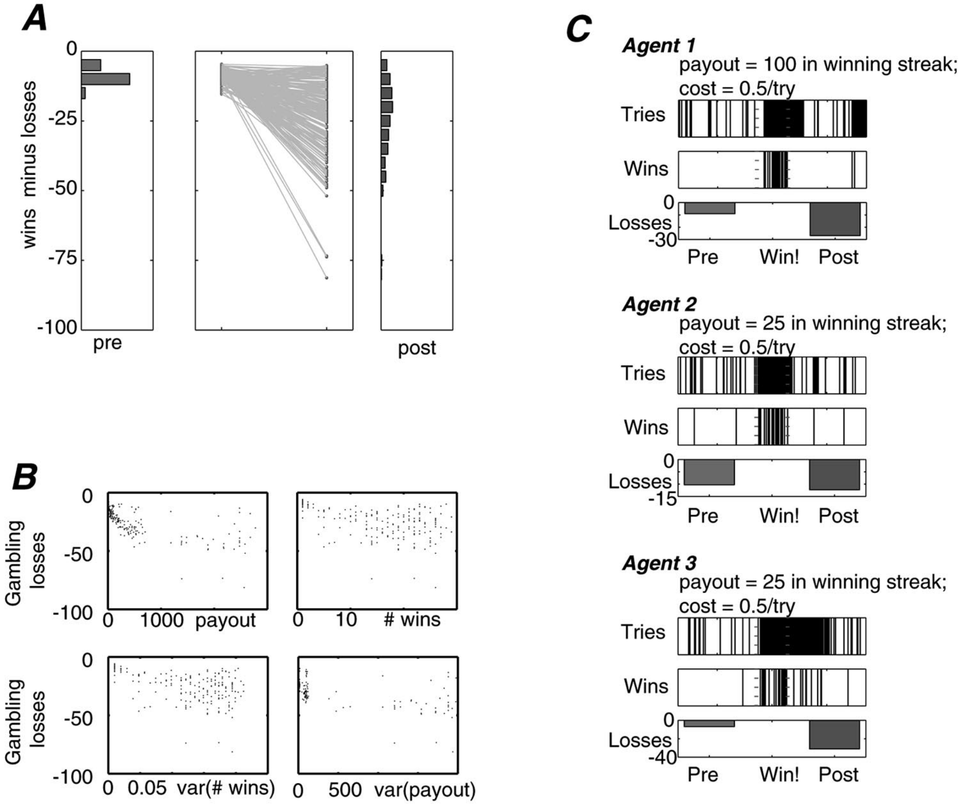

Figure 8. A model of problem gambling. A: distributions of costs before and after wins. Because there is a cost with each try, naive agents will show a small loss (pre) as a result of exploration. The histograms show the distribution of responding for pre (before the winning sequence) and post (after the winning sequence). B: relationship between postwinning losses and parameters of the winning sequence. C: Three examples of agents. var = variance. A color version of this figure is available on the Web at http://dx.doi.org/10.1037/[articleDOI].supp

Unlike addicts, who know how to get back to their high-value goals but do not wish to, gamblers do not know how to get back to their high-value goals but are continuously trying to do so. Gamblers believe there is a way to achieve their earlier success if they could only find it. A stereotypical problem gambler is always trying to recapture his original success. The gambler is stuck, trying to differentiate *s* from other states, never able to achieve that magical, original state *s* that once led to the large reward.

Summary and Conclusion

TDRL models require two processes: a situation-categorization process, in which cues are categorized into situations, and an association process, in which values are associated with those situations. We hypothesize that the situation-categorization process reacts to the lack of delivery of expected reward by a recategorization process (i.e., a "splitting" of the state space). We show that the inclusion of this explicit situation-categorization process enables models of acquisition, extinction, and renewal compatible with the experimental literature. It provides explanations for effects of protein synthesis inhibitors on reconsolidation and extinction. It also provides explanations for relapse in addiction as well as problem gambling. This model requires a signal identifying the lack of delivered reward. We hypothesize that that signal is most likely to be reflected in tonic levels of dopamine.

References

Arbib, M. (Ed.). (1995). *The handbook of brain theory and neural networks.* Cambridge, MA: MIT Press.

Arnsten, A. F. T., Cai, J. X., Murphy, B. L., & Goldman-Rakic, P. S. (1994). Dopamine d_1 receptor mechanisms in the cognitive performance of young adult and aged monkeys. *Psychopharmacology, 116,* 143–151.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Reviews: Psychology, 56,* 149–178.

Aston-Jones, G., Chiang, C., & Alexinsky, T. (1991). Discharge of noradrenergic locus coeruleus neurons in behaving rats and monkeys suggests a role in vigilance. *Progress in Brain Research, 88,* 501–520.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28,* 403–450.

Aston-Jones, G., Delfs, J. M., Druhan, J., & Zhu, Y. (1999). The bed nucleus of the stria terminalis: A target site for noradrenergic actions in opiate withdrawal. *Annals of the New York Academy of Sciences, 877,* 486–498.

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *Journal of Neuroscience, 14,* 4467–4480.

Bao, S., Chan, V. T., & Merzenich, M. M. (2001, July 5). Cortical remodeling induced by activity of ventral tegmental dopamine neurons. *Nature, 412,* 79–83.

Barnes, C. A., Suster, M. S., Shen, J., & McNaughton, B. L. (1997, July 17). Multistability of cognitive maps in the hippocampus of old rats. *Nature, 388,* 272–275.

Barnes, T. D., Kubota, Y., Hu, D., Jin, D. Z., & Graybiel, A. M. (2005, October 20). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature, 437,* 1158–1161.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Bayer, H. M., & Glimcher, P. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47,* 129–141.

Berman, D. E., & Dudai, Y. (2001, March 23). Memory extinction, learning anew, and learning the new: Dissociations in the molecular machinery of learning in cortex. *Science, 291,* 2417–2419.

Bishop, C. M. (1995). *Neural networks for pattern recognition.* New York: Oxford University Press.

Bostock, E., Muller, R. U., & Kubie, J. L. (1991). Experience-dependent modifications of hippocampal place cell firing. *Hippocampus, 1,* 193–206.

Bouret, S., & Sara, S. J. (2005). Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences, 28,* 574–582.

Bourke, J. P., Turkington, D., Thomas, G., McComb, J. M., & Tynan, M. (1997). Florid psychopathology in patients receiving shocks from implanted cardioverter-defibrillators. *Heart, 78,* 581–583.

Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry, 52,* 976–986.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning and Memory, 11,* 485–494.

Bouton, M. E., Westbrook, R. F., Corcoran, K. A., & Maren, S. (2006). Contextual and temporal modulation of extinction: Behavioral and biological mechanisms. *Biological Psychiatry, 60,* 352–360.

Capaldi, E. J. (1957). The effect of different amounts of alternating partial reinforcement on resistance to extinction. *American Journal of Psychology, 70,* 451–452.

Capaldi, E. J. (1958). The effect of different amounts of training on the resistance to extinction of different patterns of partially reinforced responses. *Journal of Comparative and Physiological Psychology, 51,* 367–371.

Capaldi, E. J., & Birmingham, K. M. (1998). Reward-produced memories regulate memory-discrimination learning, extinction, and other forms of discrimination learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 254–264.

Capaldi, E. J., & Lynch, A. D. (1968). Magnitude of partial reward and resistance to extinction: Effect of n–r transitions. *Journal of Comparative and Physiological Psychology, 65,* 179–181.

Carelli, R. M., & West, M. O. (1991). Representation of the body by single neurons in the dorsolateral striatum of the awake, unrestrained rat. *Journal of Comparative Neurology, 309,* 231–249.

Cassaday, H. J., & Rawlins, J. N. P. (1997). The hippocampus, objects, and their contexts. *Behavioral Neuroscience, 111,* 1228–1244.

Childress, A. R., Ehrman, R., Rohsenow, D. J., Robbins, S. J., & O'Brien, C. P. (1992). Classically conditioned factors in drug dependence. In J. H. Lowinson, P. Ruiz, & R. B. Millman (Eds.), *Substance abuse: A comprehensive textbook* (pp. 56–69). Baltimore: Williams & Wilkins.

Childress, A. R., Hole, A. V., Ehrman, R. N., Robbins, S. J., McLellan, A. T., & O'Brien, C. P. (1993). Cue reactivity and cue reactivity interventions in drug dependence. *NIDA Research Monographs, 137,* 73–94.

Childress, A. R., McLellan, A. T., Ehrman, R., & O'Brien, C. P. (1988). Classically conditioned responses in opioid and cocaine dependence: A role in relapse? *NIDA Research Monographs, 84,* 25–43.

Clayton, E. C., Rajkowski, J., Cohen, J. D., & Aston-Jones, G. (2004). Phasic activation of monkey locus coeruleus neurons by simple decisions in a forced-choice task. *Journal of Neuroscience, 24,* 9914–9920.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system.* Cambridge, MA: MIT Press.

Collett, T. S. (1987). The use of visual landmarks by gerbils: Reaching a goal when landmarks are displaced. *Journal of Comparative Physiology, 160*(A), 109–113.

Custer, R. L. (1984). Profile of the pathological gambler. *Journal of Clinical Psychiatry, 45*(12), 35–38.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications.* Unpublished doctoral dissertation, Carnegie Mellon University.

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., Courville, A. C., & Touretzky, D. S. (2002). Timing and partial observability in the dopamine system. *Neural Information Processing Systems, 15,* 99–106.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks, 15,* 603–616.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006, June 15). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience.* Cambridge, MA: MIT Press.

Delameter, A. R. (2004). Experimental extinction in Pavlovian conditioning: Behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 57*(B), 97–132.

Delfs, J. M., Zhu, Y., Druhan, J. P., & Aston-Jones, G. (2000, January 27). Noradrenaline in the ventral forebrain is critical for opiate withdrawal-induced aversion. *Nature, 403,* 430–434.

Domjan, M. (1998). *The principles of learning and behavior* (4th ed.). Boston: Brooks/Cole.

Doya, K. (2000a). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology, 10,* 732–739.

Doya, K. (2000b). Meta-learning, neuromodulation, and emotion. In G. Hatano, N. Okada, & H. Tanabe (Eds.), *Affective minds* (pp. 101–104). Amsterdam: Elsevier.

Doya, K. (2000c). Reinforcement learning in continuous time and space. *Neural Computation, 12,* 219–245.

Doya, K. (2002). Meta-learning and neuromodulation. *Neural Networks, 15,* 495–506.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification.* New York: Wiley.

Durstewitz, D., Kelc, M., & Gunturkun, O. (1999). A neurocomputational theory of the dopaminergic modulation of working memory functions. *Journal of Neuroscience, 19,* 2807–2822.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of Neurophysiology, 83,* 1733–1750.

Eisenberg, M., Kobilo, T., Berman, D. E., & Dudai, Y. (2003, August 22). Stability of retrieved memory: Inverse correlation with trace dominance. *Science, 301,* 1102–1104.

Elster, J. (1999). Gambling and addiction. In J. Elster & O.-J. Skog (Eds.), *Getting hooked* (pp. 208–234). New York: Cambridge University Press.

Everitt, B. J., Dickinson, A., & Robbins, T. W. (2001). The neuropsychological basis of addictive behavior. *Brain Research Reviews, 36,* 129–138.

Ferino, F., Thierry, A. M., & Glowinski, J. (1987). Anatomical and electrophysiological evidence for a direct projection from Ammon's horn to the medial prefrontal cortex in the rat. *Experimental Brain Research, 65,* 421–426.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement.* New York: Appleton-Century-Crofts.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003, March 21). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299,* 1898–1902.

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

Floresco, S. B., & Phillips, A. G. (2001). Delay-dependent modulation of memory retrieval by infusion of a dopamine Dl agonist into the rat medial prefrontal cortex. *Behavioral Neuroscience, 115,* 934–939.

Foster, D. J., Morris, R. G. M., & Dayan, P. (2000). A model of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus, 10,* 1–6.

Fuhs, M. C., VanRhoads, S. R., Casale, A. E., McNaughton, B., & Touretzky, D. S. (2005). Influence of path integration versus environmental orientation on place cell remapping between visually identical environments. *Journal of Neurophysiology, 94,* 2603–2616.

Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe* (3rd ed.). Baltimore: Lippincott Williams & Wilkins.

Gais, S., & Born, J. (2006). Sleep after learning aids memory recall. *Learning and Memory, 13,* 259–262.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences, USA, 101,* 13124–13131.

Gardiner, T. W., & Kitai, S. T. (1992). Single-unit activity in the globus pallidus and neostriatum of the rat during performance of a trained head movement. *Experimental Brain Research, 88,* 517–530.

Gawin, F. H. (1991, April 11). Cocaine addiction: Psychology and neuropsychology. *Science, 251,* 1580–1586.

Godemann, F., Ahrens, B., Behrens, S., Berthold, R., Gandor, C., Lampe, F., & Linden, M. (2001). Classic conditioning and dysfunctional cognitions in patients with panic disorder and agoraphobia treated with an implantable cardioverter/defibrillator. *Psychosomatic Medicine, 63,* 231–238.

Goto, Y., & Grace, A. A. (2005). Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. *Nature Neuroscience, 8,* 805–812.

Grace, A. A. (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, A. A. (1995). The tonic/phasic model of dopamine system regulation: Its relevance for understanding how stimulant abuse can alter basal ganglia function. *Drug and Alcohol Dependence, 37,* 111–129.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23,* 121–134.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics, 84,* 401–410.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics, 84,* 411–423.

Gutkin, B. S., Dehaene, S., & Changeux, J.-P. (2006). A neurocomputational hypothesis for nicotine addiction. *Proceedings of the National Academy of Sciences, USA, 103,* 1106–1111.

Hamner, M., Hunt, N., Gee, J., Garrell, R., & Monroe, R. (1999). PTSD and automatic implantable cardioverter defibrillators. *Psychosomatics, 40,* 82–85.

Hasselmo, M. E. (1993). Acetylcholine and learning in a cortical associative memory. *Neural Computation, 5,* 32–44.

Hasselmo, M. E. (2005). Expecting the unexpected: Modeling of neuromodulation. *Neuron, 46,* 526–528.

Hasselmo, M. E., & Bower, J. M. (1993). Acetylcholine and memory. *Trends in Neurosciences, 16,* 218–222.

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation.* Reading, MA: Addison Wesley.

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., et al. (1999). Parallel neural networks for learning sequential procedures. *Trends in Neurosciences, 22,* 464–471.

Hikosaka, O., Nakamura, K., & Nakahara, H. (2006). Basal ganglia orient eyes to reward. *Journal of Neurophysiology, 95,* 567–584.

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioral Biology, 12,* 421–444.

Hirsh, R., Leber, B., & Gillman, K. (1978). Fornix fibers and motivational states as controllers of behavior: A study stimulated by the contextual retrieval theory. *Behavioral Biology, 22,* 463–478.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA, 79,* 2554–2558.

Houk, J. C., Davis, J. L., & Beiser, D. G. (Eds.). (1995). *Models of information processing in the basal ganglia.* Cambridge, MA: MIT Press.

Huang, Y., & Kandel, E. (1995). D1/D5 receptor agonists induce a protein synthesis-dependent late potentiation in the CA1 region of the hippocampus. *Proceedings of the National Academy of Sciences, USA, 92,* 2446–2450.

Hyman, S. E. (2005). Addiction: A disease of learning and memory. *American Journal of Psychiatry, 162,* 1414–1422.

Isaacson, R. L. (1974). *The limbic system.* New York: Plenum Press.

Jaffe, J. H. (1992). Current concepts of addiction. In C. P. O'Brien & J. H. Jaffe (Eds.), *Research publications: Association for research in nervous and mental disease* (Vol. 70, pp. 1–21). New York: Raven Press.

Jay, T. M., Burette, F., & Laroche, S. (1995). NMDA receptor-dependent long-term potentiation in the hippocampal afferent fibre system to the prefrontal cortex in the rat. *European Journal of Neuroscience, 7,* 247–250.

Jay, T. M., & Witter, M. P. (2004). Distribution of hippocampal CA1 and subicular efferents in the prefrontal cortex of the rat studied by means of anterograde transport of *Phaseolus vulgaris*-leucoagglutinin. *Journal of Comparative Neurology, 313,* 574–586.

Jenkins, H. M. (1962). Resistance to extinction when partial reinforcement is followed by regular reinforcement. *Journal of Experimental Psychology, 64,* 441–450.

Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V., & Graybiel, A. M. (1999, November 26). Building neural representations of habits. *Science, 286,* 1746–1749.

Johnson, A., & Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks, 18,* 1163–1171.

Jones, B. E., & Moore, R. Y. (1977). Ascending projections of the locus coeruleus in the rat. ii. Autoradiographic study. *Brain Research, 127,* 23–53.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review, 109,* 544–553.

Kalaria, R. N., Andorn, A. C., Tabaton, M., Whitehouse, P. J., Harik, S. I., & Unnerstall, J. R. (1989). Adrenergic receptors in aging and Alzheimer's disease: Increased beta 2-receptors in prefrontal cortex and hippocampus. *Journal of Neurochemistry, 53,* 1772–1881.

Kalivas, P. W., & Volkow, N. D. (2005). The neural basis of addiction: A pathology of motivation and choice. *American Journal of Psychiatry, 162,* 1403–1413.

Kamien, J. B., Bickel, W. K., Hughes, J. R., Higgins, S. T., & Smith, B. J. (1993). Drug discrimination by humans compared to nonhumans: Current status and future directions. *Psychopharmacology, 111,* 259–270.

Katz, J. L., & Higgins, S. T. (2003). The validity of the reinstatement model of craving and relapse. *Psychopharmacology, 168,* 21–30.

Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Reward-predicting activity of dopamine and caudate neurons—A possible mechanism of motivational control of saccadic eye movement. *Journal of Neurophysiology, 91,* 1013–1024.

Kelley, A. E. (2004). Memory and addiction: Shared neural circuitry and molecular mechanisms. *Neuron, 44,* 161–179.

Kentros, C. G., Agnihotri, N. T., Streater, S., Hawkins, R. D., & Kandel, E. R. (2004). Increased attention to spatial context increases both place field stability and spatial memory. *Neuron, 42,* 283–295.

Kéri, S. (2003). The cognitive neuroscience of category learning. *Brain Research Reviews, 43,* 85–109.

Kermadi, I., & Joseph, J. P. (1995). Activity in the caudate nucleus of monkey during spatial sequencing. *Journal of Neurophysiology, 74,* 911–933.

Kermadi, I., Jurquet, Y., Arzi, M., & Joseph, J. (1993). Neural activity in the caudate nucleus of monkeys during spatial sequencing. *Experimental Brain Research, 94,* 352–356.

Knierim, J. J., Kudrimoti, H. S., & McNaughton, B. L. (1995). Place cells, head direction cells, and the learning of landmark stability. *Journal of Neuroscience, 15,* 1648–1659.

Kohonen, T. (1980). *Content-addressable memories.* New York: Springer.

Koob, G. F., & Le Moal, M. (2001). Drug addiction, dysregulation of reward, and allostasis. *Neuropsychopharmacology, 24,* 97–129.

Laing, C. R., & Chow, C. C. (2001). Stationary bumps in networks of spiking neurons. *Neural Computation, 13,* 1473–1494.

Lakoff, G. (1990). *Women, fire, and dangerous things.* Chicago: University of Chicago Press.

Langer, E. J., & Roth, J. (1975). Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. *Journal of Personality and Social Psychology, 32,* 951–955.

Lebron, K., Milad, M. R., & Quirk, G. J. (2004). Delayed recall of fear extinction in rats with lesions of ventral medial prefrontal cortex. *Learning and Memory, 11,* 544–548.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology, 67,* 145–163.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Reviews: Neuroscience, 19,* 577–621.

Lowinson, J. H., Ruiz, P., Millman, R. B., & Langrod, J. G. (Eds.). (1997). *Substance abuse: A comprehensive textbook* (3rd ed.). Baltimore: Williams & Wilkins.

Lynch, G. (1998). Memory and the brain: Unexpected chemistries and a new pharmacology. *Neurobiology of Learning and Memory, 70,* 82–100.

Lynch, G., Rex, C. S., & Gall, C. M. (2007). LTP consolidation: Substrates, explanatory power, and functional significance. *Neuropharmacology, 52,* 12–23.

Marr, D. (1971). Simple memory: A theory of archicortex. *Philosophical Transactions of the Royal Society of London, 262,* 23–81.

Matsumoto, N., Hanakawa, T., Maki, S., Graybiel, A. M., & Kimura, M. (1999). Role of nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *Journal of Neurophysiology, 82,* 978–998.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron, 38,* 339–346.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004, October 15). Separate neural systems value immediate and delayed monetary rewards. *Science, 306,* 503–507.

McNaughton, B. L., & Nadel, L. (1990). Hebb-Marr networks and the neurobiological representation of action in space. In M. A. Gluck & D. E. Rumelhart (Eds.), *Neuroscience and connectionist theory* (pp. 1–63). Hillsdale, NJ: Erlbaum.

Meyer, R., & Mirin, S. (1979). *The heroin stimulus.* New York: Plenum Press.

Milad, M. R., & Quirk, G. J. (2002, November 7). Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature, 420,* 70–74.

Milad, M. R., Vidal-Gonzalez, I., & Quirk, G. J. (2004). Electrical stimulation of medial prefrontal cortex reduces conditioned fear in a temporally specific manner. *Behavioral Neuroscience, 118,* 389–394.

Miranda, M. I., LaLumiere, R. T., Buen, T. V., Bermudez-Rattoni, F., & McGaugh, J. L. (2003). Blockade of noradrenergic receptors in the basolateral amygdala impairs taste memory. *European Journal of Neuroscience, 18,* 2605–2610.

Mirenowicz, J., & Schultz, W. (1996, February 1). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature, 379,* 449–451.

Miyachi, S., Hikosaka, O., Miyashita, K., Kárádi, Z., & Rand, M. K. (1997). Differential roles of monkey striatum in learning of sequential hand movement. *Experimental Brain Research, 115,* 1–5.

Moita, M. A., Rosis, S., Zhou, Y., LeDoux, J. E., & Blair, H. T. (2004). Putting fear in its place: Remapping of hippocampal place cells during fear conditioning. *Journal of Neuroscience, 24,* 7015–7023.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995, October 26). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature, 377,* 725–728.

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.

Montague, P. R., McClure, S. M., Baldwin, P. R., Phillips, P. E. M., Budygin, E. A., Stuber, G. D., et al. (2004). Dynamic gain control of dopamine delivery in freely moving animals. *Journal of Neuroscience, 24,* 1754–1759.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience, 9,* 1057–1063.

Murphy, B. L., Arnsten, A. F. T., Goldman-Rakic, P. S., & Roth, R. H. (1996). Increased dopamine turnover in the prefrontal cortex impairs spatial working memory performance in rats and monkeys. *Proceedings of the National Academy of Sciences, USA, 93,* 1325–1329.

Myers, K. M., & Davis, M. (2002). Behavioral and neural analysis of extinction. *Neuron, 36,* 567–584.

Myers, K. M., Ressler, K. J., & Davis, M. (2006). Different mechanisms of fear extinction dependent on length of time since fear acquisition. *Learning and Memory, 13,* 216–233.

Nadel, L. (1994). Multiple memory systems: What and why, an update. In D. L. Schacter & E. Tulving (Eds.), *Memory systems* (pp. 39–64). Cambridge, MA: MIT Press.

Nadel, L. (1995). The role of the hippocampus in declarative memory: A commentary on Zola-Morgan, Squire, and Ramus, 1994. *Hippocampus, 5,* 232–234.

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7,* 217–227.

Nadel, L., & Willner, J. (1980). Context and conditioning: A place for space. *Physiological Psychology, 8,* 218–228.

Nadel, L., Willner, J., & Kurz, E. M. (1985). Cognitive maps and environmental context. In P. D. Balsam & A. Tomie (Eds.), *Context and learning* (pp. 385–406). Hillsdale, NJ: Erlbaum.

Nader, K., Schafe, G. E., & LeDoux, J. E. (2000, August 17). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature, 406,* 722–726.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron, 41,* 269–280.

Niv, Y., Daw, N. D., & Dayan, P. (2006a). Choice values. *Nature Neuroscience, 9,* 987–988.

Niv, Y., Daw, N. D., & Dayan, P. (2006b). How fast to work: Response vigor, motivation and tonic dopamine. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems* (pp. 1019–1026). Cambridge, MA: MIT Press.

O'Brien, C. P., Childress, A. R., McLellan, A. T., & Ehrman, R. (1992). A learning model of addiction. In C. P. O'Brien & J. H. Jaffe (Eds.), *Research Publications: Association for Research in Nervous and Mental Disease* (Vol. 70, pp. 157–177). New York: Raven Press.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology, 14,* 769–776.

O'Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron, 49,* 157–166.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004, April 16). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304,* 452–454.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map.* Oxford, England: Clarendon Press.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation, 18,* 282–328.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience.* Cambridge, MA: MIT Press.

Ouyang, M., & Thomas, S. A. (2005). A requirement for memory retrieval during and after long-term extinction learning. *Proceedings of the National Academy of Sciences, USA,102,* 9347–9352.

Paladini, C. A., & Williams, J. T. (2004). Noradrenergic inhibition of midbrain dopamine neurons. *Journal of Neuroscience, 24,* 4568–4575.

Pan, W., Schmidt, R., Wickens, J., & Hyland, B. I. (2005). *A unified theory of extinction: Suggested by midbrain dopamine cell activity and temporal difference model*. Society for Neuroscience Annual Meeting. Program No. 69.18.

Paulus, M. P., Feinstein, J. S., Tapert, S. F., & Liu, T. T. (2004). Trend detection via temporal difference model predicts inferior prefrontal cortex activation during acquisition of advantageous action selection. *NeuroImage, 21,* 733–743.

Pavlov, I. (1927). *Conditioned reflexes.* New York: Oxford University Press.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101,* 587–607.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology, 52,* 111–139.

Quirk, G. J. (2002). Memory for extinction of conditioned fear is long-lasting and persists following spontaneous recovery. *Learning and Memory, 9,* 402–407.

Quirk, G. J., Garcia, R., & González-Lima, F. (2006). Prefrontal mechanisms in extinction of conditioned fear. *Biological Psychiatry, 60,* 337–343.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience, 89,* 1009–1023.

Redish, A. D. (1999). *Beyond the cognitive map: From place cells to episodic memory.* Cambridge, MA: MIT Press.

Redish, A. D. (2004, December 10). Addiction as a computational process gone awry. *Science, 306,* 1944–1947.

Rescorla, R. A. (2003). Elemental and configural encoding of the conditioned stimulus. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 56*(B), 161–176.

Rescorla, R. A. (2004). Spontaneous recovery. *Learning and Memory, 11,* 501–509.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokesy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Robbins, S. J. (1990). Mechanisms underlying spontaneous recovery in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes, 16,* 235–249.

Robbins, T. W. (2005). Chemistry of the mind: Neurochemical modulation of prefrontal cortical function. *Journal of Comparative Neurology, 493,* 140–146.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Sajikumar, S., & Frey, J. U. (2004). Late-associativity, synaptic tagging, and the role of dopamine during LTP and LTD. *Neurobiology of Learning and Memory, 82,* 12–25.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005, November 25). Representation of action-specific reward values in the striatum. *Science, 310,* 1337–1340.

Schmitzer-Torbert, N. C., & Redish, A. D. (2004). Neuronal activity in the rodent dorsal striatum in sequential navigation: Separation of spatial and reward responses on the multiple-T task. *Journal of Neurophysiology, 91,* 2259–2272.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron, 36,* 241–263.

Seamans, J. K., Floresco, S. B., & Phillips, A. G. (1998). D1 receptor modulation of hippocampal-prefrontal cortical circuits integrating spatial memory with executive functions in the rat. *Journal of Neuroscience, 18,* 1613–1621.

Seamans, J. K., Gorelova, N., Durstewitz, D., & Yang, C. R. (2001). Bidirectional dopamine modulation of GABAergic inhibition in prefrontal cortical pyramidal neurons. *Journal of Neuroscience, 21,* 3628–3638.

Seamans, J. K., & Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology, 74,* 1–57.

Seligman, M. E. (1972). Learned helplessness. *Annual Review of Medicine, 23,* 407–412.

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990, August 24). A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science, 249,* 892–895.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., et al. (2004, June 10). Temporal difference models describe higher-order learning in humans. *Nature, 429,* 664–667.

Shaham, Y., Shalev, U., Lu, L., de Wit, H., & Stewart, J. (2003). The reinstatement model of drug relapse: History, methodology and major findings. *Psychopharmacology, 168,* 3–20.

Shalev, U., Highfield, D., Yap, J., & Shaham, Y. (2000). Stress and relapse to drug seeking in rats: Studies on the generality of the effect. *Psychopharmacology, 150,* 337–346.

Sharp, P. E., Blair, H. T., Etkin, D., & Tzanetos, D. B. (1995). Influences of vestibular and visual motion information on the spatial firing patterns of hippocampal place cells. *Journal of Neuroscience, 15,* 173–189.

Sinha, R., Catapano, D., & O'Malley, S. (1999). Stress-induced craving and stress response in cocaine dependent individuals. *Psychopharmacology, 142,* 343–351.

Smith, C. (1995). Sleep states and memory processes. *Behavioural Brain Research, 69,* 137–145.

Sotres-Bayon, F., Cain, C. K., & LeDoux, J. E. (2006). Brain mechanisms of fear extinction: Historical perspectives on the contribution of prefrontal cortex. *Biological Psychiatry, 60,* 329–336.

Squire, L. R. (1987). *Memory and brain.* New York: Oxford University Press.

Stefani, M. R., & Moghaddam, B. (2006). Rule learning and reward contingency are associated with dissociable patterns of dopamine activation in the rat prefrontal cortex, nucleus accumbens, and dorsal striatum. *Journal of Neuroscience, 26,* 8810–8818.

Stuber, G. D., Wightman, R. M., & Carelli, R. M. (2005). Extinction of cocaine self-administration reveals functionally and temporally distinct dopaminergic signals in the nucleus accumbens. *Neuron, 46,* 661–669.

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 model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience, 91,* 871–890.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88,* 135–170.

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

Suzuki, A., Josselyn, S. A., Frankland, P. W., Masushige, S., Silva, A. J., & Kida, S. (2004). Memory reconsolidation and extinction have distinct temporal and biochemical signatures. *Journal of Neuroscience, 24,* 4787–4795.

Szepesvári, C., & Littman, M. L. (1999). Unified analysis of value-function-based reinforcement learning algorithms. *Neural Computation, 8,* 2017–2060.

Tanaka, S. (2006). Dopaminergic control of working memory and its relevance to schizophrenia: A circuit dynamics perspective. *Neuroscience, 139,* 153–171.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience, 7,* 887–893.

Tarter, R. E., Ammerman, R. T., & Ott, P. J. (Eds.). (1998). *Handbook of substance abuse: Neurobehavioral pharmacology.* New York: Plenum.

Teng, E., & Squire, L. R. (1999, August 12). Memory for places learned long ago is intact after hippocampal damage. *Nature, 400,* 675–677.

Theios, J. (1962). The partial reinforcement effect sustained through blocks of continued reinforcement. *Journal of Experimental Psychology, 64,* 1–6.

Touretzky, D. S., & Redish, A. D. (1996). A theory of rodent navigation based on interacting representations of space. *Hippocampus, 6,* 247–270.

Tremblay, L., Hollerman, J. R., & Schultz, W. (1998). Modifications of reward expectation-related neuronal activity during learning in primate striatum. *Journal of Neurophysiology, 80,* 964–977.

Ungless, M. A. (2004). Dopamine: The salient issue. *Trends in Neurosciences, 27,* 702–706.

Ungless, M. A., Magill, P. J., & Bolam, J. P. (2004, March 26). Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science, 303,* 2040–2042.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999, January 22). The role of locus coeruleus in the regulation of cognitive performance. *Science, 283,* 549–554.

Vianna, M. R. M., Szapiro, G., McGaugh, J. L., Medina, J. H., & Izquierdo, I. (2001). Retrieval of memory for fear-motivated training initiates extinction requiring protein synthesis in the rat hippocampus. *Proceedings of the National Academy of Sciences, USA, 98,* 12251–12254.

Waelti, P., Dickinson, A., & Schultz, W. (2001, July 5). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412,* 43–48.

Wagenaar, W. A. (1988). *Paradoxes of gambling behavior.* Hillsdale, NJ: Erlbaum.

Wang, S., Marin, M., & Nader, K. (2005). *Memory strength as a transient boundary condition on reconsolidation of auditory fear memories and its molecular correlates*. Society for Neuroscience Annual Meeting. Program No. 650.2.

Weinberger, N. M. (1998). Physiological memory in primary auditory cortex: Characteristics and mechanisms. *Neurobiology of Learning and Memory, 70,* 226–251.

White, N. M., & Hiroi, N. (1998). Preferential localization of self-stimulation sites in striosomes/patches in the rat striatum. *Proceedings of the National Academy of Sciences, USA, 95,* 6486–6491.

Wills, T. J., Lever, C., Cacucci, F., Burgess, N., & O'Keefe, J. (2005, May 6). Attractor dynamics in the hippocampal representation of the local environment. *Science, 308,* 873–876.

Willshaw, D. J., Bruneman, O. P., & Longuet-Higgins, H. C. (1969, June 7). Non-holographic associative memory. *Nature, 222,* 960–962.

Wilson, H. R., & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic tissue. *Kybernetik, 13,* 55–80.

Wise, R. A. (2004). Dopamine, learning, and motivation. *Nature Reviews Neuroscience, 5,* 1–12.

Young, A. M., Moran, P. M., & Joseph, M. H. (2005). The role of dopamine in conditioning and latent inhibition: What, when, where, and how? *Neuroscience and Biobehavioral Reviews, 29,* 963–976.

Yu, A. J., & Dayan, P. (2002). Acetylcholine in cortical inference. *Neural Networks, 15,* 719–730.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46,* 681–692.

Zahrt, J., Taylor, J. R., Mathew, R. G., & Arnsten, A. F. T. (1997). Supranormal stimulation of D1 dopamine receptors in the rodent prefrontal cortex impairs spatial working memory performance. *Journal of Neuroscience, 17,* 8528–8535.

Received January 27, 2006
Revision received February 2, 2007
Accepted February 2, 2007 ■