Behavioral Neuroscience
2007, Vol. 121, No. 1, 31–49

0735-7044/07/$12.00 DOI: 10.1037/0735-7044.121.1.31

PVLV: The Primary Value and Learned Value Pavlovian Learning Algorithm

Randall C. O'Reilly
University of Colorado at Boulder

Michael J. Frank
University of Arizona

Thomas E. Hazy and Brandon Watz
University of Colorado at Boulder

The authors present their primary value learned value (PVLV) model for understanding the reward-predictive firing properties of dopamine (DA) neurons as an alternative to the temporal-differences (TD) algorithm. PVLV is more directly related to underlying biology and is also more robust to variability in the environment. The primary value (PV) system controls performance and learning during primary rewards, whereas the learned value (LV) system learns about conditioned stimuli. The PV system is essentially the Rescorla–Wagner/delta-rule and comprises the neurons in the ventral striatum/nucleus accumbens that inhibit DA cells. The LV system comprises the neurons in the central nucleus of the amygdala that excite DA cells. The authors show that the PVLV model can account for critical aspects of the DA firing data, making a number of clear predictions about lesion effects, several of which are consistent with existing data. For example, first- and second-order conditioning can be anatomically dissociated, which is consistent with PVLV and not TD. Overall, the model provides a biologically plausible framework for understanding the neural basis of reward learning.

Keywords: basal ganglia, dopamine, reinforcement learning, Pavlovian conditioning, computational modeling

An important and longstanding challenge for both the cognitive neuroscience and artificial intelligence communities has been to develop an adequate understanding (and a correspondingly robust model) of Pavlovian learning. Such a model should account for the full range of signature findings in the rich literature on this phenomenon. *Pavlovian conditioning* refers to the ability of previously neutral stimuli that reliably co-occur with primary rewards to elicit new conditioned behaviors and to take on reward value themselves (e.g., Pavlov's famous case of the bell signaling food for hungry dogs; Pavlov, 1927).

Pavlovian conditioning is distinguished from *instrumental conditioning* in that the latter involves the learning of new behaviors that are reliably associated with reward, either first order (US), or second order (CS). Although Pavlovian conditioning also involves behaviors (conditioned and unconditioned responses), reward delivery is not contingent on behavior but is instead reliably paired with a stimulus regardless of behavior. In contrast, instrumental conditioning explicitly makes reward contingent on a particular "operant" or "instrumental" response. Both stimulus–reward (Pavlovian) and stimulus–response–reward (instrumental) associations, however, are thought to be trained by the same phasic dopamine signal that occurs at the time of primary reward (US) as described below. In practice, the distinction is often blurry as the two types of conditioning interact (e.g., second-order instrumental conditioning and so-called Pavlovian instrumental transfer effects).

The dominant theoretical perspective for both Pavlovian and instrumental conditioning since the seminal Rescorla and Wagner (1972) model, is that learning is based on the discrepancy between actual rewards received and predictions thereof (i.e., reward prediction error). Currently, the temporal differences (TD) reward prediction framework (Sutton, 1988; Sutton & Barto, 1998) is by far the most widely adopted computational level account of Pavlovian (and instrumental) conditioning and dopamine firing (e.g., Barto, 1995; Daw, Courville, & Touretzky, 2003; Daw, Kakade, & Dayan, 2002; Dayan, 2001; Dayan & Balleine, 2002; Houk, Adams, & Barto, 1995; Kakade & Dayan, 2002a, 2002b; Montague, Dayan, & Sejnowski, 1996; Suri & Schultz, 1999, 2001; see Brown, Bullock, & Grossberg, 1999; Contreras-Vidal & Schultz, 1999; Sporns & Alexander, 2002, for alternative models, and Joel, Niv, & Ruppin, 2002, for a biologically oriented review).

One important reason for the popularity of TD is that a reward prediction error signal has been established in the brain, in the pattern of midbrain dopamine neuron activation (e.g., Schultz, 1998; Schultz, Apicella, & Ljungberg, 1993; see Figure 1). These neurons initially fire short phasic bursts of activity for primary rewards and over the course of learning come to fire similarly at

Randall C. O'Reilly, Thomas E. Hazy, and Brandon Watz, Department of Psychology, University of Colorado at Boulder; Michael J. Frank, Department of Psychology and Program in Neuroscience, University of Arizona.

This work was supported by Office of Naval Research Grant N00014-03-1-0428 and National Institute of Mental Health Grants MH069597 and MH64445. We thank Peter Dayan, Nathaniel Daw, Yael Niv, Eric Claus, and the CCN lab for discussions of these ideas.

Correspondence concerning this article should be addressed to Randall C. O'Reilly, Department of Psychology, University of Colorado at Boulder, 345 UCB, Boulder, CO 80309. E-mail: oreilly@psych.colorado.edu

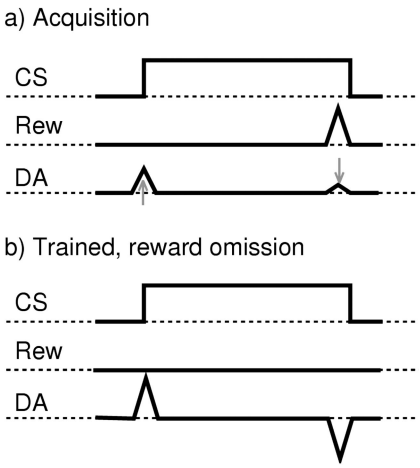

Figure 1. Schematic of dopamine (DA) recording data for a simple conditioning experiment in which a conditioned stimulus (CS) reliably precedes the delivery of a reward (Rew). During acquisition (a), DA initially bursts at the time of reward delivery but then starts spiking at stimulus onset, diminishing at the time of reward delivery. Note that there is no strong evidence of a backward-propagating burst over training, as predicted by some versions of the temporal-differences model but not by primary value learned value (PVLV). After training (b), if reward is omitted, a dip in DA below baseline tonic levels is observed.

the onset of previously neutral, reward predictive stimuli (i.e., conditioned stimuli; CS), and no longer to the reward itself. Generally, there is a time period when both CS and reward-related firing is occurring (Pan, Schmidt, Wickens, & Hyland, 2005; Schultz, 2002).

However, it remains unclear exactly what brain mechanisms lead to this behavior on the part of dopamine cells. Most researchers agree that the critical learning processes are taking place upstream from the midbrain dopamine neurons themselves. But which areas are doing what? Because it is an abstract, unitary (and elegant) framework, the TD model does not map directly onto the relatively large collection of neural substrates known to be involved in reinforcement learning, including areas of the basal ganglia, amygdala, midbrain dopamine nuclei, and ventromedial prefrontal cortex (Cardinal, Parkinson, Hall, & Everitt, 2002). Indeed, relatively few specific proposals have been made for a biological mapping of the TD model (Houk et al., 1995; Joel et al., 2002).

In this article, we offer a multicomponent model of Pavlovian learning called *PVLV*, which provides a more direct mapping onto the underlying neural substrates. PVLV is composed of two subsystems: *primary value* (PV) and *learned value* (LV). The PV system is engaged by primary reward (i.e., an unconditioned stimulus; US) and learns to expect the occurrence of a given US, thereby inhibiting the dopamine burst that would otherwise occur for it. The LV system learns about conditioned stimuli that are reliably associated with primary rewards, and it drives phasic dopamine burst firing at the time of CS onset. This decomposition is similar to the model of Brown et al. (1999), but as we discuss later, there are several important functional and anatomical differences between the two models.

The PV and LV systems are further subdivided into excitatory and inhibitory subcomponents, which provide a good fit with a wide range of data (reviewed in detail later) on three different brain areas. The excitatory subcomponent of PV (denoted PVe) is associated with the reward-driven excitatory projections from the lateral hypothalamus onto midbrain dopamine neurons in the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) as we discuss in more detail later in the section "Biological Mapping of PVLV." The inhibitory subcomponent of PV (PVi) is associated with neurons in the ventral striatum/nucleus accumbens (VS/NAc) that have direct GABAergic projections to the SNc and VTA and fire just in advance of primary rewards. The excitatory subcomponent of the LV system (LVe) is associated with neurons in the central nucleus of the amygdala (CNA), which have a net excitatory effect on the SNc and VTA. Thus, we suggest that these CNA neurons learn to associate CSs with reward and drive excitatory dopamine bursts at CS onset. Finally, there is an inhibitory component of the LV (LVi) that is also associated with the VS/NAc, which slowly learns to inhibit the excitatory LVe drive on the dopamine neurons.

In addition to these core PVLV mechanisms, a number of other brain areas play a critical role in reinforcement learning. For example, we think of the prefrontal cortex (PFC) and hippocampus as providing something akin to an eligibility trace (as in TD[λ]; Sutton & Barto, 1998; Pan et al., 2005). We believe this sort of actively maintained working memory representation is particularly crucial in trace conditioning paradigms in which there is an interval of time between CS-offset and US-onset. As we discuss later, PVLV explicitly accounts for known dissociations between delay versus trace conditioning paradigms that occur under PFC and/or hippocampal lesions, something about which TD is less explicit. In fact, PVLV actually requires that models learn to hold onto working memory representations under trace conditioning paradigms. Although for the models in this article we apply the working memory inputs directly (so as to focus on the core PVLV mechanisms), our larger prefrontal cortex-basal ganglia (PBWM) model, of which PVLV is a subcomponent, demonstrates how the system can learn to maintain task-relevant information in working memory (O'Reilly & Frank, 2006). TD does not address the learning of working memory representations explicitly, instead it finesses the issue by assuming that it is just there in the eligibility trace.

In addition, the cerebellum (and possibly other brain areas) provides a representation of time (e.g., Mauk & Buonomano, 2004; Ivry, 1996) that acts as an additional input signal that can become associated with reward, as in the framework of Savastano and Miller (1998). The basolateral nucleus of the amygdala (BLA) is important for second-order conditioning in this framework because detailed properties of the PVLV mechanism prevent the LVe (CNA) from performing both first- and second-order conditioning. This is consistent with data showing anatomical dissociations between these forms of conditioning (e.g., Hatfield, Han, Conely, & Holland, 1996). Note that this dissociation, and many others reviewed later, would not be predicted by the abstract, unitary TD mechanism. Thus, the PVLV mechanism provides an important bridge between the more abstract TD model and the details of the underlying neural systems.

The mapping between TD and PVLV is not perfect, however, and PVLV makes several distinctive predictions relative to TD in various behavioral paradigms. For example, PVLV strongly predicts that higher order conditioning beyond second-order should be

weak to nonexistent, whereas TD makes no distinction between these different levels of conditioning. There is a remarkable lack of published work on third or higher levels of conditioning, and the two references we were able to find indicate that it is nonexistent or weak at best (Denny & Ratner, 1970; Dinsmoor, 2001). Another difference comes from paradigms with variable CS–US intervals. As we show later, PVLV is very robust to this variability but TD is not. The data indicate that animals are also robust to this form of variability (H. Davis, McIntire, & Cohen, 1969; Kamin, 1960; Kirkpatrick & Church, 2000). PVLV also makes a very strong distinction between delay and trace conditioning, as do animals, whereas this distinction in TD is considerably more arbitrary.

The remainder of the article is organized as follows. First we develop the PVLV algorithm at a computational level and provide several demonstrations of basic Pavlovian learning phenomena by using the PVLV model. Next, we discuss the mapping of PVLV onto the brain areas as summarized above and review a range of empirical data that are consistent with this model. We conclude by comparing our model with other models in the literature, including the Brown et al. (1999) model, which has several important similarities and several differences relative to our model.

The PVLV Algorithm

The PVLV algorithm starts with the basic Rescorla and Wagner (1972) learning rule (which is formally identical to the earlier delta rule; Widrow & Hoff, 1960, originally pointed out by Sutton & Barto, 1981), which captures the core principle that learning should be based on the discrepancy between predictions and actual outcomes:

$$\delta^t = r^t - \hat{r}^t, \tag{1}$$

where r^t is the current reward value at time t, $\hat{r}^t$ is the expected or predicted reward value, and δ^t is the discrepancy or error between the two. This δ^t value then drives synaptic weight changes for the system computing $\hat{r}^t$. For example, a simple neural model would involve a single neural unit that computes the estimated value $\hat{r}^t$ by using synaptic weights w_i from a set of sensory inputs x_i:

$$\hat{r}^t = \sum_i w_i^t x_i^t. \tag{2}$$

The change in the weight values needed to improve the estimated reward value is simply

$$\Delta w_i^t = \varepsilon \delta^t x_i^t. \tag{3}$$

This model does an excellent job of learning to expect primary rewards, and, if we take the δ^t to represent the dopamine firing deviations from baseline, it can explain the cancellation of dopamine bursting at the onset of the US in a classic Pavlovian paradigm (Figure 1). However, it cannot account for the firing of dopamine bursts at the earlier onset of a CS because in fact there is no actual primary reward (r^t) present at that time, and thus the system will not learn to expect anything at that time. This CS-triggered dopamine firing plays a critical functional role in learning because it allows the system to learn which situations and actions can lead to subsequent reward. For example, initial exposure to the presence of cookies in a cookie jar can enable a subsequent dopamine-reinforced approach and opening of the jar.

The TD algorithm corrects this critical limitation of the Rescorla–Wagner algorithm by adopting a temporally extended prediction framework, where the objective is to predict future rewards not just present rewards. The consequence of this is that the δ^t at one point in time drives learning based on the immediately prior sensory input state x_i^{t-1}. This produces a chain-reaction effect in which a reward prediction error at one point in time propagates earlier and earlier in time, to the earliest reliable predictor of a subsequent reward. Hence, the δ^t value, and thus the dopamine bursting, can move earlier in time to the onset of the CS.

The PVLV algorithm takes a different approach: The basic Rescorla–Wagner learning rule is retained as the PV (primary value) system, and an additional system (LV, learned value) is added to learn about reward associations for conditioned stimuli. In addition to the biological motivations for such a division of labor mentioned earlier (and elaborated below), there are some computational advantages for adopting this approach. Principally, the relationship between a CS and a subsequent US is not always very reliable, and having separate PV and LV systems enables the system to be very robust to such variability. In contrast, the chaining mechanism present in the TD algorithm is designed to work optimally when there is a very reliable sequence of events leading from the CS to the US. Intuitively, the chain between CS and US must remain unbroken for the predictive signal to propagate backward over learning, and this chain is only as strong as its weakest link. This problem can be mitigated to some extent by using an eligibility trace as in TD(λ), where $0 < \lambda < 1$ parameterizes an exponentially decaying trace of the input stimuli used for learning. This can smooth over rough spots in the chain but at the potential cost of reducing the temporal precision of reward predictions as a result of excessive smearing. In contrast, PVLV avoids this problem entirely by not relying on a chaining mechanism at all.

There are many situations in which the CS–US relationship is unreliable. For example, in many working memory tasks, a highly variable number of distractor stimuli can intervene between a stimulus to be encoded in working memory and the subsequent demand to recall that stimulus (Hochreiter & Schmidhuber, 1997; O'Reilly & Frank, 2006). Any dog owner knows that dogs come to associate the jingling of a leash with the idea that they will soon be going on a walk, despite a variable amount of time and intervening events between the leash jingle and the walk itself (e.g., the owner may go to the bathroom, turn off the television, and check e-mail). In the animal learning literature, there are (only) a few experiments in which the CS–US relationship is variable (H. Davis et al., 1969; Kamin, 1960; Kirkpatrick & Church, 2000), but it is clear that conditioning is very robust in this case, equivalent to comparison conditions that have fixed CS–US intervals. This finding is consistent with PVLV and poses a challenge to TD-based approaches.

In short, we think the PVLV mechanism has the simplicity and robustness that are often characteristic of biological systems, with the cost of being less elegant than the TD system (two systems are

required instead of one). In the subsequent sections, we provide the details for how the PV and LV systems operate.

The PV System

We can rewrite the Rescorla–Wagner equation in terms of the excitatory (PVe) and inhibitory (PVi) subcomponents of the PV system. The excitatory PV system represents the value implicitly hardwired into a primary reward (US), $PV_e^t = r^t$ in the notation of Rescorla–Wagner, whereas the inhibitory system learns to cancel out these rewards, $PV_i^t = \hat{r}^t$. Thus, in this terminology, the PV delta is

$$\delta_{pv}^t = PV_e^t - PV_i^t = r^t - \hat{r}^t, \tag{4}$$

and this value is used to train the PVi system as described earlier (Equation 3). As a consequence, when primary rewards are delivered, the PVi system associates the current state of the system with the US (reward). This current state information includes any sensory inputs that coincide with reward, together with internally generated timing signals (e.g., if rewards are always delivered precisely 2 s following an input stimulus, then the 2-s timing signal becomes associated with the US just as an external sensory stimulus can become associated with it; Savastano & Miller, 1998). As these associations increase, PV_i^t at the time of primary reward increases to match PV_e^t, and the δ_{pv}^t value (i.e., dopamine bursting) decreases, which is the observed pattern.

The LV System

The LV system also uses the Rescorla–Wagner learning rule but has a few key differences that enable it to signal reward associations at the time of CS onset. Like the PV system, the LV system has two components, excitatory (LVe) and inhibitory (LVi). We focus first on the LVe component, which learns CS associations and drives the excitatory dopamine bursts at CS onset. The most important property of the LVe system is that it only learns when primary rewards are present or expected. In contrast, the PVi system learns at all times about the current primary reward status (PVe or r^t). This difference protects the LVe system from having to learn that there are no actual primary rewards present at the time of CS onset. Therefore, unlike the PV system, it is able to signal the reward association of a CS and not have this signal (dopamine burst) trained away, as otherwise it would be if pure Rescorla–Wagner learning were at work.

More formally, the LVe learning is conditioned on the state of the PV system, according to the following filtering condition:

$$PV_{filter} = PV_i^t > \theta_{pv} \text{ or } PV_e^t > \theta_{pv}, \tag{5}$$

where θ_{pv} is a threshold on PV activation, above which it is considered that the PV system is expecting or receiving a reward at this time (in the Appendix we present a more general condition that allows for representation of both reward and punishment expectations).

For clarity, note that PV_{filter} is thus a boolean variable such that

$$PV_{filter} = \begin{cases} 1 & \text{if primary reward present or expected} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The boolean value of PV_{filter} then regulates the learning of the LVe system,

$$\Delta w_i^t = \begin{cases} \varepsilon(PV_e^t - LV_e^t)x_i^t, & \text{if } PV_{filter} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The dependence of the secondary LV system on the primary PV system for learning ensures that actual reward outcomes have the final say in shaping all of the reward associations learned by the system. Also, note that it is the primary reward value itself (PV_e^t or r^t) that drives the learning of the LV system, not the PV or LV delta value, which is defined next. These features have important implications for understanding various conditioning phenomena, as elaborated below.

The LVi system performs a similar role for a learned CS as the PVi system does for the US: It learns to cancel out dopamine bursts for a highly learned CS. The LVi system is essentially the same as the LVe system, except that it uses a slower learning rate (ε), and it produces a net inhibitory drive on the dopamine system like the PVi system. The LV delta is then the difference between the excitatory and inhibitory components (just as with the PV delta),

$$\delta_{lv}^t = LV_e^t - LV_i^t. \tag{8}$$

Because of its slower learning rate, LVi slowly learns which CSs are reliably associated with reward and decreases the dopamine bursts for such CSs relative to those that have more recently become associated with reward (which have been learned by the faster LVe but not the slower LVi system). Furthermore, if a CS that has been reliably associated with reward subsequently becomes less strongly associated with reward, the LV delta can become negative (because LVe has learned this new lower reward association, but LVi retains the previous more positive association), indicating the change in reward association. Thus, consistent with the computational motivation for the delta rule, the LV delta in Equation 8 represents the discrepancy between what was previously known or expected (as encoded in the LVi weights of the system through prior learning) and what is more recently happening (encoded through the LVe weights). This LVi value does not much affect the simple conditioning simulations shown below, but it is more important for the efficacy of PVLV in training an actor (in our case for working memory updating; O'Reilly & Frank, 2006). Specifically, without LVi a stimulus associated with reward would always drive a DA burst (even if its reward association had recently decreased), and it would always reinforce actions with a constant dopamine burst, to the point that such actions would be massively overlearned.

How do the PV and LV systems each contribute to the dopamine output signal? Because there are two delta signals in PVLV, from PV and LV, these need to be combined to produce an overall delta value that can be used as a global dopamine signal (e.g., to train an actor system in an actor–critic architecture). The most functionally transparent mechanism is to

have the PV delta apply whenever there is a primary reward present or expected by the PV system. But when no rewards are present, the LV delta can still drive dopamine firing. As before (see Equation 5), PVLV implements this by using the boolean variable, PV_{filter}, where

$$PV_{filter} = \begin{cases} 1 & \text{if primary reward present or expected} \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

and PV_{filter} is evaluated,

$$PV_{filter} = PV_i^t > \theta_{pv} \text{ or } PV_e^t > \theta_{py}. \quad (10)$$

Thus,

$$\delta^t = \begin{cases} \delta_{pv}, & \text{if } PV_{filter} \\ \delta_{lv} & \text{otherwise.} \end{cases} \quad (11)$$

This is also consistent with the equation that determines when the LV system learns according to PV expectations and actual reward delivery (Equation 7).[1]

Figure 2 summarizes the PVLV system's operation in the simple CS–US conditioning paradigm we have been considering. The PV continuously learns about the occurrence of primary rewards (both presence and absence), and as it learns to expect reward delivery it cancels the dopamine burst (i.e., PV delta value) that would otherwise occur at that time. The reward also trains the LV system, which produces increasing weights from the CS (as long as this is active in the input at this time). On subsequent trials, the LV system is then able to fire naturally at CS onset, producing a dopamine burst (i.e., LV delta value). By this mechanism, the time gap between CS-onset and US is bridged automatically by the CS–US association, without recourse to the kind of explicit prediction that is central to the TD model. The biological mapping of the PVLV mechanisms shown in the figure are discussed in detail below.

Additional Mechanisms

There are two additional mechanisms required for the overall system to function (and to be consistent with available data). First (as previously noted), the PVi system must take advantage of some kind of timing signal that enables it to fire at the expected time of actual reward input and not otherwise. In Figure 2B, we illustrate a ramping timing signal triggered by CS onset, which is intended to represent the kind of interval timing signal provided by the cerebellum (e.g., Ivry, 1996; Mauk & Buonomano, 2004), but any kind of regular activity pattern would work just as well for our model (see Lustig, Matell, & Meck, 2005, for a model of timing signals within the basal ganglia). We discuss this issue further in comparison with an alternative model of DA firing by Brown et al. (1999) below, which depends on an intrinsic timing mechanism as an integral part of their system.

The second additional mechanism required is a novelty detection (and familiarity suppression) mechanism, so that the LV system does not continue to trigger dopamine spiking during the entire duration of CS input. With such a mechanism in place, the first onset of a stimulus input triggers a burst of LV firing, but this then decreases as the stimulus stays on. One solution to this problem is to use a habituation mechanism on the LV system to achieve this effect (e.g., Brown et al., 1999), but this would generalize across various different stimuli and would therefore prevent a second stimulus that could be associated with a different or larger reward from evoking DA firing. Instead, in our implementation we have adopted a synaptic depression mechanism (e.g., Abbott, Varela, Sen, & Nelson, 1997; Markram & Tsodyks, 1996; Zucker & Regehr, 2002; Huber & O'Reilly, 2003), which causes a habituation of the LV DA-burst firing response only to the stimulus that was initially active (i.e., only active synapses are depressed). With this mechanism in place, the LVe system accommodates to any constant sensory inputs and responds only to changes in input signals, causing it to fire only at the onset of a stimulus tone. Such synaptic depression mechanisms are ubiquitous throughout the vertebrate and invertebrate brain (Zucker & Regehr, 2002). Nevertheless, there are a large number of ways of implementing such an overall function, so we are confident that, if our overall hypothesis about the PVLV mechanism is correct, the brain will have found a means of achieving this function.[2] For full details about the PVLV algorithm and implementation, see the Appendix.

Application to Conditioning Data

At the level of the basic DA firing data represented in Figure 1, both TD and PVLV account for the most basic findings of DA bursting at tone onset and cancellation of the burst at reward delivery. However, as noted earlier, simple TD models (but not PVLV) also predict a chaining of DA bursts "backward in time" from the reward to the stimulus onset, which has not been reliably observed empirically (Fiorillo, Tobler, & Schultz, 2005; Pan et al., 2005). However, this particular aspect of the data is still controversial (e.g., Niv, Duff, & Dayan, 2005) and also depends critically on the way that the input environment is represented. For example, Pan et al. (2005) recently showed how a TD(λ) model with a high lambda value could reproduce the empirically observed pattern (i.e., no evidence of backward marching dopamine bursts). Furthermore, the data often show dopamine bursts at both the CS and US (Pan et al., 2005; Schultz, 2002)—this is incompatible with

[1] A simpler possible implementation would be to just add the two delta values to produce a summed DA value, but this double counts the reward-related deltas because both the LV and PV contribute in this case. Nevertheless, because LV and PV deltas otherwise occur at different times, Equation 11 is very similar to adding the deltas; the PV system just dominates when external rewards are presented or expected. It is also possible to consider an additive equation that also conditionalizes the contribution of the PV component; this was found in O'Reilly and Frank (2006) to work slightly better than Equation 11 in a working memory model (see Appendix for details).

[2] Available evidence suggests that a mechanism such as proposed here most likely exists in the pathway somewhere distal to the LVe representations themselves (which PVLV proposes to be in the central nucleus of the amygdala, see below) as electrophysiological recording data show sustained (i.e., not onset-only) firing in CNA cells throughout CS duration (Ono, Nishijo, & Uwano, 1995). For example, downstream synaptic depression/habituation may occur in either the pedunculopontine nucleus, or it could be intrinsic to local dynamics in the midbrain dopamine nuclei themselves.

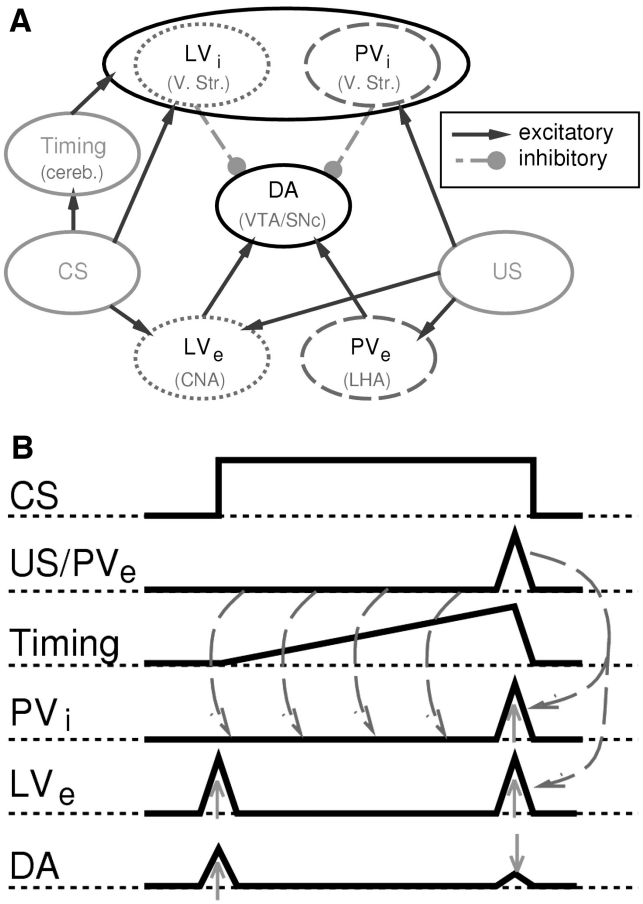

Figure 2. The primary value learned value (PVLV) learning mechanism. A: The structure of PVLV. The primary value (PV) system learns about primary rewards and contains two subsystems: The excitatory (PVe) drives excitatory dopamine (DA) bursts from primary rewards (US = unconditioned stimulus), and the inhibitory (PVi) learns to cancel these bursts (by using timing or other reliable signals). The learned value (LV) system learns to fire for conditioned stimuli (CS) that are reliably associated with reward. The excitatory component (LVe) drives DA bursting, whereas the inhibitory component (LVi) is just like the PVi, except it inhibits CS-associated bursts. The PVe corresponds anatomically to the lateral hypothalamus (LHA), which has excitatory projections to the midbrain DA nuclei, and responds to primary rewards. The PVi and LVi correspond to the striosome/patch neurons in the ventral striatum (V. Str.), which have direct inhibitory projections onto the DA system and learn to fire at the time of expected rewards. The LVe corresponds to the central nucleus of the amygdala (CNA), which has excitatory DA projections and learns to respond to CSs. B: Application to the simple conditioning paradigm, in which the PVi learns (on the basis of the PVe reward value at each time step) to cancel the DA burst at the time of reward, whereas the LVe learns a positive CS association (only at the time of reward) and drives DA bursts at CS onset. The phasic nature of CS firing, despite a sustained CS input, requires a novelty detection mechanism of some form; we suggest a synaptic depression mechanism has beneficial computational properties. Timing is thought to represent distributed drifting pattern of activity as computed by several cerebellum models, represented here in one dimension as a simple ramp. VTA = ventral tegmental area; SNc = substantia nigra pars compacta; cereb. = cerebellum.

TD(0) but is consistent with both PVLV and TD(λ). Therefore, although it would be nice support for TD if chaining were reliably observed, its apparent absence is not particularly strong evidence against the overall TD framework.

In the following sections, we show that PVLV can account for other basic features of Pavlovian conditioning, including extinction, blocking, overshadowing/summation, conditioned inhibition, trace versus delay conditioning, and second-order conditioning. We further demonstrate that PVLV can robustly learn to associate stimuli with reward values even when the delay and number of intervening nonrewarding stimuli are randomized—a situation that challenges TD, which depends on a predictable delay and sequence of events between the CS and the US.

Model Implementation

Figure 3A shows the PVLV model implementation used for these simulations. Stimulus inputs are simple localist units, with an additional set of stimulus-specific timing inputs that are thought to represent signals generated by the cerebellum (e.g., Ivry, 1996; Mauk & Buonomano, 2004). Each stimulus triggers a sequence of unit activations that simply progress across the input from left to right at a rate of one unit per time step. These signals enable the PVi layer to learn when to expect the external reward, thereby canceling the DA burst at that time. The estimated value representations for the PV and LV systems are encoded as a distributed activity pattern across three value units with preferred activations of (0, .50, 1), so that different values can be encoded by using different weights.[3]

The PVe layer is clamped to provide a representation of the external reward value, r^t, as an input. We use .50 to represent no reward information, 1 for rewards, and 0 for negative feedback/punishment. The remaining PVi, LVe, and LVi layers learn weights from the sensory/timing inputs to produce the corresponding reward expectation values. The LVi layer learns too slowly to affect any of the results presented here, but as noted earlier it is critical for more complex reinforcement learning conditions that require integrating feedback across multiple trials and stimulus combinations (as in the O'Reilly & Frank, 2006, working memory model), and it is included for completeness. The units update their activations by using a point neuron model with simulated excitatory, inhibitory, and leak currents, as defined in the Leabra model (see Appendix for equations).

Acquisition and Extinction

Figure 3C shows the pattern of activation from the model after acquisition for the basic Pavlovian conditioning paradigm we have considered throughout the article. As expected, the model shows the observed pattern of DA firing, initially at the time of reward and then increasingly at the time of CS onset and not at the time of reward. In addition, this pattern extinguishes if reward is no longer provided, a straightforward result in the model due to the

[3] A single unit value representation would be constrained by virtue of having one set of weights to have a monotonic mapping onto scalar values. This is not critical for the simple demonstrations here, but it is more generally useful (e.g., for our working memory model; O'Reilly & Frank, 2006).

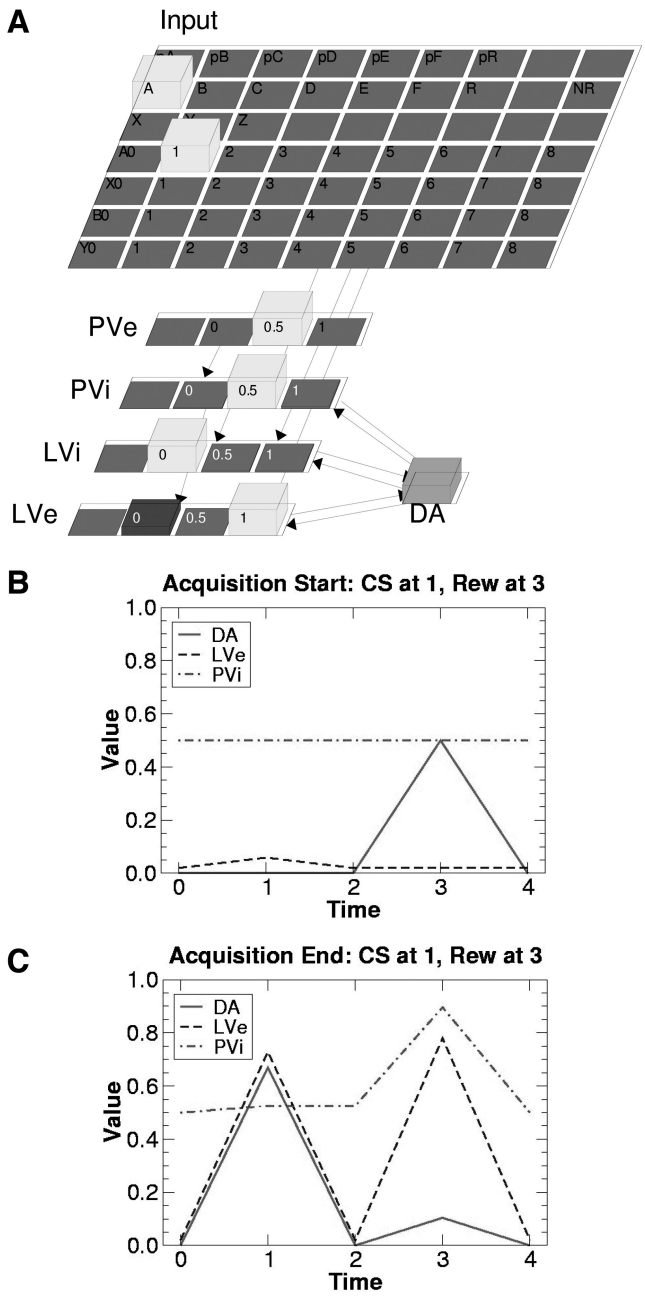

Figure 3. Primary value learned value (PVLV) network implementation (A) and basic patterns of activation (B and C) for a standard delay conditioning trial in which a conditioned stimulus (CS) comes on at Time 1 and stays on until Time 3, at which point a reward (Rew) is delivered. The overall dopamine output (DA) firing shifts from the point of reward at the start of training (B) to the onset of the CS at the end of training (C). The model uses localist representations of value (0, .50, 1) so different values can be encoded with different sets of weights; overall value is activation-weighted average across units, shown in the first unit in layer. LVe = learned value excitatory layer; LVi = learned value inhibitory layer; PVi = primary value inhibitory; PVe = primary value excitatory (external reward).

PVi mechanism (results not shown). However, it is important that the LVe have a slightly faster learning rate than PVi for this extinction to occur—otherwise the PVi can extinguish before the LVe, causing PV_{filter} to go below threshold and leaving LVe with some residual firing.

Blocking

The blocking effect (Kamin, 1968) was one of the main motivations leading to the development of the Rescorla–Wagner learning rule. In these studies, an animal first learned to associate a single stimulus (A) with reward (A + trial). Subsequently, a second stimulus (X) was paired with A, and the compound stimulus was rewarded (AX + trial). *Blocking* refers to the phenomenon that prior learning of a reward association for A blocks the acquisition of association for X. By virtue of having the Rescorla–Wagner rule drive the engine, PVLV naturally produces this effect. We first trained the model for 50 "epochs" (one pass through the training patterns) on A+ and B− (+ = positive reward association, − = null reward association). For the next 50 epochs, we added AX+ and BY+ trials while continuing to train A+ and B−. Finally, we tested X and Y, measuring the CS onset DA value. As shown in Figure 4A, the acquisition of reward association to X was blocked by the prior A+ association relative to Y. This is because PVi has learned to inhibit the DA burst that would otherwise have occurred for rewards based on the presence of the A stimulus. As a result, there is minimal DA-driven learning about X. This same pattern of DA firing (and the lack thereof) replicates physiological recordings seen in experiments that have used the same blocking paradigm simulated here (Waelti, Dickinson, & Schultz, 2001).

Overshadowing and Summation

Training on a compound stimulus (e.g., AX+) causes the acquisition of associative strength to the individual elements to be reduced relative to conditioning to each of the elements alone and relative to strength to the compound itself. The actual pattern of results that is seen depends on the preexisting salience of the two stimuli, with the more salient stimulus overshadowing the lesser. When both stimuli are of equal salience, learning about both is reduced roughly equally. This later condition is referred to as *mutual overshadowing*. The complement of overshadowing is known as *summation*, when two elements are conditioned separately (e.g., A, X) and the compound (AX) is then tested, exhibiting greater conditioning. Both of these effects can be explained through the Rescorla–Wagner rule, in which elements combine additively in producing the overall reward expectation. Therefore, individual elements get less reward association when trained in combination, and the combination of individually trained elements produces greater activation. Figures 4B and C show that these effects hold in the PVLV model.

Conditioned Inhibition

Stimuli that predict an absence of reward in the presence of other stimuli that previously predicted reward take on a negative reward association, known as *conditioned inhibition*. To demonstrate this, we first trained the model on A+, followed by AX− (while continuing to train on A+). Then, we tested A and X, and,

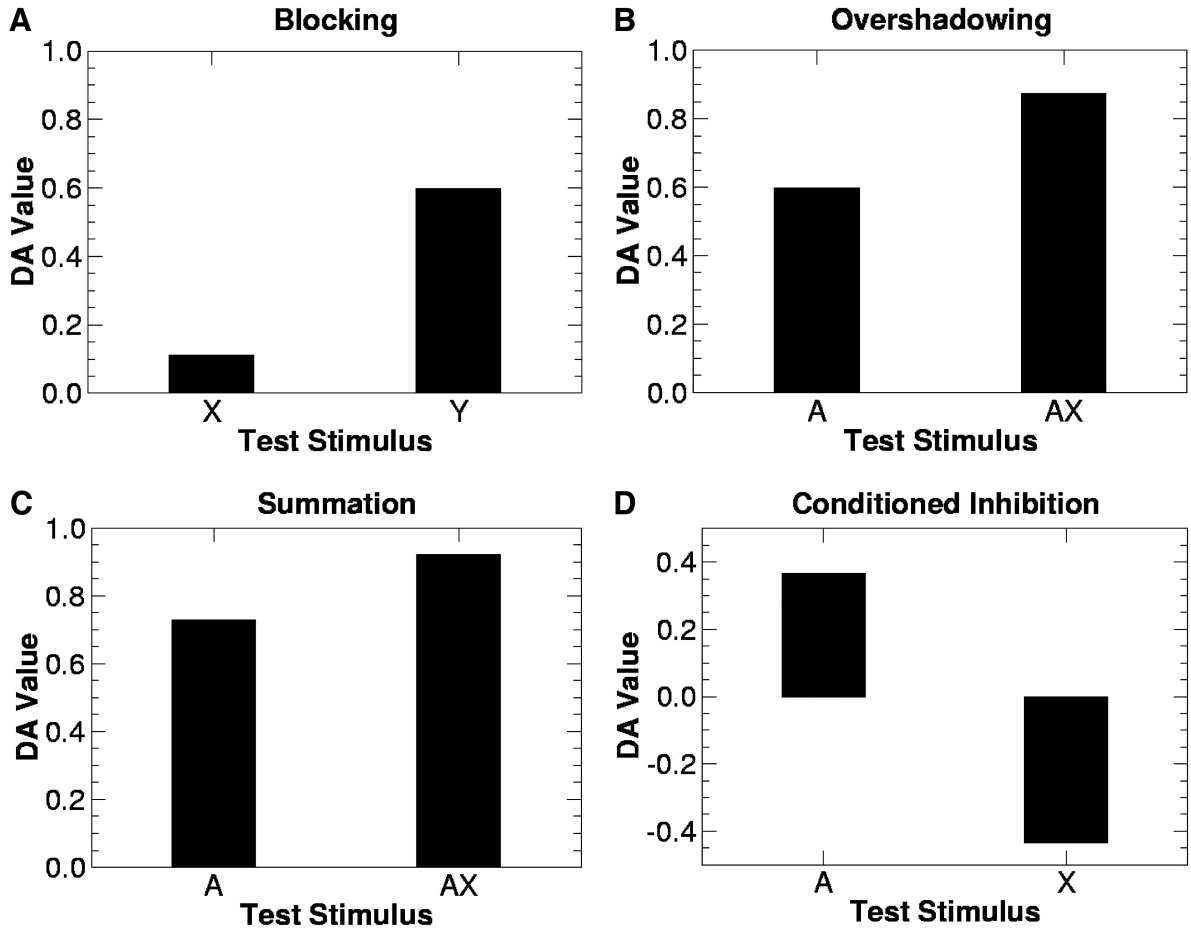

Figure 4. Results from basic conditioning paradigms, showing overall dopamine output (DA) value to onset of indicated conditioned stimulus. A: Blocking in training of X in AX+ from prior training on A+ compared with Y in BY+ where B− was previously trained. B: Overshadowing of elemental A associations by training on an AX compound. C: Summation of elemental associations (A+, X+) in testing the compound AX. Note that the difference in response to A in Panels B and C reflects the overshadowing effect as well. D: Conditioned inhibition for X produced by initial conditioning on A+ followed by AX−.

as shown in Figure 4D, X took on an overall negative (i.e., less than the .50 neutral baseline) reward association. This is consistent with dopamine recording data showing a decrease in firing to the X (Tobler, Dickinson, & Schultz, 2003).

Trace Versus Delay Conditioning

Delay conditioning refers to the type of conditioning problems we have considered so far, in which the CS input remains active through the time of reward. In the trace conditioning paradigm, the CS is instead removed before the reward is delivered, thereby requiring the animal to maintain some kind of internal "trace" of the stimulus. It is well established that trace conditioning depends on the integrity of the hippocampus and prefrontal cortex (PFC), whereas delay conditioning does not (e.g., Kronforst-Collins & Disterhoft, 1998; Weible, McEchron, & Disterhoft, 2000). The standard account of these findings is that these additional brain systems are necessary for maintaining the stimulus trace through to the point of reward so that a reward association can be established. This is exactly what the PVLV model requires, by virtue of all PVLV learning occurring at the time of (expected) rewards. Thus, if a neural representation of the stimulus is not active at the time of reward, no conditioned association can develop in the PVLV system.

However, it is not as clear how TD models would account for the need for PFC or hippocampus in trace conditioning. In principle, TD should be able to use chaining to bridge the gap between CS offset and reward onset, as long as there is some reliable sequence of neural activation states that follows from the CS presentation. This is especially true when lambda is high in TD(λ) models because of the correspondingly long eligibility traces of the CS, which should persist after the CS goes off. Although it is possible to propose that the eligibility trace is somehow implemented in PFC or hippocampus, this would then imply that the system without these areas behaves like TD(0). Such a system would likely be less robust to variability in the CS–US relationship, as explored below. This could be easily tested by using

published paradigms (e.g., Kirkpatrick & Church, 2000). If variability in the CS–US timing (all within delay conditioning) did not affect conditioning with PFC and/or hippocampal lesions, then this would present a challenge for the TD framework.

Second-Order Conditioning

S*econd-order conditioning* refers to the ability of a conditioned stimulus (X) that has become associated with reward (US) to then drive the learning of behavioral associations for a second CS (A) that reliably predicts the occurrence of the first one (X). That is, after second-order conditioning, if the X had acquired the ability to elicit a particular conditioned response (CR; e.g., salivation in Pavlov's dogs), the A will now elicit a very similar response. This paradigm provides an interesting functional distinction between PVLV and TD. In TD, the extension of reward association to the A from the DA burst at the onset of the X occurs naturally through the chaining mechanism, which just moves the burst backward in time to the A onset. Thus, in TD, CS-driven bursts are no different than reward-driven bursts in their ability to train further reward associations. This in turn implies that third-order and higher conditioning ought to occur as easily as does second-order conditioning, which apparently is not the case. We were unable to find a single report of third-order conditioning in the literature, and only one article that even mentions the issue of third-order conditioning, stating that it must be very weak if it exists at all (Dinsmoor, 2001; see also Denny & Ratner, 1970, for a brief discussion on lack of third-order and higher conditioning). This should perhaps be of concern for advocates of the TD model.

PVLV, in contrast, makes a strong distinction between first- and second-order conditioning (which is supported by lesion data, as discussed later). First-order conditioning (e.g., to the X) occurs because the X is present at the time of the primary reward, and thus the PV filtering on the LV system allows the LVe to learn positive reward associations to the X. Critically, the CS-driven bursts from the LVe system do not cause further training of the LVe representation: both LVe and PVi learn only at the time of actual (primary) rewards (and recall that LVe is trained not by a delta-like dopamine value but rather directly by the US/PVe signal). If the LVe-driven dopamine bursts at the time of CS onset were able to drive learning in the LVe itself, this would result in a positive feedback loop that would quickly saturate weights at some maximum value and effectively make extinction impossible (see later discussion for more on this issue). Nonetheless, CS-driven dopamine bursting does directly train up association weights in the rest of the system, including those between second-order CSs (e.g., A) and conditioned responses (CR), and between sensory representations themselves (i.e., CS–CS associations, as in A and X becoming associated in cortex).

One critical piece of data about second-order conditioning is that extinguishing the first-order CS (X) does not typically affect the ability of the second-order CS (A) to activate the conditioned response (Rizley & Rescorla, 1972). This finding contrasts with sensory preconditioning, which involves associating A and X repeatedly prior to conditioning, then doing first-order conditioning of X. This results in a reward association for A that is mediated entirely through X. Thus, when X is extinguished, it takes A down with it (Rizley & Rescorla, 1972). Taken together, this pattern of results suggests that CS–CR associations are more important than CS–CS associations in second-order conditioning. All of this is compatible with PVLV, in that the first-order CS dopamine bursts drive learning of both types of associations, but it is easy to imagine that CS–CR associations are more quickly learned and more dominant.

However, the clear prediction from PVLV is that a second-order CS will be dissociable from a first-order one in its ability to drive the LVe system: Only first-order CSs will have direct learned weights into the LVe system and will be capable of driving dopamine firing at CS onset. Thus, to the extent that a second-order CS does drive dopamine firing (which has not been tested to our knowledge), this dopamine firing will be mediated through the first-order CS and will thus extinguish if this first-order CS is extinguished (as in sensory preconditioning). There is no reason to expect TD to make such a distinction, and therefore this stands as an important distinguishing testable prediction. Furthermore, the apparent lack of higher order conditioning above second-order is consistent with the idea that the second-order CS does not typically acquire the same dopamine-firing ability as the first-order CS. Finally, another class of predictions, which we discuss more fully later, is that other brain areas outside of those in the core PVLV system should be critical for second-order conditioning.

To provide a simple demonstration of the principle of second-order conditioning in PVLV, we implemented the CS–CS association version of this effect (which we know is not the dominant case) because we do not have a good representation of conditioned responding in this simple model to capture the learning of CS–CR associations (other models, e.g., of the basal ganglia system could be used to model this effect; Frank, 2005). We included simulated cortical areas (and possibly basolateral amygdala, as discussed later) in our model that learn CS–CS associations via DA modulation produced by PVLV (Figure 5A). When the already conditioned X stimulus is activated in conjunction with the A in the A→X training case, the X drives a LVe-mediated DA burst that then increases cortical associations among the A and X representations. When the A is later presented alone, these cortical weights drive the reactivation of the X representation, leading to a DA burst at A onset (Figure 5B).

CS–US Timing Variability

As mentioned earlier, one feature of PVLV relative to TD is that it should be more robust to variability in the CS–US relationship. TD depends on a precise chain of events from CS onset to US onset, whereas PVLV simply learns that a CS has been reliably associated with a US and does not require such a causal chain of events between the two. This difference can be examined by introducing variability in the timing of the CS–US relationship, as has been done in behavioral studies showing no detriments in conditioning from this variability (H. Davis et al., 1969; Kamin, 1960; Kirkpatrick & Church, 2000).

We constructed two different kinds of environments that have randomized sequences of events between the CS and US: a basic randomized CS–US interval task (replicating the behavioral studies) and a simulated working memory task (O'Reilly & Frank, 2006). For the first test, we introduced a random delay between CS and US that varied in the set (3, 6, 12). In addition, the probability of reward was manipulated in the set (.10, .20, .50, 1.00), to explore sensitivity to the strength of association. A control stim-

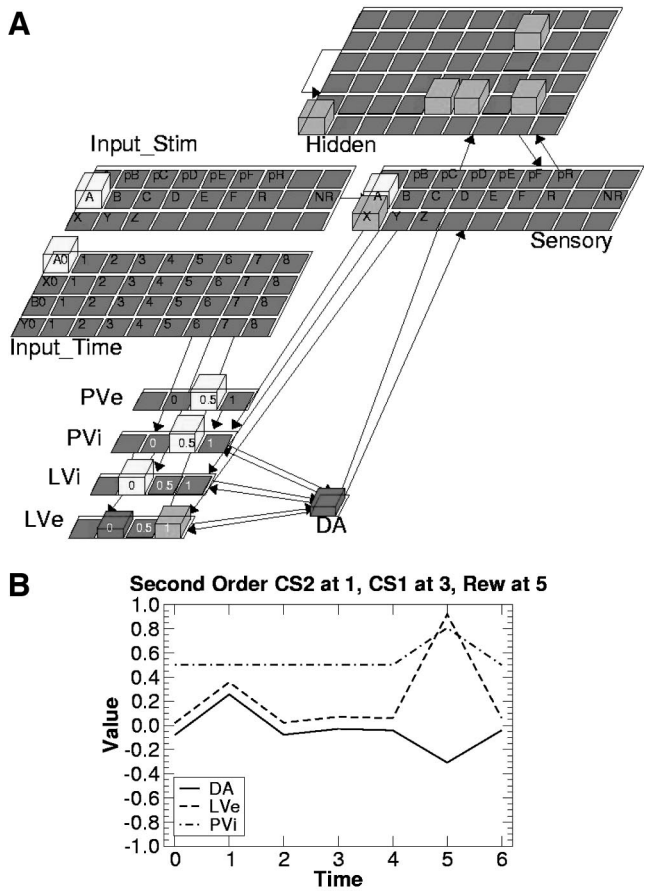

Figure 5. Second-order conditioning (A) model and (B) activation patterns. Second-order conditioning occurs in primary value learned value (PVLV) via dopaminergic or other modulation of learning in cortical systems and not directly within the PVLV system itself (unlike temporal-differences and other chaining algorithms). The sensory and hidden layers represent these cortical areas (sensory cortex and higher order association cortex) and are modulated by the overall dopamine output (DA) layer. After training (first X+ then AX intermixed with X+), the A input activates cortical representations that were originally associated with the X stimulus, causing the representation of the X to be activated in the sensory layer, leading to a learning-value-mediated DA burst at the onset of the A. This is the pattern of activation shown in the model, corresponding to Time 1 in the activation plot. The X stimulus is activated at Time 3, and no reward (Rew) is delivered in this second-order trial (but is expected on the basis of the X+ training) at Time 5 (2 steps after X onset). Stim = stimulus; PVe = primary value excitatory; PVi = primary value inhibitory; LVi = learned value inhibitory layer; LVe = learned value excitatory layer; CS = conditioned stimulus.

ulus that did not predict reward was also included. Furthermore, the delay interval was filled with randomly selected "distractor" stimuli (in addition to the maintained representation of the CS), to simulate ongoing neural activity and other sensory events in the world. A random selection of distractors was also presented between one CS–US trial and the next.

We compared the PVLV model against a TD(0) model implemented in the same basic neural framework as our PVLV model. To give the TD mechanism as good an opportunity to learn as possible, we also included a timing-like signal across a set of units that counted up (i.e., moved from one unit to the next) starting at the onset of the CS. A range of different TD learning parameters were explored, including discount factor gamma and learning rate epsilon. To keep things simple and the input representations the same across the two models, we did not introduce a delta parameter as in TD(λ). However, the CS remained fully active during the entire CS–US interval, so this would not be affected by the delta parameter. A simple implementation of the delta effect would simply smear out the stimuli over time, and while this would smooth over some of the noise, it will also smooth over the signal (CS onset) as well. Furthermore, the optimal delta value will vary depending on the length of the CS–US interval. Although there may be better solutions to these problems, the main point is to demonstrate that PVLV is naturally robust to this variability, not that it is impossible to construct some version of TD to deal with it.

Figure 6 shows the average DA (δ) value at the onset of the reward predictive CS across 10 runs for the best-performing TD parameters compared with those of the PVLV network, for reward probabilities of .50 and .20. Although the TD network can learn about the predictive nature of the CS for the higher probability reward and shorter maximum delay lengths, it fails to consistently deliver a positive DA burst as the probability of reward goes down and maximum delay length increases. In contrast, PVLV learns reliably at the longest delays and the lowest probabilities of reward.

In the second test, we simulated a Pavlovian version of the 1–2–AX working memory task explored in O'Reilly and Frank (2006). In this task, the network has to respond to target sequences (e.g., A–X or B–Y), and the currently relevant target sequence is determined by a control signal (1 or 2). That is, if a 1 is presented, the target sequence is A–X and continues to be A–X until a further control signal is presented. If a 2 is presented, then the target is B–Y until the next control signal. Thus there are several "inner loops" (A−X, A−Y, etc.) for each "outer loop" (1 or 2). In our Pavlovian version, the outer loop is either an A or a B, which determines whether an X or a Y will be rewarding over a series of subsequent trials (if A, then X is rewarding; if B, then Y is rewarding). Random distractor stimuli can appear at any time, and the number of inner loops of X, Y, or distractor stimuli was randomly varied from 1–4. The outer loop stimulus was maintained in the input for the duration of these inner loops (to remove working memory demands from the simple Pavlovian model which is not equipped with the prefrontal mechanisms simulated in O'Reilly and Frank, 2006, to support working memory). The probability of activating one of the target stimuli (X for A, Y, for B) relative to the distractors was varied in the set (.10, .20, .50, 1.00).

Figure 7 shows similar results to the random delay condition, plotting the DA burst for the outer loop stimuli (A or B). Here, the maintenance duration was only a maximum of four trials but because randomly triggered rewards from the variable number of inner loop targets occurred during the "maintenance" period of the outer loop, this introduced greater unpredictability in the TD chaining mechanism for associating reward with the outer loop stimulus. This task closely mimics the reward structure of the 1–2–AX working memory task, and clearly demonstrates the problems that the TD chaining mechanism causes in such tasks. Although some improvement in the performance of TD can almost

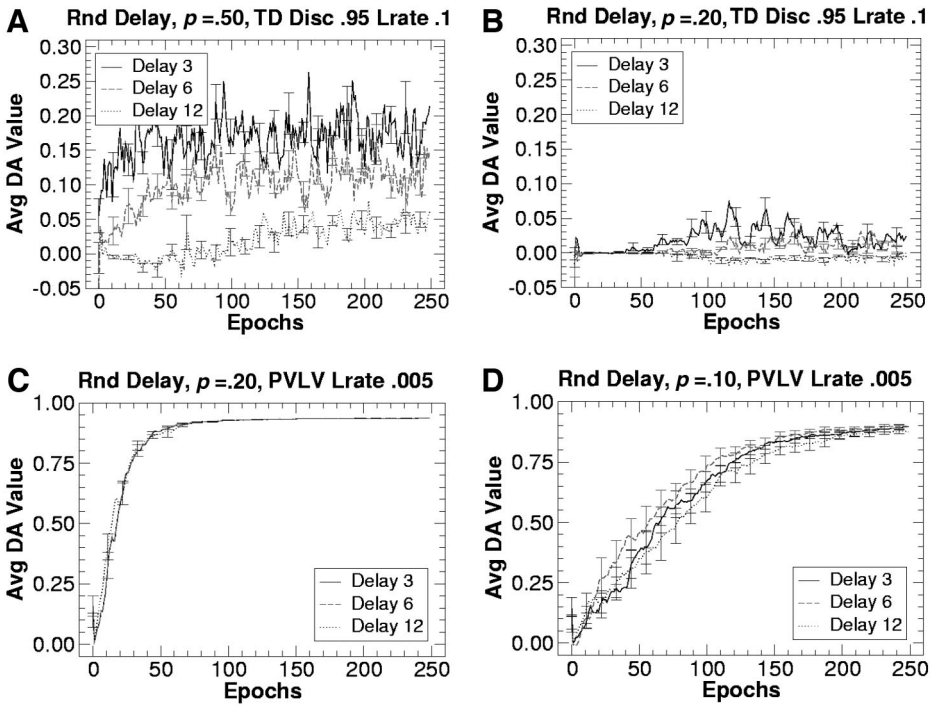

Figure 6. Average (Avg) conditioned stimulus (CS) overall dopamine output (DA) temporal-differences (TD) values for the random delay task, comparing TD (A, B) with primary value learned value (PVLV; C, D). The different delays (3, 6, 12) indicate the maximum number of intervening "distractor" stimulus trials that come between a CS and subsequent reward, with the actual number of trials determined randomly between one and the maximum. The reward is given with the probability indicated (p = .50, .20, or .10, respectively). TD breaks down with longer delays and lower probabilities of reward, being unable to reliably propagate the reward prediction back to the CS onset to trigger a DA burst. In contrast, PVLV remains reliable even for p = .10 and a delay of 12. Note the differences in scale. Results are averages over 10 runs. Rnd Delay = random delay; Disc = discount rate; Lrate = learning rate; Epoch = one pass through the training patterns.

certainly be achieved through various means as mentioned earlier, it is clear that the PVLV mechanism is intrinsically more reliable for these kinds of tasks. We argue that this robustness may be critical for functioning in the real world.

Biological Mapping of PVLV

The above simulations show that PVLV can account for a range of Pavlovian behavioral phenomena and that its simple learning mechanisms can be more robust than the predictive temporal learning in the TD algorithm. Perhaps the strongest support for this framework as a model of reward learning in the brain comes from the good fit between the functional properties of the PV and LV components of PVLV and those of brain areas that are known to support reward learning. We summarize this below (Figure 8; see Hazy, Frank, and O'Reilly (2006) for a more detailed discussion).

PVe = Lateral Hypothalamic Area (LHA)

The PVe component of the model is responsible for providing excitatory drive to dopamine cells in response to primary rewards. There is considerable evidence that neurons in the LHA reliably respond to these primary rewards without habituation (Nakamura & Ono, 1986; Ono, Nakamura, Nishijo, & Fukada, 1986) and have excitatory projections into the midbrain dopaminergic nuclei (VTA/SNc) both directly and even more strongly via the pedunculopontine tegmental nucleus (PPTN) (Semba & Fibiger, 1992). There is also direct evidence that activation of the PPTN leads to dopamine bursting (Floresco, West, Ash, Moore, & Grace, 2003).

PVi = Patch-Like Neurons in the Ventral Striatum

The PVi system learns to expect primary rewards and to inhibit the dopamine firing that would otherwise occur in response to them. Furthermore, when expected rewards are withheld, their inhibitory influence results in a dip or pause in dopamine firing. Striosome/patch-like neurons in the ventral striatum (which appear to be especially prevalent in the shell of the nucleus accumbens in rats but that are also distributed throughout the ventral striatum and nucleus accumbens) send direct inhibitory projections into the VTA and SNc (for a review, see Joel & Weiner, 2000). Unlike striosomes of the dorsal striatum, which project only to ventral tier dopamine cells mostly in the SNc, those in the ventral striatum project globally to the entire VTA and SNc putting them in a unique position to inhibit burst firing globally. Electrophysiologically, some neurons in the ventral striatum have been shown to fire immediately prior to primary rewards (Cromwell & Schultz, 2003;

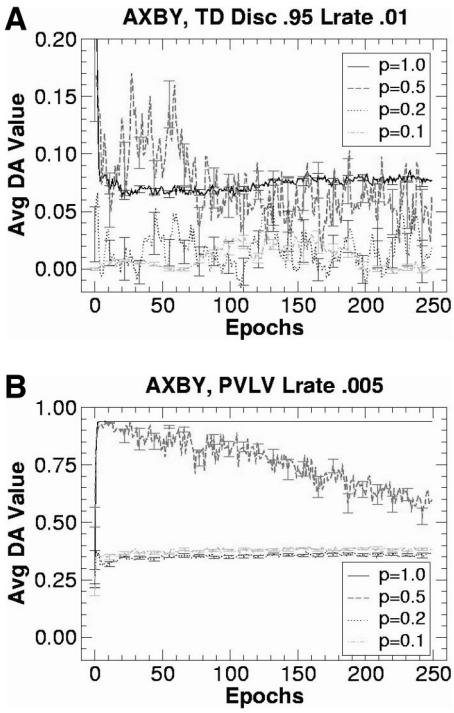

Figure 7. Average (Avg) "outer loop" conditioned stimulus (CS) overall dopamine output (DA) temporal-differences (TD) values for the AXBY task, comparing TD, (A) with primary value learned value (PVLV), (B) for different probabilities of the target stimulus (X for A, Y for B) appearing within the inner loop and triggering a reward (otherwise, a randomly chosen distractor would appear). Note the differences in scale. As in the random delay task shown in the previous figure, the PVLV algorithm provides much more reliable DA bursts compared with the TD mechanism. Results are averages over 10 runs. Disc = discount rate; Lrate = learning rate; Epoch = one pass through the training patterns.

Deadwyler, Hayashizaki, Cheer, & Hampson, 2004; Schultz, Apicella, Scarnati, & Ljungberg, 1992), critically including firing at the expected time of reward when no rewards are actually presented (i.e., in "extinction" trials).

LVe = Central Nucleus of the Amygdala (CNA)

The LVe component of the model learns reward associations from sensory inputs (CSs), and drives excitatory dopamine bursts in response to these inputs. The CNA receives broad projections from all over the cortex, both directly and indirectly (via the subnuclei of the BLA; Amaral, Price, Pitkanen, & Carmichael, 1992). These CNA neurons can then drive excitatory dopamine firing in the VTA and the SNc (Ahn & Phillips, 2003; Fudge & Haber, 2000; Rouillard & Freeman, 1995). There are direct and indirect (e.g., via PPTN) projections from CNA to VTA/SNc (Fudge & Haber, 2000; Wallace, Magnuson, & Gray, 1992), but the exact nature of these projections remains somewhat unclear. The lateral CNA contains GABA spiny neurons much like the striatum (M. Davis, Rainnie, & Cassell, 1994), but the medial CNA neurons have a different morphology and at least some are glutametergic (Takayama & Miura, 1991). Thus, the mechanism by which CNA excites the midbrain dopamine neurons could be either direct excitation or indirect disinhibition. Electrophysiologically, multimodal CNA neurons in Pavlovian conditioning paradigms (initially responsive to only a US) learn to fire also for an associated CS (Ono et al., 1995). In response to a visual stimulus predictive of reward, immediate early gene expression was observed in CNA cells, particularly those that project to SNc (Lee et al., 2005). Further, disconnection of the CNA and the SNc prevented the acquisition of conditioned responses, without affecting the acquisition of food-related responses (which would be served by the PV system). Other lesion studies show that the CNA is important for supporting Pavlovian learning, but not the expression of learned behavior, to a CS (e.g., El-Amamy & Holland, 2006; Groshek et al., 2005; Killcross, Robbins, & Everitt, 1997). The CNA is also critical for a nonspecific form of Pavlovian instrumental transfer (PIT), whereby a Pavlovian CS can provide a nonspecific enhancement of instrumental responding for a different US than that paired with the CS (Corbit & Balleine, 2005). These PIT data are consistent with PVLV in that the CNA (LVe), having been activated by the CS, can drive global dopamine firing that can then facilitate instrumental responding in a nonspecific way.

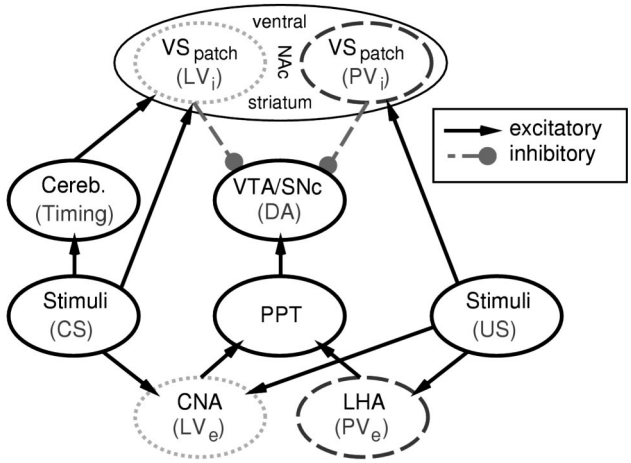

Figure 8. Biological mapping of the primary value learned value (PVLV) algorithm. Excitatory drive on the midbrain dopamine system comes from the lateral hypothalamus (LHA) and central nucleus of the amygdala (CNA). These project (via the pedunculopontine tegmental nucleus [PPT]) to the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) midbrain dopamine neurons. The CNA learns to associate arbitrary stimuli (CSs) with rewards (unconditioned stimuli [USs]), and drives dopamine firing at the onset of CSs, consistent with the properties of the learned value excitatory (LVe) system. The LHA responds to rewards and drives dopamine bursting at the time of US onset, consistent with the primary value excitatory (PVe) system. The patch/striosome-like neurons of the ventral striatum (VS patch) send direct inhibitory projections to the entire VTA and SNc and learn to fire at the time of US and CS onset. Thus, they correspond to the primary value inhibitory (PVi; firing at US onset) and learned value inhibitory (LVi; firing at CS onset) systems. Note that striosome/patch cells in the dorsal striatum have more focused connectivity with the dopamine nuclei, whereas these in the ventral striatum project globally, as is required for the PVi/LVi system.

LVi = Patch-Like Neurons in the Ventral Striatum

The LVi neurons perform the same role for learned CSs as the PVi neurons do for primary rewards: They learn (very slowly) about the reward association of these stimuli and serve to cancel out excitatory dopamine bursts that would otherwise occur. The above anatomical data on patch-like neurons in the ventral striatum are consistent with this LVi role. Furthermore, subpopulations of these ventral striatum neurons exhibit CS-related firing (Schultz et al., 1992; Cromwell & Schultz, 2003), whereas others exhibit the US-related firing noted above for PVi.

Beyond these core PVLV-associated areas, the biology of reinforcement learning ultimately encompasses much of the brain. In closely related work, we have developed a model of how the dorsal areas of the basal ganglia interact with the frontal cortex to control both motor outputs and working memory updating (Frank, 2005, 2006; O'Reilly & Frank, 2006; Frank, Loughry, & O'Reilly, 2001). The frontal working memory system is critical for a variety of reinforcement learning functions. For example, as noted earlier, it can support trace conditioning by maintaining active representations of stimuli after they disappear from the environment (e.g., Kronforst & Collins & Disterhoft, 1998; Weible et al., 2000). Without this system, the basic PVLV model with immediate sensory inputs can only support delay conditioning. The orbital prefrontal cortex may also be critical for supporting working memory of reward-related information (e.g., Frank & Claus, 2006; Hikosaka & Watanabe, 2000; Schoenbaum & Roesch, 2005; Tremblay & Schultz, 1999), which can then provide top-down goal-driven influences on the PVLV mechanisms. Simulations of these functions and their interactions with the basal ganglia demonstrated that they can potentially account for various other reinforcement learning phenomena, such as devaluation and decision-making deficits under orbital lesions, and are also important for reversal learning (Frank & Claus, 2006).

In addition, the hippocampus can support the formation of conjunctive stimulus representations that play a role in various forms of reinforcement learning, including nonlinear discrimination learning (e.g., Alvarado & Rudy, 1995), sensory preconditioning (Port & Patterson, 1984), and transitive inference (Dusek & Eichenbaum, 1997). We have developed computational models of these phenomena (e.g., Frank, Rudy, & O'Reilly, 2003; O'Reilly & Rudy, 2001), and future work will integrate these models with the PVLV mechanisms.

Finally, the role of the BLA in second-order conditioning and a variety of other reinforcement learning paradigms has yet to be fully integrated within our framework. Our overall position is that the BLA plays a somewhat similar role to the CNA in that it is important for associating CSs with USs, but it does so in a more specific way that involves direct synaptic connections with neurons in the ventral striatum and the limbic prefrontal cortex. In contrast, the CNA has a more generalized effect consistent with its role in driving global dopaminergic firing. In other words, we think of BLA as more like a "cortical" area, whereas CNA is more clearly subcortical. This is generally consistent with the ideas and data of Cardinal, Parkinson, Hall, and Everitt (2002); Swanson (2003); and Corbit and Balleine (2005).

Testable Predictions of the Model

We previously discussed several testable predictions that would discriminate between PVLV and TD. On the basis of the above anatomical mappings, we can make a number of additional predictions about possible experiments that would strongly test the PVLV model.

The striosome/patch neurons in the ventral striatum should be specifically responsible for producing the dip in dopamine firing at the time of expected rewards, when rewards are not delivered. This could be tested by lesioning the nucleus accumbens shell in rats (where such neurons are concentrated) and recording from VTA/SNc neurons in a standard Pavlovian paradigm with extinction.

At a behavioral level, the VS striosome/patch neurons should be specifically important for extinction and reversal learning, which should depend on the dopamine dip for expected but not delivered rewards. There is some data consistent with this prediction (Ferry, Lu, & Price, 2000; Robbins, Giardini, Jones, Reading, & Sahakian, 1990).

The CNA should be specifically responsible for producing the dopamine burst to CS onset. This could be directly tested by lesioning the CNA and recording from VTA/SNc in a standard Pavlovian paradigm.

At a behavioral level, CNA should be critical for learning working memory tasks, in which working memory update actions (driven by more central regions of the dorsal striatum that project to prefrontal cortex) must be activated at the onset of task-relevant stimuli to encode these stimuli into working memory (O'Reilly & Frank, 2006).

The CNA should also be important for autoshaping, where an animal learns to approach and/or act on a conditioned stimulus (CS) after Pavlovian conditioning. According to the PVLV model, CS- and CNA-driven dopamine bursts should be particularly critical in reinforcing motor actions directed toward the CS because dopamine bursts at the time of US delivery generally occur too late to have an effect on CS-directed behaviors. This is consistent with the account provided by Cardinal, Parkinson, Hall, and Everitt (2002), where it was found that CNA lesions severely impair autoshaping conditioned responses (Parkinson, Robbins, & Everitt, 2000). Critically, CNA lesions after training (i.e., autoshaping has already taken place) do not impair expression, which is also consistent with our learning-based model (Cardinal, Parkinson, Lachenal, et al., 2002).

As noted earlier, the PVLV model strongly predicts that CNA (LVe) should not be important for second-order conditioning. The LV system only learns from primary rewards (i.e., the PV system). If this were not the case, then the dopamine burst driven by the LVe to a CS onset would cause the LVe system to increase its reward association to that same CS, leading to a runaway positive feedback loop. The consequence is that the LVe-driven dopamine bursting cannot drive learning within the LVe itself, and thus the LVe should not be able to support second-order conditioning. This prediction is consistent with data showing that the BLA, but not the CNA, is important for second-order conditioning (Hatfield et al., 1996). More rigorous tests could solidify this prediction and provide a clearer picture of the precise role of the BLA in this case. Again, see Hazy et al. (2006) for more detailed biological discussion and further predictions.

Comparison With Brown, Bullock, and Grossberg (1999) and Other Models

The PVLV model shares several features in common with the dopamine bursting in the basal ganglia (BBG) model (Brown et al., 1999). The BBG model features two pathways that converge on the substantia nigra pars compacta (SNc) midbrain dopamine nucleus. One pathway goes via the ventral striatum through the ventral pallidum and PPTN, whereas the other goes via the striosomes of the dorsal striatum. External rewards directly excite the SNc via the ventral striatal pathway through nonmodifiable projections. The ventral striatal pathway produces a net excitatory effect on the SNc dopamine neurons, whereas the striosomal pathway is adaptively delayed in time (via an intracellular spectral timing mechanism) and has an inhibitory connection. Both of these pathways are trained by dopamine from the SNc. The net effect is that, after standard Pavlovian training (e.g., Figure 1), a CS input produces an initial excitatory dopamine burst via the ventral striatum pathway, whereas the striosomal pathway produces an adaptively delayed inhibition of the SNc that cancels out the dopamine burst that otherwise would have been caused by the external reward input.

Thus, this ventral striatum pathway achieves an effect similar to the LV system in PVLV (i.e., CS-onset firing), whereas the striosomal pathway achieves an effect similar to the PV system (i.e., reward burst canceling). In this respect of having two separable systems achieving these two functions, the BBG model and PVLV are similar. Furthermore, the ventral striatum pathway in the BBG model also has a habituation mechanism to prevent continued dopamine firing to the CS, similar to the synaptic depression mechanism in the LV system of PVLV.

Despite these similarities, the two models also have some important differences. Perhaps the most obvious is that the reward-canceling striosome pathway in the BBG model operates only on the basis of time since CS onset, whereas the PV system in PVLV is a more generalized system for canceling reward bursts, which can use timing signals or any other form of external or internally generated inputs (as advocated by Savastano & Miller, 1998). As noted earlier, animals can learn just fine with variable CS–US delays (H. Davis et al., 1969; Kamin, 1960; Kirkpatrick & Church, 2000). Indeed, by virtue of its strong reliance on time in predicting rewards, the BBG model suffers from much of the same dependence on future predictability as does the TD model, which the PVLV model avoids.

There are several other differences between the two models. For example, the SNc dopamine in the BBG model trains up both the ventral striatum and striosome pathways in the same way, whereas PVLV has an asymmetry in the way that the PV and LV systems are trained (PV filters the training of the LV). One consequence of this is that the CS-onset dopamine burst in the BBG model reinforces the very weights that generated this burst, causing a positive feedback loop that drives weights to their high-saturation value. This would interfere with the system's ability to accurately represent statistical frequency information with graded weight values. Also, it is not clear how the BBG model would account for the blocking effect because the dopamine burst triggered by the onset of the already conditioned CS should drive the system to learn about the new to-be-blocked CS when it is subsequently introduced. In other words, blocking occurs in these dopamine-based models by virtue of the dopamine burst being canceled when it is expected, but the CS-onset burst produced by the ventral striatum in the BBG model is never canceled and thus should not produce blocking. The PVLV model avoids this problem by virtue of the LV-generated burst not training itself: All learning in the PVLV system itself is driven by the PV-modulated dopamine bursts, which are subject to cancellation and thus blocking.

In terms of biological mapping, the two models are also fairly different. We associate the ventral striatum striosomal neurons with the PVi system, whereas the BBG model attributes this same basic function to dorsal striatum striosomal neurons. The BBG model holds that the ventral striatum nonstriosomal neurons drive excitatory CS-related dopamine bursts, whereas this LVe-like function is associated with the CNA in our model. We think there is considerably more evidence for the CNA's role in this type of function, as elaborated earlier. Furthermore, our more general basal ganglia model suggests that the disinhibitory pathway from the striatum to the SNc via the ventral pallidum is responsible for disinhibiting dopamine release for actions that were initiated by "go" signals in the striatum and not for directly activating dopamine bursting (O'Reilly & Frank, 2006). This disinhibitory, but not bursting, role for the pallidal connections is consistent with data from Floresco et al. (2003), who showed that activation of the ventral pallidum increased the overall amount of dopamine activation (which we suggest results from disinhibiting more dopamine neurons) but did not increase the extent of bursting. In contrast, activation of the PPTN resulted in increased bursting, which is consistent with the idea that the CNA and LHA drive bursting via their projections to the PPTN.

A model sharing some general properties of the BBG model was proposed by Contreras-Vidal and Schultz (1999), who also posited a spectral timing function for the striosomes with inhibitory projections to the dopamine neurons to cancel expected rewards. They attributed the excitatory function more to the limbic prefrontal cortex, though the exact biological mechanisms of this are less clearly tied to the known anatomy of the basal ganglia. Nevertheless, overall, this model shares the notion that there are two systems driving dopamine firing with both the BBG and PVLV models.

In terms of specific hypotheses as to how the unitary TD computation is thought to be computed neurally, Houk et al. (1995) proposed the first model, and most of the other TD-based models have adopted their general ideas (to the extent that they are explicit about the underlying biological mechanisms). Houk et al. (1995) proposed that the striosome/patch neurons in the striatum of the basal ganglia represent the current value estimate $\hat{V}^t$, and that two different projections to the midbrain dopamine neurons compute the delta value as the temporal derivative of the value estimate as it changes over time (i.e., from t to $t + 1$). The temporal derivative can be computed in terms of a fast, indirect, net-excitatory projection (via the subthalamic nucleus) and a slower direct inhibitory projection. If the striosome firing rate is steady, the two pathways cancel each other out. An increase in striosome activation leads to more excitation initially (via the fast indirect pathway), which is then canceled out by the slower inhibition. The opposite occurs for reductions in striosome activation. In evaluating this proposal, Joel et al. (2002) concluded that it was not particularly consistent with the known anatomy of DA connectivity. Also, existing data appear to suggest that, if anything, this

circuit would more likely compute fast inhibition and slower disinhibition (excitation), which would compute the negative of the temporal derivative. Furthermore, we argue that the requirement that the TD model has a single system as the source of both CS-related and US-related dopamine firing is inconsistent with the data reviewed above, which instead seems more consistent with (at least) two components, as in PVLV.

Finally, Dayan (2001; see also Dayan & Balleine, 2002) presented an important model that extends the standard TD-based reinforcement learning paradigm to include motivational systems, to account for a variety of behavioral findings that are inconsistent with existing TD-based models. Although beyond the scope of this article, this model posits distinct roles for the amygdala, orbital frontal cortex (OFC), and the core and shell of the nucleus accumbens. The amygdala and OFC in this model learn to estimate the value function, which is more consistent with the biological mapping in the PVLV model than prior TD-based models. The issues of stimulus substitution and devaluation addressed by this Dayan (2001) model are discussed from the PVLV framework in Hazy et al. (2006; see also Frank & Claus, 2006), and future work should focus on addressing these important issues more concretely in the model.

Conclusion

In this article, we present a new computational model that can account for observed patterns of dopamine firing that are thought to underlie Pavlovian conditioning and other related forms of reinforcement learning. This model avoids the dependence on predictable chains of events that the temporal differences model has, making it better able to produce useful learning signals in more chaotic environments. We argue that such environments are more characteristic of the natural world in which such learning mechanisms need to operate. Furthermore, the two components of the PVLV mechanism provide a good fit with the underlying neurobiological substrates of the central nucleus of the amygdala and the neurons of the ventral striatum that provide inhibition of the midbrain dopaminergic nuclei. Considerable additional work remains to be done to apply this framework to the vast literature on reinforcement learning paradigms and their biological substrates.

References

Abbott, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997, January 10). Synaptic depression and cortical gain control. *Science, 275,* 221–224.

Ahn, S., & Phillips, A. G. (2003). Independent modulation of basal and feeding-evoked dopamine efflux in the nucleus accumbens and medial prefrontal cortex by the central and basolateral amygdalar nuclei in the rat. *Neuroscience, 116,* 295–305.

Alvarado, M. C., & Rudy, J. W. (1995). A comparison of configural discrimination problems: Implications for understanding the role of the hippocampal formation in learning and memory. *Psychobiology, 23,* 178–184.

Amaral, D. G., Price, J. L., Pitkanen, A., & Carmichael, S. T. (1992). Anatomical organization of the primate amygdaloid complex. In *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 1–66). New York: Wiley.

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.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience, 19,* 10502–10511.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews, 26,* 321–352.

Cardinal, R. N., Parkinson, J. A., Lachenal, G., Halkerston, K. M., Rudarakanchana, N., Hall, J. et al. (2002). Effects of selective excitotoxic lesions of the nucleus accumbens core, anterior cingulate cortex, and central nucleus of the amygdala on autoshaping performance in rats. *Behavioral Neuroscience, 116,* 553–567.

Contreras-Vidal, J. L., & Schultz, W. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of Computational Neuroscience, 6,* 191–214.

Corbit, L. H., & Balleine, B. W. (2005). Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of Pavlovian instrumental transfer. *Journal of Neuroscience, 25,* 962–970.

Cromwell, H. C., & Schultz, W. (2003). Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *Journal of Neurophysiology, 89,* 2823–2838.

Davis, H., McIntire, R. W., & Cohen, S. I. (1969). Fixed and variable duration warning stimuli and conditioned suppression. *Journal of Psychology, 73,* 19–25.

Davis, M., Rainnie, D., & Cassell, M. (1994). Neurotransmission in the rat amygdala related to fear and anxiety. *Trends in Neuroscience, 17,* 208–214.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2003). Timing and partial observability in the dopamine system. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 99–106). Cambridge, MA: MIT Press.

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

Dayan, P. (2001). Motivated reinforcement learning. In D. S. Touretzky, M. C. Mozer, & ME. Hasselmo (Eds.), *Advances in neural information processing systems 13* (pp. 11–18). Cambridge, MA: MIT Press.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron, 36,* 285–298.

Deadwyler, S. A., Hayashizaki, S., Cheer, J., & Hampson, R. E. (2004). Reward, memory and substance abuse: Functional neuronal circuits in the nucleus accumbens. *Neuroscience and Biobehavioral Reviews, 27,* 703–711.

Denny, M. R., & Ratner, S. C. (1970). *Comparative psychology: Research in animal behavior* (rev. ed.). Homewood, IL: The Dorsey Press.

Dinsmoor, J. A. (2001). Stimuli inevitably generated by behavior that avoids electric shock are inherently reinforcing. *Journal of the Experimental Analysis of Behavior, 75,* 311–333.

Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences, 94,* 7109–7114.

El-Amamy, H., & Holland, P. C. (2006). Substantia nigra pars compacta is critical to both the acquistion and expression of learned orienting in rats. *European Journal of Neuroscience, 24,* 270–276.

Ferry, A. T., Lu, X. C., & Price, J. L. (2000). Effects of excitotoxic lesions in the ventral striatopallidal–thalamocortical pathway on odor reversal learning: Inability to extinguish an incorrect response. *Experimental Brain Research, 131,* 320–335.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2005). Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating td errors. *Behavioral and Brain Functions, 1,* 1–5.

Floresco, S. B., West, A. R., Ash, B., Moore, H., & Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nature Neuroscience, 6,* 968–973.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and non-medicated Parkinsonism. *Journal of Cognitive Neuroscience, 17,* 51–72.

Frank, M. J. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks, 19,* 1120–1136.

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review, 113,* 300–326.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, and Behavioral Neuroscience, 1,* 137–160.

Frank, M. J., Rudy, J. W., & O'Reilly, R. C. (2003). Transitivity, flexibility, conjunctive representations and the hippocampus: II. A computational analysis. *Hippocampus, 13,* 341–354.

Fudge, J. L., & Haber, S. N. (2000). The central nucleus of the amygdala projection to dopamine subpopulations in primates. *Neuroscience, 97,* 479–494.

Groshek, F., Kerfoot, E., McKenna, V., Polackwich, A. S., Gallagher, M., & Holland, P. C. (2005). Amygdala central nucleus function is necessary for learning, but not expression, of conditioned auditory orienting. *Behavioral Neuroscience, 119,* 202–212.

Hatfield, T., Han, J.-S., Conely, M., & Holland, P. (1996). Neurotoxic lesions of basolateral, but not central, amygdala interfere with Pavlovian second-order conditioning and reinforcer devaluation effects. *Journal of Neuroscience, 16,* 5256–5265.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). *The biological basis of Pavlovian learning: A computational framework*. Manuscript in preparation.

Hikosaka, K., & Watanabe, M. (2000). Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. *Cerebral Cortex, 10,* 263–271.

Hochreiter, S., & Schmidhuber, J. (1997). Long short term memory. *Neural Computation, 9,* 1735–1780.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 233–248). Cambridge, MA: MIT Press.

Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cognitive Science, 27,* 403–430.

Ivry, R. (1996). The representation of temporal information in perception and motor control. *Current Opinion in Neurobiology, 6,* 851–857.

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

Joel, D., & Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: An analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience, 96,* 451.

Kakade, S., & Dayan, P. (2002a). Acquisition and extinction in autoshaping. *Psychological Review, 109,* 533–544.

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

Kamin, L. J. (1960). Acquisition of avoidance with a variable CS–US interval. *Canadian Journal of Psychology, 15,* 176–188.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive simulation* (pp. 9–31). Miami, FL: University of Miami Press.

Killcross, S., Robbins, T. W., & Everitt, B. J. (1997, July 24). Different types of fear-conditioned behavior mediated by separate nuclei within amygdala. *Nature, 388,* 377–380.

Kirkpatrick, K., & Church, R. M. (2000). Stimulus and temporal cues in classical conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 26,* 206–219.

Kronforst-Collins, M. A., & Disterhoft, J. F. (1998). Lesions of the caudal area of rabbit medial prefrontal cortex impair trace eyeblink conditioning. *Neurobiology of Learning and Memory, 69,* 147–162.

Lee, H. J., Groshek, F., Petrovich, G. D., Cantalini, J. P., Gallagher, M., & Holland, P. C. (2005). Role of amygdalo-nigral circuitry in conditioning of a visual stimulus paired with food. *Journal of Neuroscience, 25,* 3881–3888.

Lustig, C., Matell, M. S., & Meck, W. H. (2005). Not "just" a coincidence: Frontal-striatal interactions in working memory and interval timing. *Memory, 3–4,* 441–448.

Markram, H., & Tsodyks, M. (1996, August 29). Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature, 382,* 759–760.

Mauk, M. D., & Buonomano, D. V. (2004). The neural basis of temporal processing. *Annual Review of Neuroscience, 27,* 307–340.

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.

Nakamura, K., & Ono, T. (1986). Lateral hypothalamus neuron involvement in integration of natural and artificial rewards and cue signals. *Journal of Neurophysiology, 55,* 163–181.

Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions, 1*(6), 1–9.

Ono, T., Nakamura, K., Nishijo, H., & Fukada, M. (1986). Hypothalamic neuron involvement in integration of reward, aversion, and cue signals. *Journal of Neurophysiology, 56,* 63–79.

Ono, T., Nishijo, H., & Uwano, T. (1995). Amygdala role in conditioned associative learning. *Progress in Neurobiology, 46,* 401–422.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation, 13,* 1199–1242.

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,* 283–328.

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

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review, 108,* 311–345.

Pan, W.-X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience, 25,* 6235–6242.

Parkinson, J., Robbins, T. W., & Everitt, B. J. (2000). Dissociable roles of the central and basolateral amygdala in appetitive emotional learning. *European Journal of Neuroscience, 12,* 405–413.

Pavlov, I. P. (Ed.). (1927). *Conditioned reflexes.* London: Oxford University Press.

Port, R., & Patterson, M. (1984). Fibrial lesions and sensory preconditioning. *Behavioral Neuroscience, 98,* 584–589.

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

Rizley, R. C., & Rescorla, R. A. (1972). Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology, 81,* 1–11.

Robbins, T. W., Giardini, V., Jones, G. H., Reading, P., & Sahakian, B. J. (1990). Effects of dopamine depletion from the caudate-putamen and nucleus accumbens esepti on the acquisition and performance of a conditional discrimination task. *Behavioural Brain Research, 38,* 243–261.

Rouillard, C., & Freeman, A. S. (1995). Effects of electrical stimulation of the central nucleus of the amygdala on the in vivo electrophysiological activity of rat nigral dopaminergic neurons. *Synapse, 21,* 348–356.

Savastano, H. I., & Miller, R. R. (1998). Time as content in Pavlovian conditioning. *Behavioural Processes, 44,* 147–162.

Schoenbaum, G., & Roesch, M. (2005). Orbitofrontal cortex, associative learning, and expectancies. *Neuron, 47,* 633–636.

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

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

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience, 13,* 900–913.

Schultz, W., Apicella, P., Scarnati, D., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience, 12,* 4595–4610.

Semba, K., & Fibiger, H. (1992). Afferent connections of the laterodorsal and the pedunculopontine tegmental nuclei in the rat: A retro- and anterograde transport and immunohistochemical study. *Journal of Comparative Neurology, 323,* 387–410.

Sporns, O., & Alexander, W. H. (2002). Neuromodulation and plasticity in an autonomous robot. *Neural Networks, 15,* 761–774.

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.

Suri, R. E., & Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Computation, 13,* 841.

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning, 3,* 9–44.

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.

Swanson, L. W. (2003). The amygdala and its place in the cerebral hemisphere. *Annals of the New York Academy of Science, 985.*

Takayama, K., & Miura, M. (1991). Glutamate-immunoreactive neurons of the central amygdaloid nucleus projecting to the subretrofacial nucleus of SHR and WKY rats: A double-labeling study. *Neuroscience Letters, 134,* 62–66.

Tobler, P. N., Dickinson, A., & Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience, 23,* 10402–10410.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature, 398,* 704.

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

Wallace, D. M., Magnuson, D. J., & Gray, T. S. (1992). Organization of amygdaloid projections to brainstem dopaminergic, noradrenergic, and adrenergic cell groups in the rat. *Brain Research Bulletin, 28,* 447–454.

Weible, A. P., McEchron, M. D., & Disterhoft, J. F. (2000). Cortical involvement in acquisition and extinction of trace eyeblink conditioning. *Behavioral Neuroscience, 114,* 1058–1067.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record* (pp. 96–104). New York: IRE.

Zucker, R. S., & Regehr, W. G. (2002). Short-term synaptic plasticity. *Annual Review of Physiology, 64,* 355–405.

Appendix
Implementational Details

The model (summarized here) was implemented using the Leabra framework that is described in detail in O'Reilly and Munakata (2000) and in O'Reilly (2001). These same parameters and equations have been used to simulate over 40 different models in O'Reilly and Munakata and in a number of other research models. Thus, the model can be viewed as an instantiation of a systematic modeling framework that uses standardized mechanisms instead of new mechanisms constructed for each model. The model can be obtained by e-mailing Randall C. O'Reilly.

Point Neuron Activation Function

Leabra uses a point neuron activation function that models the electrophysiological properties of real neurons, while simplifying their geometry to a single point. This function is nearly as simple computationally as the standard sigmoidal activation function, but the more biologically based implementation makes it considerably easier to model inhibitory competition, as described below. Further, the use of this function enables cognitive models to be more easily related to more physiologically detailed simulations, thereby facilitating bridge-building between biology and cognition.

The membrane potential V_m is updated as a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\Delta V_m(t) = \tau \sum_c g_c(t)\bar{g}_c(E_c - V_m(t)), \tag{A1}$$

with 3 channels (c), corresponding to excitatory input (e), leak current (l), and inhibitory input (i). Following electrophysiological convention, the overall conductance is decomposed into a time-varying component computed as a function of the dynamic state of the network and a constant $\bar{g}_c$ that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_c) to 1 and the leak and inhibitory driving potentials (E_l and E_i) of 0,

$$V_m^{\infty} = \frac{g_e\bar{g}_e}{g_e\bar{g}_e + g_l\bar{g}_l + g_i\bar{g}_i}, \tag{A2}$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This equilibrium form of the equation can be understood in terms of a

(*Appendix continues*)

Bayesian decision-making framework (O'Reilly & Munakata, 2000).

The excitatory net input/conductance $g_e(t)$ or η_j is computed as the proportion of open excitatory channels as a function of sending activations times the weight values,

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij}. \quad \text{(A3)}$$

The inhibitory conductance is computed via the kWTA function described in the next section, and leak is a constant.

Activation communicated to other cells (y_j) is a thresholded (Θ) sigmoidal function of the membrane potential with the gain parameter (γ),

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma [V_m(t) - \Theta]_+}\right)}, \quad \text{(A4)}$$

where $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $X > 0$. Note that if it returns 0, we assume $y_j(t) = 0$, to avoid dividing by 0. As it is, this function has a very sharp threshold, which interferes with graded learning mechanisms (e.g., gradient descent). To produce a less discontinuous deterministic function with a softer threshold, the function is convolved with a Gaussian noise kernel ($\mu = 0$, $\sigma = .005$), which reflects the intrinsic processing noise of biological neurons,

$$y_j^*(x) = \int \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/(2\sigma^2)} y_j(z - x) dz, \quad \text{(A5)}$$

where x represents the $[V_m - \Theta]_+$ value, and y_j^* is the noise-convolved activation for that value. In the simulation, this function is implemented using a numerical lookup table.

k-Winners-Take-All Inhibition

Leabra uses a k-Winners-Take-All (kWTA) function to achieve inhibitory competition among units within a layer (area). The kWTA function computes a uniform level of inhibitory current for all units in the layer such that the $k + 1$th most excited unit within a layer is generally below its firing threshold, whereas the kth is typically above threshold. Activation dynamics similar to those produced by the kWTA function have been shown to result from simulated inhibitory interneurons that project both feedforward and feedback inhibition (O'Reilly & Munakata, 2000). Thus, although the kWTA function is somewhat biologically implausible in its implementation (e.g., requiring global information about activation states and using sorting mechanisms), it provides a computationally effective approximation to biologically plausible inhibitory dynamics.

kWTA is computed via a uniform level of inhibitory current for all units in the layer as follows:

$$g_i = g_{k+1}^\Theta + q(g_k^\Theta - g_{k+1}^\Theta), \quad \text{(A6)}$$

where $0 < q < 1$ (.25 default used here) is a parameter for setting the inhibition between the upper bound of g_k^Θ and the lower bound of g_{k+1}^Θ. These boundary inhibition values are computed as a function of the level of inhibition necessary to keep a unit right at threshold,

$$g_i^\Theta = \frac{g_e^* \bar{g}_e (E_e - \Theta) + g_l \bar{g}_l (E_l - \Theta)}{\Theta - E_i}, \quad \text{(A7)}$$

where g_e^* is the excitatory net input without the bias weight contribution—this allows the bias weights to override the kWTA constraint.

In the basic version of the kWTA function, which is relatively rigid about the kWTA constraint and is therefore used for output layers, g_k^Θ and g_{k+1}^Θ are set to the threshold inhibition value for the kth and $k + 1$th most excited units, respectively. Thus, the inhibition is placed exactly to allow k units to be above threshold, and the remainder below threshold. For this version, the q parameter is almost always .25, allowing the kth unit to be sufficiently above the inhibitory threshold.

In the average-based kWTA version, g_k^Θ is the average g_i^Θ value for the top k most excited units, and g_{k+1}^Θ is the average of g_i^Θ for the remaining n–k units. This version allows for more flexibility in the actual number of units active depending on the nature of the activation distribution in the layer and the value of the q parameter (which is typically .60) and is therefore used for hidden layers.

PVLV Equations

The PVLV value layers use standard Leabra activation and kWTA dynamics as described above, with the following modifications. They have a three-unit distributed representation of the scalar values they encode, where the units have preferred values of (0, .50, 1). The overall value represented by the layer is the weighted average of the unit's activation times its preferred value, and this decoded average is displayed visually in the first unit in the layer. The activation function of these units is a "noisy" linear function (i.e., without the $x/(x+1)$ nonlinearity, to produce a linear value representation, but that is still convolved with Gaussian noise to soften the threshold, as for the standard units, Equation A5), with gain $\gamma = 220$, noise variance $\sigma = .01$, and a lower threshold $\Theta = .17$. The k for kWTA (average based) is 1, and the q value is .90 (instead of the default of .60). These values were obtained by optimizing the match for value represented with varying frequencies of 0–1 reinforcement (e.g., the value should be close to .40 when the layer is trained with 40% 1 values and 60% 0 values). Note that having different units for different values, instead of the typical use of a single unit with linear activations, allows much more complex mappings to be learned. For example, units representing high values can have completely different patterns of weights than those encoding low values, whereas a single unit is constrained by virtue of having one set of weights to have a monotonic mapping onto scalar values.

Learning Rules

The PVe layer does not learn, and is always just clamped to reflect any received reward value (r). By default we use a value of 0 to reflect negative feedback, .50 for no feedback, and 1 for positive feedback (the scale is arbitrary). The PVi layer units (y_j) are trained at every point in time to produce an expectation for the amount of reward that will be received at that time. In the minus phase of a given trial, the units settle to a distributed value

representation based on sensory inputs. This results in unit activations y_j^-, and an overall weighted average value across these units denoted PV_i. In the plus phase, the unit activations (y_j^+) are clamped to represent the actual reward r (a.k.a., PV_e). The weights (w_{ij}) into each PVi unit from sending units with plus-phase activations x_i^+, are updated using the delta rule between the two phases of PVi unit activation states:

$$\Delta w_{ij} = \varepsilon(y_j^+ - y_j^-)x_i^+. \tag{A8}$$

This is equivalent to saying that the US/reward drives a pattern of activation over the PVi units, which then learn to activate this pattern based on sensory inputs.

The LVe and LVi layers learn in much the same way as the PVi layer (Equation A8), except that the PV system filters the training of the LV values, such that they only learn from actual reward outcomes (or when reward is expected by the PV system, but is not delivered) and not when no rewards are present or expected. This condition is as follows:

$$PV_{filter} = PV_i < \theta_{min} \vee PV_e < \theta_{min} \vee PV_i > \theta_{max} \vee PV_e > \theta_{max}, \tag{A9}$$

$$\Delta w_i = \begin{cases} \varepsilon(y_j^+ - y_j^-)x_i^+, & \text{if } PV_{filter} \\ 0 & \text{otherwise,} \end{cases} \tag{A10}$$

where θ_{min} is a lower threshold (.20 by default), below which negative feedback is indicated, and θ_{max} is an upper threshold (.80), above which positive feedback is indicated (otherwise, no feedback is indicated). Biologically, this filtering requires that the LV systems be driven directly by primary rewards (which is reasonable, and required by the basic learning rule anyway) and that they learn from DA dips driven by high PVi expectations of reward that are not met. The only difference between the LVe and LVi systems is the learning rate ε, which is .05 for LVe and .001 for LVi. Thus, the inhibitory LVi system serves as a slowly integrating inhibitory cancellation mechanism for the rapidly adapting excitatory LVe system.

The four PV and LV distributed value representations drive the dopamine layer (VTA/SNc) activations in terms of the difference between the excitatory and inhibitory terms for each. Thus, there is a PV delta and an LV delta:

$$\delta_{pv} = PV_e - PV_i \tag{A11}$$

$$\delta_{lv} = LV_e - LV_i. \tag{A12}$$

With the differences in learning rate between LVe (fast) and LVi (slow), the LV delta signal reflects recent deviations from expectations and not the raw expectations themselves, just as the PV delta reflects deviations from expectations about primary reward values. This is essential for learning to converge and stabilize when the network has mastered the task (as the results presented in this article show). We also impose a minimum value on the LVi term of .10, so that there is always some expectation—this ensures that low LVe learned values result in negative deltas.

These two delta signals need to be combined to provide an overall DA delta value, as reflected in the firing of the VTA and SNc units. One sensible way of doing so is to have the PV system dominate at the time of primary rewards, whereas the LV system dominates otherwise, by using the same PV-based filtering as holds in the LV learning rule (Equation A10):

$$\delta = \begin{cases} \delta_{pv}, & \text{if } PV_{filter} \\ \delta_{lv} & \text{otherwise.} \end{cases} \tag{A13}$$

It turns out that a slight variation of this where the LV always contributes works slightly better, and is what is used in this paper:

$$\delta = \delta_{lv} + \begin{cases} \delta_{pv}, & \text{if } PV_{filter} \\ 0 & \text{otherwise.} \end{cases} \tag{A14}$$

Synaptic Depression of LV Weights

The weights into the LV units are subject to synaptic depression, which makes them sensitive to changes in stimulus inputs, and not to static, persistent activations (Abbott et al., 1997). Each incoming weight has an effective weight value w^* that is subject to depression and recovery changes as follows:

$$\Delta w_i^* = R(w_i - w_i^*) - Dx_i w_i, \tag{A15}$$

where R is the recovery parameter, and D is the depression parameter, and w_i is the asymptotic weight value. For simplicity, we compute these changes at the end of every trial instead of in an online manner, using $R = 1$ and $D = 1$, which produces discrete one-trial depression and recovery.

Cortical DA Modulated Learning (Second-Order Conditioning)

For the second-order conditioning model, the overall PVLV DA signal (Equation A13) modulated the cortical units by adding an additional excitatory (for positive DA) or inhibitory (for negative DA) current to the overall membrane potential computation, with the conductance proportional to the DA value. This additional current was only added during the plus phase (second phase of activation settling), and learning was then computed by using the standard contrastive Hebbian learning (CHL) algorithm (a standard part of the Leabra algorithm),

$$\Delta w_{ij} = (x_i^+ y_j^+) - (x_i^- y_j^-). \tag{A16}$$

TD Implementation

Our implementation of the TD algorithm uses a distributed coarse-coded representation of value over 12 units, covering the range between −.50 and 3.50 in increments of .20. Activation of a given value is encoded as a Gaussian bump over adjacent units, such that the activation-weighted average represents the appropriate value. This Gaussian model is necessary for TD because it does not use the same discrete value representations as the PVLV model (i.e., 0, .50, 1), so it needs a representation that is capable of better representing a range of values. As with the PVLV value representations, this mechanism was extensively tested and parameterized to produce accurate value representations across a wide range of conditions. The results show that it works well when reliable signals are available.

Received December 14, 2005
Revision received August 18, 2006
Accepted September 7, 2006 ■