Cerebral Cortex October 2007;17:2443–2452
doi:10.1093/cercor/bhl152
Advance Access publication January 13, 2007

Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling

Eugene M. Izhikevich

The Neurosciences Institute, 10640 John Jay Hopkins Drive, San Diego, CA 92121, USA

In Pavlovian and instrumental conditioning, reward typically comes seconds after reward-triggering actions, creating an explanatory conundrum known as "distal reward problem": How does the brain know what firing patterns of what neurons are responsible for the reward if 1) the patterns are no longer there when the reward arrives and 2) all neurons and synapses are active during the waiting period to the reward? Here, we show how the conundrum is resolved by a model network of cortical spiking neurons with spike-timing–dependent plasticity (STDP) modulated by dopamine (DA). Although STDP is triggered by nearly coincident firing patterns on a millisecond timescale, slow kinetics of subsequent synaptic plasticity is sensitive to changes in the extracellular DA concentration during the critical period of a few seconds. Random firings during the waiting period to the reward do not affect STDP and hence make the network insensitive to the ongoing activity—the key feature that distinguishes our approach from previous theoretical studies, which implicitly assume that the network be quiet during the waiting period or that the patterns be preserved until the reward arrives. This study emphasizes the importance of precise firing patterns in brain dynamics and suggests how a global diffusive reinforcement signal in the form of extracellular DA can selectively influence the right synapses at the right time.

Keywords: classical conditioning, dopamine, instrumental conditioning, reward, simulation, spike-timing–dependent plasticity (STDP)

Introduction

Learning the associations between cues and reward (classical or Pavlovian conditioning) or between cues, actions, and reward (instrumental or operant conditioning) involves reinforcement of neuronal activity by reward or punishments (Pavlov 1927; Hull 1943; Houk, Davis, Beiser 1995; Schultz 1998; Dayan and Abbott 2001). Typically, the reward comes seconds after reward-predicting cues or reward-triggering actions, creating an explanatory conundrum known in the behavioral literature as the "distal reward problem" (Hull 1943) and in the reinforcement learning literature as the "credit assignment problem" (Minsky 1963; Barto et al. 1983; Houk, Adams, Barto 1995; Sutton and Barto 1998; Dayan and Abbott 2001; Worgotter and Porr 2005). Indeed, how does the animal know which of the many cues and actions preceding the reward should be credited for the reward? In neural terms, in which sensory cues and motor actions correspond to neuronal firings, how does the brain know what firing patterns, out of an unlimited repertoire of all possible patterns, are responsible for the reward if the patterns are no longer there when the reward arrives? How does it know which spikes of which neurons result in the reward if many neurons fire during the waiting period to the reward? Finally, how does the common reinforcement signal in the form of the neuromodulator dopamine (DA) (Schultz 1998, 2002; Seamans and Yang 2004; Schultz 2007a, 2007b) influences the right synapses at the right time, if DA is released globally to many synapses? In this paper, we show how the credit assignment problem can be solved in a simulated network of cortical spiking neurons with DA-modulated plasticity.

An important aspect of DA modulation of synaptic plasticity is its enhancement of long-term potentiation (LTP) and long-term depression (LTD): In hippocampus, DA D1-receptor agonists enhance tetanus-induced LTP, but the effect disappears if the agonist arrives at the synapses 15–25 s after the tetanus (Otmakhova and Lisman 1996, p. 7481; see also Impey et al. 1996; Barad et al. 1998), thereby suggesting the existence of a short window of opportunity for the enhancement. LTP in the hippocampal → prefrontal cortex pathway is enhanced by direct application of DA in vivo (Jay et al. 1996) or by burst stimulation of the ventral tegmental area (VTA), which releases DA (Gurden et al. 2000). Correspondingly, D1-receptor antagonists prevent the maintenance of LTP (Frey et al. 1990; Impey et al. 1996), whereas agonists promote it via blocking depotentiation (Otmakhova and Lisman 1998) even when they are applied after the plasticity-triggering stimuli. DA is also shown to enhance tetanus-induced LTD in layer 5 pyramidal neurons of prefrontal cortex (Otani et al. 2003), and it gates corticostriatal LTP and LTD in striatal projection neurons (Choi and Lovinger 1997; Centonze et al. 1999; Calabresi et al. 2000).

Spike-timing–dependent synaptic plasticity (STDP) involves both LTP and LTD of synapses: Firing of a presynaptic neuron immediately before a postsynaptic neuron results in LTP of synaptic transmission, and the reverse order of firing results in LTD, as shown in Figure 1*a*,*b* (Levy and Steward 1983; Markram et al. 1997; Bi and Poo 1998; see also theoretical paper by Gerstner et al. 1996). It is reasonable to assume that the LTP and LTD components of STDP are modulated by DA the same way as they are in the classical LTP and LTD protocols (Houk, Adams, Barto 1995; Seamans and Yang 2004). That is, a particular order of firing induces a synaptic change (positive or negative), which is enhanced if extracellular DA is present during the critical window of a few seconds.

In this article, we show that DA modulation of STDP has a built-in property of instrumental conditioning: It can reinforce firing patterns occurring on a millisecond timescale even when they are followed by reward that is delayed by seconds. This property relies on the existence of slow synaptic processes that act as "synaptic eligibility traces" (Klopf 1982; Sutton and Barto 1998) or "synaptic tags" (Frey and Morris 1997). These processes are triggered by nearly coincident spiking patterns, but due to a short temporal window of STDP, they are not affected by random firing during the waiting period to the reward. To

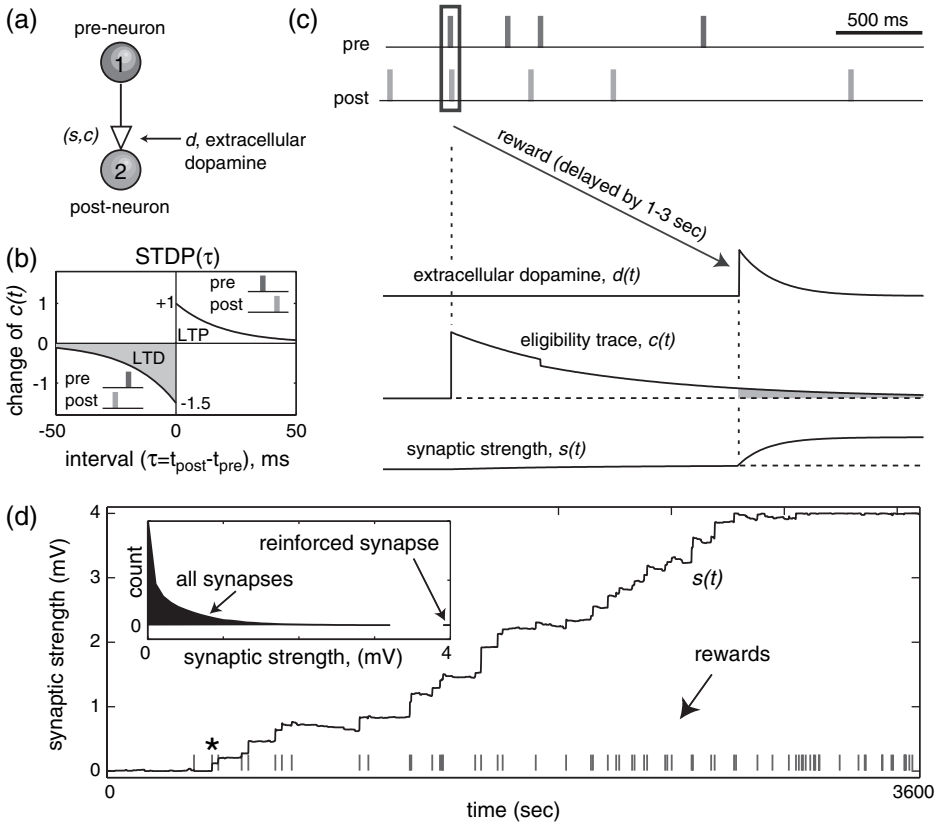

Figure 1. Instrumental conditioning of a synapse. (*a*) The dynamics of each synapse is described by 2 phenomenological variables governed by equations (1) and (2): synapse strength *s* and eligibility trace *c*, which are gated by the extracellular DA *d*. Firings of the pre- and postsynaptic neurons induce changes to the variable *c* according to the STDP rule (shown in *b*). These changes result in modification of the synaptic strength, *s*, only when extracellular DA is present ($d > 0$) during the critical window of a few seconds while the eligibility trace *c* decays to zero. (*c*) The magnification of the region in (*d*) marked by *. To reinforce coincident firings of 2 coupled neurons, we deliver a reward (step-increase of variable *d*) with a random delay (between 1 and 3 s) each time a postsynaptic firing occurs within 10 ms after a presynaptic firing (marked by a rectangle in *c*). This rare event increases *c* greater than any random firings of the same neurons during the delayed period. (*d*) Consistent rewarding of each such event results in the gradual increase of synaptic strength, *s*, which increases the probability of coincident firings and brings even more reward. The time course of a typical unreinforced synapse (not shown here) looks like a random walk near 0. The inset shows the distribution of all synaptic weights in the network. The reinforced synapse is potentiated to the maximal allowable value 4 mV (42 out of 50 experiments) whereas the other synapses are not.

illustrate this point, consider 2 neurons, each firing 1 spike per second, which is comparable to the spontaneous firing rate of neocortical pyramidal neurons (all layers: less than 1 Hz and often less than 0.1 Hz, layer 5: 4.1 Hz; Swadlow 1990, 1994). A nearly coincident firing will trigger STDP and change the synaptic tag. However, the probability that subsequent random spikes with the same firing frequency will fall within 50 ms of each other to trigger more STDP and alter the synaptic tag is quite small—on average once per 20 s (we elaborate this point in Reinforcing a Synapse). This "insensitivity" of the synaptic tags to the random ongoing activity during the waiting period is the key feature that distinguish our approach from previous studies (see e.g., Houk, Adams, Barto 1995; Seung 2003), which requires that the network be quiet during the waiting period or that the patterns are preserved as a sustained response (Drew and Abbott 2006). In this paper, we show how DA-modulated STDP can selectively reinforce precise spike-timing patterns that consistently precede the reward and ignore the other firings that do not cause the reward. At the end of the article, we discuss why this mechanism works only when precise firing patterns are embedded into the sea of noise and why it fails in the mean firing rate models.

We also present a spiking network implementation of the most important aspect of the temporal difference (TD) reinforcement learning rule (Sutton 1988)—the shift of reward-triggered release of DA from unconditional stimuli (US) to reward-predicting conditional stimuli (CS) (Ljungberg et al. 1992; Montague et al. 1996; Schultz et al. 1997; Schultz 1998, 2002, 2006b; Pan et al. 2005). Our simulations demonstrate how DA modulation of STDP could play an important role in the reward circuitry and solve the distal reward problem.

Materials and Methods

Because details of the kinetics of the intracellular processes triggered by STDP and DA are unknown, we suggest the simplest phenomenological model that captures the essence of DA modulation of STDP. Following Izhikevich et al. (2004), we describe the state of each synapse using 2 phenomenological variables (Fig. 1*a*): synaptic strength/weight, *s*, and activation of an enzyme important for plasticity, *c*, for example, autophosphorylation of CaMK-II (Lisman 1989), oxidation of PKC or PKA, or some other relatively slow process acting as a synaptic tag

$$\dot{c} = -c/\tau_c + \mathrm{STDP}(\tau)\delta(t - t_{\mathrm{pre/post}}), \quad (1)$$

$$\dot{s} = cd. \quad (2)$$

Here and below, *d* describes the extracellular concentration of DA and $\delta(t)$ is the Dirac delta function that step-increases the variable *c*. Firings of pre- and postsynaptic neurons, occurring at times $t_{\mathrm{pre/post}}$, respectively, change *c* by the amount STDP(τ) depicted in Figure 1*b*,

where $\tau = t_{post} - t_{pre}$ is the interspike interval. This variable decays to $c = 0$ exponentially with the time constant $\tau_c = 1$ s, as in Figure 1*c*. The decay rate controls the sensitivity of plasticity to delayed reward. Indeed, *c* acts as the "eligibility trace" for synaptic modification (Houk, Adams, Barto 1995) because it allows change of the synaptic strength *s* via equation (2) gated by *d*. A better description of the decay of the synaptic tag *c* may be provided by detailed biophysical/kinetic models. Notice that the decay of eligibility trace is relatively fast, so that the effect of DA is negligible 5 s after the STDP-triggered event, which is consistent with the experimental results of Otmakhova and Lisman (1996) who observed no effect when DA was delivered 15–25 s after induction of plasticity.

The model integrates, in a biophysically plausible way, the millisecond timescale of spike interactions in synapse-specific STDP with the slow trace modulated by global reward signal corresponding to the behavioral timescale. There is no direct experimental evidence for or against our model; thus, the model makes a testable prediction, rather than a postdiction, on the action of DA on STDP based on purely theoretical considerations.

The variable *d* describes the concentration (μM) of extracellular DA, and it is the same for all synapses in our model (whereas *c* and *s* are different for different synapses). We assume that

$$\dot{d} = -d/\tau_d + \mathrm{DA}(t) \quad , \tag{3}$$

where τ_d is the time constant of DA uptake and $\mathrm{DA}(t)$ models the source of DA due to the activity of dopaminergic neurons in the midbrain structures VTA and substantia nigra pars compacta. A better description of DA kinetics, based on Michaelis–Menten formalism, was recently suggested by Montague et al. (2004).

In the simulations below, we take $\tau_d = 0.2$ s, which is greater than the experimentally measured time constant of DA uptake in striatum (around 0.1 s, Wightman and Zimmerman 1990; Garris et al. 1994) but smaller than that in the prefrontal cortex (seconds, see Cass and Gerhardt 1995). We take tonic source of DA to be $\mathrm{DA}(t) = 0.01$ μM/s so that the baseline (tonic) concentration of DA is 2 nM as measured by microdialysis in the striatum and prefrontal cortex (Seamans and Yang 2004, p. 31). We simulate the delivery of the reward in Figure 1*c* as a burst of activity of dopaminergic neurons which step-increases the concentration of DA by 0.5 μM (i.e., $\mathrm{DA}(t) = 0.5\delta(t - t_{rew})$ of μM/s at the moment of reward t_{rew}), which is in the range measured in by Garris et al. (1994). Because the tonic level of DA is much lower than the phasic level during the reward, no significant modification of synaptic strength occurs ($d \approx 0$) unless the reward is delivered (*d* is large). In Figure 4, we use $\mathrm{DA}(t) = 0.004\delta(t)$ μM/s for each spike fired by the neurons in group $\mathrm{VTA_p}$. A possible extension of equations (1) and (2) is to consider a vector of synaptic tags corresponding to a cascade of processes (Fusi et al. 2005). In this case, the STDP-triggered increase of the synaptic eligibility trace would not be instantaneous, as in Figure 1*c*. Instead, it would slowly increase and then decrease, like the synaptic alpha function but on a longer timescale. The slow increase would create a "refractory period" corresponding to the insensitivity to reward that comes too early.

All simulations are carried out using a network of 1000 spiking neurons described in detail by Izhikevich (2006), who provides the MATLAB and C code. The code is also available on the author's web page www.izhikevich.com. The network has 80% excitatory neurons of the regular spiking type and 20% inhibitory neurons of the fast spiking type (Connors and Gutnick 1990), representing the layer 2/3 part of a cortical minicolumn. Neurons are randomly connected with 10% probability so that there are 100 synapses per averaged neuron. The connections, combined with the random input simulating noisy miniature PSPs, make neurons fire Poisson-like spike trains with an average frequency around 1 Hz. This low frequency of firing is important for the "low probability" of sequential spikes to fall within the STDP time window by chance (the firing rate in neocortical layer 2/3 is much less than 1 Hz, Swadlow 1990, 1994). The maximal axonal conduction delay is taken to be 1 ms. Each excitatory synapse is modified according to equations (1) and (2) with STDP depicted in Figure 1*b*, but the weights are limited to the range 0 to 4 mV (i.e., clipped at 0 and 4 mV). Both excitatory-to-excitatory and excitatory-to-inhibitory synaptic connections are subject to the same STDP rule. One could use a different, more physiological STDP rule for the latter, or even keep them fixed (nonplastic). Our choice was done for the sake of simplicity and to be consistent with previous implementations of the spiking model (Izhikevich 2006). Inhibitory synapses are not plastic in the model. The LTD area in Figure 1*b* is 50% greater than the LTP area so that uncorrelated firing of any 2 neurons results in the decrease of synaptic strength between them (Kempter et al. 1999a, 1999b; Song et al. 2000). As a result of spontaneous activity, the strengths of excitatory synapses in the network converge to the exponential distribution depicted in the inset in Figure 1*d*. Notice that all synapses are much weaker than the maximal allowable value of 4 mV, and the majority is less than 0.1 mV.

Results

Below we use the spiking network of 1000 cortical neurons with DA-modulated STDP to illustrate various aspects of reinforcement of precise firing patterns embedded into the sea of noise.

Reinforcing a Synapse

In Figure 1, we reinforce contingent firing of 2 neurons by delayed reward to illustrate how DA-modulated STDP addresses the credit assignment problem on the synaptic level. This experiment is motivated by the in vivo monkey experiment of Ahissar et al. (1992). The experiment might look artificial in the context of animal learning, but it explains the mechanism responsible for reinforcement of more complicated spiking patterns, as we demonstrate later.

In the network of 1000 neurons and 100 000 synaptic interconnections, we randomly choose a synapse that connects 2 excitatory neurons, as in Figure 1*a*, and manually set its synaptic strength to zero ($s = 0$). The firing rate in the network is around 1 Hz, so every few minutes the postsynaptic neuron fires by chance within 10 ms after the presynaptic neuron. Every time such an event occurs, marked by the blue rectangle in Figure 1*c*, we deliver the reward to the network in the form of a spike of extracellular DA with a random delay between 1 and 3 s. Because the delivery of the reward depends on what the network is doing, the experiment in the figure could be interpreted as the simplest form of instrumental conditioning (Dayan and Abbott 2001).

In Figure 1*d*, we plot the strength of the synapse (curve) and the moments the reward are delivered (bars). At the beginning, the network receives unexpected reward every few minutes, but it "does not know" what causes the reward or when. Because of the delay to the reward, all neurons fire and all synapses are activated during the waiting period to the reward (in contrast to previous models), and all synapses receive the same amount of reward (variable *d*). As in instrumental conditioning, the system has to determine on its own what patterns of spiking bring the reward and how to reinforce them.

Each delivery of the reward potentiates the chosen synapse and brings it closer to the maximal allowable value of 4 mV. On average, the probability (frequency) of reward triples, and the chosen synapse quickly reaches the maximal allowable value of 4 mV. Other synapses change as well, but none reach 4 mV. The distribution of all synaptic weights, depicted in the inset in Figure 1*d*, remains relatively unchanged. To test the robustness of this phenomenon, we ran 50 simulated experiments, each with a randomly chosen synapse and schedule of reward delays. In 42 out of 50 experiments, the chosen synapse reached the maximal allowable value within 1-h period, requiring only 40 ± 8 reward.

Why is the "chosen" synapse consistently potentiated, but the other 79 999 excitatory synapses are not? (Only excitatory synapses are plastic in the model). Nearly, coincident pre-then-post firing of the 2 neurons in Figure 1*c* increases the value of the variable *c*, which acts as the eligibility trace (synaptic tag) for the modification of the synapse. The subsequent non-coincident firings of the 2 neurons perturb *c* slightly because the function STDP(τ) in Figure 1*b* is small for large interspike intervals τ. As a result, *c* has a residual positive value when the delayed reward arrives, so the synaptic strength *s* increases in proportion to *cd* (see Materials and Methods). Of course, a nearly coincident firing of the 2 neurons with the reverse order (post-then-pre) during the waiting period could make *c* negative, resulting in the decrease of *s* when the reward arrives, but the probability of such an adverse event during the waiting period is quite small (because the firing rate is small). Naturally, there are many other pairs of neurons that fire nearly coincident spikes by chance just before the reward, so the corresponding synapses are also modified. However, the order of firing of these neurons is random, so after many firings, the positive and negative modifications cancel each other out, resulting in a net decrease of the synaptic weight (because the LTD area of the STDP curve is larger than the LTP area). As a result, across many trials, each reward consistently potentiates only the chosen synapse and increases the cross-correlation between the pre- and postsynaptic neurons, thereby bringing more reward.

Classical (Pavlovian) Conditioning

In Figure 2*a*, we illustrate a classical (Pavlovian) conditioning experiment: Rewarding a CS (S_1) embedded into a continuous stream of a large number of irrelevant but equally salient stimuli. To simulate the experiment, we choose 100 random sets, S_1, $S_2, \ldots, S_{100}$, of 50 neurons each to represent 100 random stimuli. To deliver a stimulus, say S_1, we stimulate all 50 neurons in the set S_1 with a 1-ms pulse of superthreshold current. The nearly coincident firing of neurons in S_1 reveals itself as a vertical strip in Figure 2*b*. The precise firing pattern is clearly seen only when activities of all neurons are plotted, but it cannot be seen in the activity of any individual neuron, because the spike evoked by stimulus S_1 is not different from any other spike of the neuron. Next, we form a continuous input stream consisting of stimuli S_k ($1 \leqslant k \leqslant 100$) in the random order with a random

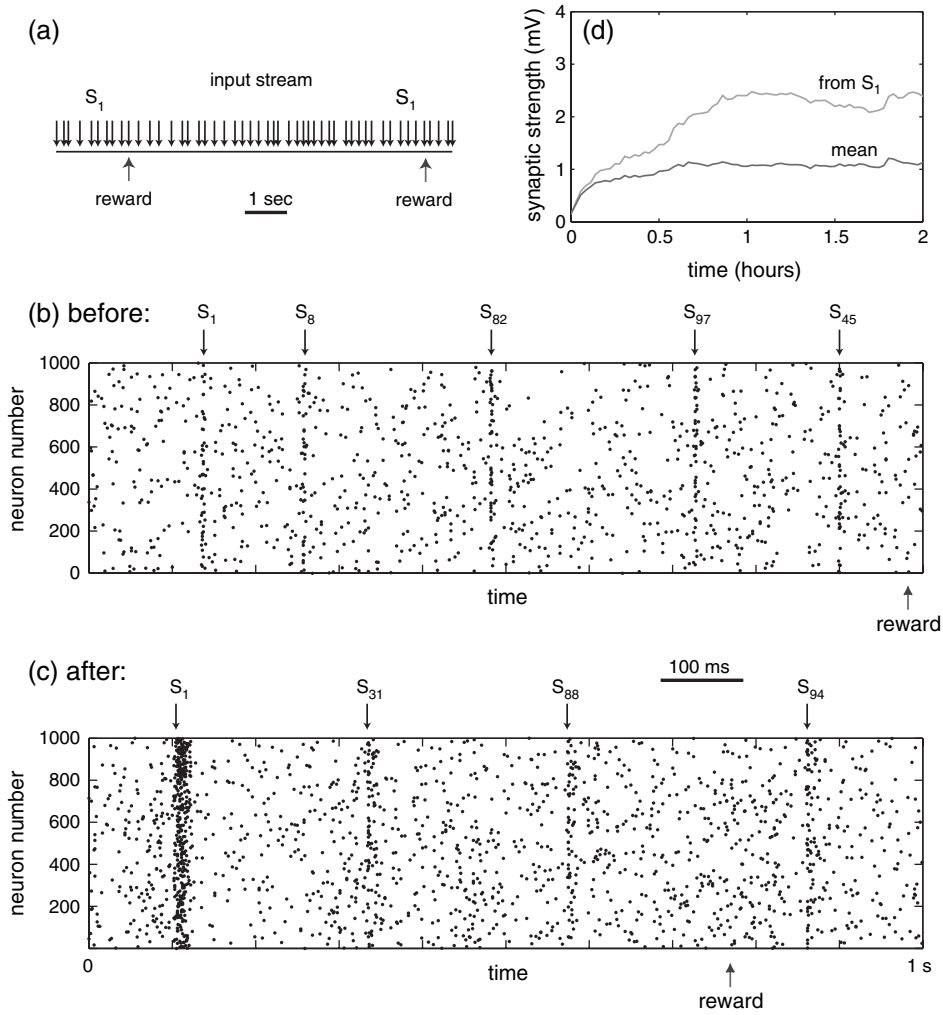

Figure 2. Classical (Pavlovian) conditioning. We select 100 groups, $S_1, S_2, \ldots, S_{100}$, of 50 randomly chosen neurons to represent 100 different input stimuli. To present a stimulus, we fire the corresponding neurons (by injecting a superthreshold current). (*a*) The network receives a continuous stream of 5 stimuli per second (on average). Every time S_1 appears in the stream (on average once per 20 s), the network receives DA reward with a random delay up to 1 s. Response of the network to stimulation at the beginning of the experiment (*b*) and after 1 h (*c*). Notice the enhanced response to S_1. (*d*) The mean strength of excitatory synapses outgoing from the neurons in S_1 increases greater than the mean excitatory synaptic strength in the rest of the network, resulting in stronger network response to S_1.

interstimulus intervals between 100 ms and 300 ms, that is, on average 5 stimuli per second. We treat S_1 as our CS and the other stimuli as distracters. For every occurrence of S_1, we deliver a reward in the form of the increase of extracellular DA with a random delay of up to 1 s, as in Figure 2*a*. The delay is large enough to allow many neurons in the network to fire a spike and to allow a few irrelevant stimuli during the waiting period, as in Figure 2*b*. Thus, the network receives reward on average every 20 s caused by an unknown (to the network) firing pattern embedded into the sea of random spikes and distracters.

At the beginning of the experiment, depicted in Figure 2*b*, all stimuli have equal salience in the sense that they evoke coincident firing of 50 stimulated neurons. However, after a hundred of CS-reward parings, that is, within the first hour, the response of the network to the CS S_1 becomes reinforced, as indicated by the thick vertical strip in Figure 2*c*. In Figure 2*d*, we show that the averaged strength of excitatory synaptic connection from neurons in S_1 becomes much stronger than the mean excitatory synaptic connection in the rest of the network. That is, neurons in S_1 can strongly influence their postsynaptic targets, or, in other words, the other neurons in the network listen more closely to neurons in S_1. The other neurons may contain motor neurons that trigger a conditional response. In this case, S_1 would trigger the response more often and stronger than any other stimulus S_k. The conditional response could be a simple movement in the anticipation of the reward or a learned motor response, as in the instrumental (operant) conditioning discussed in the next section. The other neurons may also contain neurons projecting to the midbrain dopaminergic neurons, as we discuss in Shift of DA Response from US to Reward-Predicting CS in Classical Conditioning. In this case, presentation of the CS S_1 would trigger more DA release than presentation of any other stimulus S_k, that is, S_1 would acquire a rewarding value.

How can the network select and reinforce a single firing pattern in the presence of noise and irrelevant patterns, especially in the view that the reward comes with a delay? Presentation of every stimulus S_k fires 50 neurons, which sends spikes to other neurons in the network, possibly firing a few of them by chance. Because of the order pre-then-post, the synaptic connections from the 50 neurons to the fired neurons become eligible for potentiation, that is, the corresponding tags c_{ij} increase. If no DA reward is delivered within a critical period after this event, the synaptic tags c_{ij} decay to zero, resulting in small overall potentiation (due to the tonic level of DA) which is counterbalanced by depression (due to random spikes and the fact that LTD window of STDP is greater than the LTP window). However, if DA reward comes within the critical period after the stimulation, the synapses are potentiated according to the mechanism depicted in Figure 1*c*. The stronger the synapses, the more excitation follows S_1, the more postsynaptic targets fire, leading to even greater potentiation of synapses from neurons representing the CS S_1.

Stimulus–Response Instrumental Conditioning

In Figure 3, we simulate a typical instrumental conditioning experiment: we reinforce a network of 1000 cortical spiking neurons to produce an appropriate motor response to a stimulus. First, we choose a random group of 50 neurons, called S, that represents the input stimulus to the network. We choose 2 random nonoverlapping groups of 50 excitatory neurons each, called A and B, that give rise to 2 motor responses of the network. To deliver the stimulus, we inject a strong 1-ms pulse of current into the neurons in S to make them fire, as in Figure 3*a* (the 2- to 3-ms delay is due to the spike upstroke). Their coincident firing typically evokes a few spikes in the other neurons in the network. During a 20-ms time window after the stimulation, we count the number of spikes fired by neurons in A and B, denoted as |A| and |B|, respectively. We say that the network exhibits response A when |A| > |B|, response B when |B| > |A|, and no response otherwise (e.g., when |B| = |A| = 0, a stronger requirement, e.g., |A| > 2|B| for response A, would still be effective, but it takes longer time to reinforce). One might think of neurons in groups A and B as projecting to 2 motor areas that innervate 2 antagonistic muscles; to produce a noticeable movement, one group has to fire more spikes than the other group.

The simulated experiment consists of trials separated by 10 s. In each trial, illustrated in Figure 3*a*, we deliver stimulation to neurons in S and monitor the response of the network. If the response is A (more spikes in group A than in group B), we deliver a reward in the form of the increase of extracellular DA with a delay of up to 1 s (the delay is inversely proportional to the ratio |A|/|B|, so that greater ratios result in faster movements and earlier reward). During the first few trials, the probability of response A is the same as that of B, see Figure 3*b*, but then it quickly increases to nearly 80% in less than 100 trials. As a control, after the first 400 trials, we start to reward response B. The probability of response A decreases while that of B increases, and the network switches its behavior after less than 50 trials. We repeat this experiment 20 times, selecting random sets S, A, and B. The network learns the correct response all 20 times. The only variability was the number of trials needed to reach the 80% correct probability of responses. Increasing the learning rate can decrease the number of required trials to just a few—consistent with animal experiments (Pasupathy and Miller 2005). However, the small size of the network would make the network responses less reliable (noisier) in this case.

The number of spikes fired by neurons in A and B depends on the strength of synaptic connections from S to A and B. Rewarding response A reinforces connections to A, as one can see in Figure 3*c*, according to the same mechanism as described in Figure 1 for a pair of neurons. Interestingly, it also reinforces connections from S to B (because there is no winner-take-all competition between neurons in A and B), as well as connections from S to any other neuron in the network (as in Fig. 2), though to a lesser degree. Indeed, if a neuron in B starts to fire in response to the stimulation, but there are still more spikes in A, the reward still comes and the connections from S to that neuron in B are potentiated. This may continue as long as |A| > |B|. A possible behavioral interpretation of this effect is that the network generalizes that "reward are delivered in response to stimulation S." Conversely, rewarding B after 400 trials makes connections S → B stronger than connections S → A. One could further enhance the contrast between the synaptic connections to A and B (and improve the percentage of correct choices) via anatomical constraints, such as stronger winner-take-all lateral inhibition. In this paper, we keep the anatomy simple (all-to-all with 10% connectivity) to emphasize the role of DA modulation of STDP over any other mechanism.

Notice that a simple combinatorial consideration shows that there are more than 10^{164} different choices of 2 groups of 50 neurons out of 800 excitatory neurons. The network neither knows the identity of neurons in A and B nor does it know the

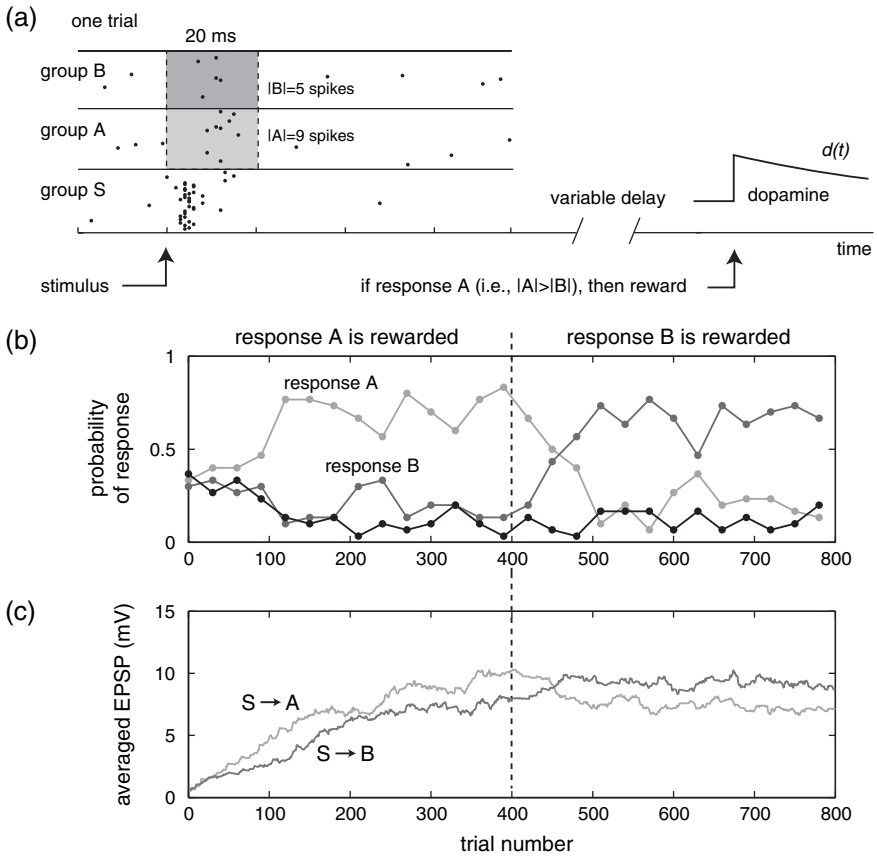

Figure 3. Instrumental conditioning. (*a*) S, A, and B are 3 groups of 50 randomly chosen neurons (out of 1000) that correspond to the representation of an input stimulus and 2 (nonantagonistic) motor responses, respectively. Each trial consists of the presentation of a stimulus that fires neurons in group S (simulated by injecting a superthreshold current). Spikes in groups A and B are counted during the 20-ms window after the stimulation. The network is said to produce response A if group A fires more spikes than group B, and vice versa. Rewarding response A is simulated by the increase of the extracellular concentration of DA *d* (as in Fig. 1*b*) with a delay of up to 1 s. In (*b*) and (*c*), the first 400 trials delivered every 10 s reinforce response A. The higher probability of response A to stimulation S results from the increase of the averaged strength of synaptic input from neurons S to neurons A. The relationship reverses when response B is rewarded (trials 401–800). Notice that while the network does not know the identity of neurons in A and B, or what is rewarded, the appropriate stimulus–response relationship is reinforced.

rules of the game or the schedule of the reward. It receives a seemingly arbitrary sequence of rewards, and it determines on its own what brings the reward and what it must do to increase the frequency of the reward.

Shift of DA Response from US to Reward-Predicting CS in Classical Conditioning

In Figure 4, we reproduce the basic phenomenology of shifting the release of DA in response to US to an earlier reward-predicting CS (Ljungberg et al. 1992; Schultz et al. 1997; Schultz 1998, 2002, 2006b; Pan et al. 2005). The shift occurs automatically when VTA-projecting neurons are part of the whole network and the synapses onto these neurons are subject to the same DA-modulated STDP. Demonstrating the shift is the first step toward a spiking network implementation of the TD error signal (Sutton 1988; Houk, Adams, Barto 1995; Montague et al. 1996; Sutton and Barto 1998; Pan et al. 2005). The full spiking implementation of TD would require modeling the looping anatomy of striatum and basal ganglia, which is outside the scope of this article.

First, we choose a random group of 100 excitatory neurons and assume that this group, called VTA_p, represents cortical projections to the VTA of the midbrain (Au-Young et al. 1999). We use different fonts: VTA for the area in midbrain and VTA_p for the group of neurons projecting to VTA (subscript "p" stands for "projecting"). Thus, we assume that the midbrain activity, and hence the amount of DA released into the network, is proportional to the firing rate of the neurons in this group. Next, we choose a random group of excitatory neurons, called *US*, that represents the US, and 2 groups, CS_1 and CS_2 that represent 2 CS; see Figure 4*a*.

To simulate the prior association between the US and the release of DA, we reset the weights of synaptic connections from the *US* group to the VTA_p group (projecting to VTA) to the maximal allowable values. (We could have achieved that by repeating the classical conditioning experiment in Fig. 2 with S_1 being the US). Thus, stimulating neurons in the *US* group would result in a strong response in the VTA-projecting neurons VTA_p, and hence would release DA, whereas stimulating any other random group of neurons would not result in significant response of the VTA_p. This is the only difference between the *US* group and the other neurons in the network. (Apparently, there are multiple pathways from US-triggered activity in the brain to the VTA; here we consider only one, cortical pathway).

During the first 100 trials, where each trial is separated by 10–30 s, we deliver the US (but not CS), that is, we inject a superthreshold current into the *US* group of neurons. Because

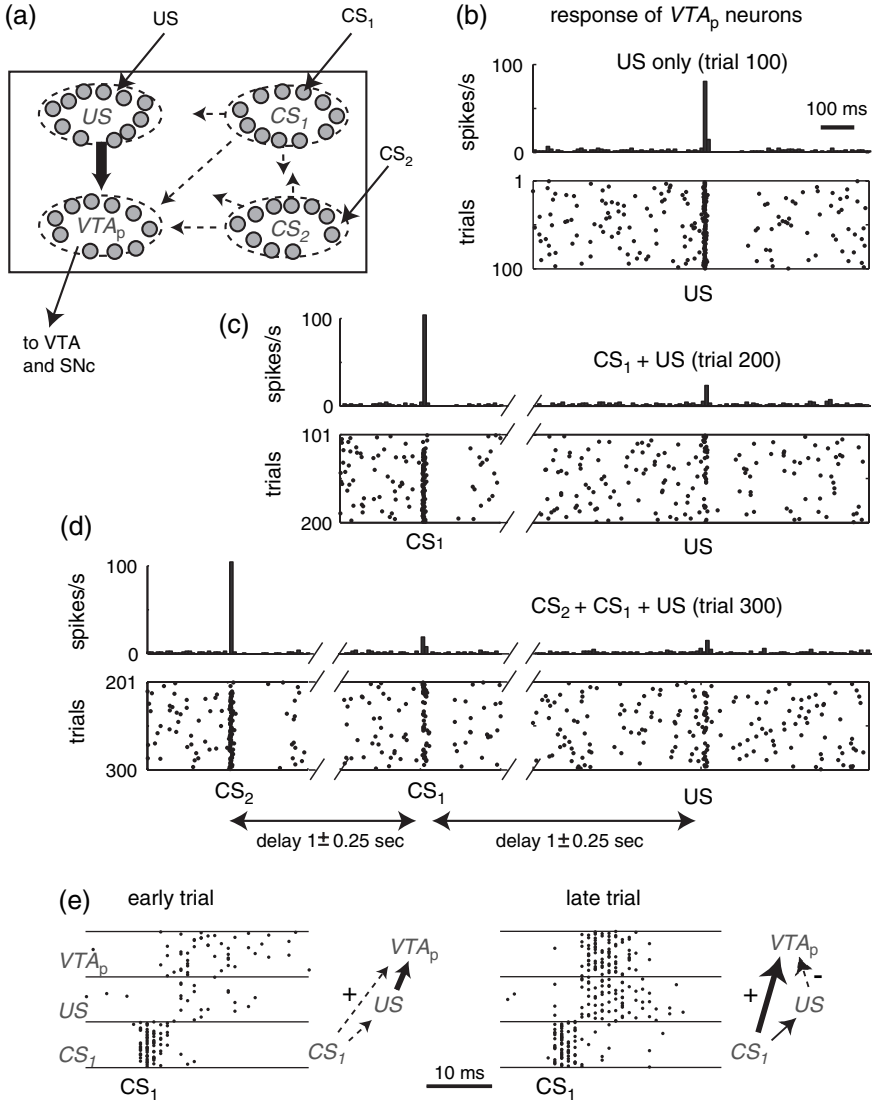

Figure 4. Spiking neuron implementation of shift of DA response from US to reward-predicting CS. (*a*) Four random groups of 100 neurons, *US*, CS_1, CS_2, VTA_p, are chosen to represent, respectively, the US, CS_1, CS_2, and neurons projecting (possibly indirectly) to the midbrain structures responsible for the release of extracellular DA (VTA and substantia nigra pars compacta). Each firing from the group VTA_p increases the extracellular concentration of DA. Initially, the synaptic connections from *US* to VTA_p are set to maximal values, reflecting the acquired association between US and reward. (*b*), (*c*), and (*d*) Spiking responses of a typical neuron in the VTA_p group (spike raster) in a series of trials separated by a pseudorandom interval of 10–30 s and the combined activity of the entire group (histogram) to the last trial in the series. (*b*) Trials 1–100: presentation of the US (brief injection of superthreshold current to the neurons in group *US*) evokes a reliable response in VTA_p neurons. (*c*) Trials 101–200: CS_1 consistently precedes US by 1 ± 0.25 s. The VTA_p response shifts to CS_1. (*d*) Trials 201–300: CS_2 precedes CS_1, which precedes US. The response shifts to the earlier stimulus, CS_2. (*e*) Explanation of the mechanism of the shift from US to CS_1. Shown are spikes in the VTA_p, *US*, and CS_1 groups in response to presentation of CS_1; see text for details.

of the strong initial projections from *US* to VTA_p group, the stimulation evokes a reliable response in the VTA_p group resulting in the elevation of extracellular DA and maintaining (reinforcing) the projections (indeed, due to the spontaneous release of DA, synapses are plastic all the time and may depress because STDP is dominated by LTD). The histogram in Figure 4*b* shows the response of the entire VTA_p group on the last trial, and the spike raster shows a typical response of a single neuron in the group in 100 consecutive trials, which is similar to the in vivo–recorded responses of midbrain neurons to unexpected reward, novel, and salient stimuli (Schultz 1998, 2002, 2006b).

During trials 101–200, we stimulate neurons in the group CS_1, then we stimulate neurons in the group *US* with a random delay 1 ± 0.25 s. As one can see, in Figure 4*c*, the VTA_p neuron starts to fire in response to the reward-predicting CS_1 just after a few trials, as was observed in vivo in monkeys and rats (see Schultz 2002; Pan et al. 2005, who showed that the switch occurs after just a few pairings). The response of the neuron to the US slowly decreases, so the response of the entire VTA_p group to the last trial (histogram in Fig. 4*c*) is diminished. During trials 201–300, we present CS_2 1 ± 0.25 s prior to CS_1, which is 1 ± 0.25 s prior to US. As one can see, in Figure 4*d*, the response of the neuron switches to the earliest reward-predicting stimulus, CS_2, though there is still some response to CS_1 and the US, again, consistent with in vivo work (Pan et al. 2005).

The mechanism of switching of the response from the US to the earlier CS relies on the sensitivity of STDP to the order of firings occurring within tens of milliseconds (despite the fact

that CS and US are separated by a second). Due to the random connections in the network, stimulation of CS_1 group of neurons causes some neurons in the *US* group to fire, which in turn causes some neurons in the VTA_p group to fire; see Figure 4*e*, earlier trial. In essence, presentation of the CS triggers the reactivation of the activity chain leading to the reward, CS_1-then-US-then-VTA, but on a compressed timescale. This property emerged in the spiking network spontaneously. Due to the same mechanism as in Figure 1, the order of firing CS_1-then-VTA, and the subsequent release of DA due to the presentation of the US, potentiates the direct synaptic projections $CS \rightarrow VTA_p$, resulting in the increased response to the CS_1 seen in Figure 4*c*, left. After many trials, neurons in group VTA_p can fire in response to firings of CS_1 neurons alone, simultaneously or often before they receive spikes from the *US* neurons, as in Figure 4*e*, late trial. As a result of jittered and often inverse order of firing, VTA_p-then-*US*, and the fact that LTD part of STDP in Figure 1*b* is dominant over the LTP part, the synaptic projections $US \rightarrow VTA_p$ depress, resulting in the decreased (unlearned) response to the US seen in Figure 4*c*, right. The same mechanism is responsible for the switching of response from CS_1 to CS_2 in Figure 4*d*. Again, this property appears spontaneously in a randomly connected network of spiking neurons with STDP.

Discussion

We present a biologically realistic implementation of Pavlovian and instrumental conditioning and some aspects of TD reinforcement learning using a spiking network with DA-modulated STDP. Based on experimental evidence that DA modulates classical LTP and LTD, we assume that DA has a permissive, enabling effect allowing STDP to take place—a testable assumption that has never been suggested before. Although STDP acts on a millisecond timescale, the slow biochemical kinetics of synaptic plasticity could make it sensitive to DA reward delayed by seconds. We interpret the spiking network as representing a small part of prefrontal cortex receiving numerous dopaminergic projections from the midbrain and projecting to the midbrain (Seamans and Yang 2004), though the theory can be applied to neostriatum and basal ganglia as well. Our simulations suggest a neurally plausible mechanism of how associations between cues, actions, and delayed reward are learned (Figs 1–3), as well as how DA response shifts from US to reward-predicting CS (Fig. 4).

Spiking Implementation of Reinforcement Learning

Spiking implementation of reinforcement learning has been suggested by Seung (2003), Hasselmo (2005), and Koene and Hasselmo (2005), and there are many more models based on synaptic eligibility traces (see e.g., Houk, Davis, Beiser 1995). All these models have one common drawback: they require the network to be relatively quiet during the waiting period to the reward. Indeed, random neuronal activity during the waiting period triggers synaptic transmission in all synapses, alters the eligibility traces, and impedes learning. In contrast, STDP is insensitive to random firings during the waiting period but sensitive only to precise firing patterns. Because the set of precise patterns is sparse in the space of all possible firing patterns, DA-modulated STDP takes advantage of this fact and renders a superior mechanism of reinforcement learning.

Rao and Sejnowski (2001) consider explicitly the relationship between STDP and TD, but they asked the opposite question: how to get STDP from TD acting on a millisecond timescale and how the resulting STDP depends on the dendritic location?

Synaptic Eligibility Traces

The slow kinetics of synaptic plasticity, modeled by the variable c (see eq. 1), results in the existence of synaptic eligibility traces (Houk, Adams, Barto 1995). This is an old idea in the classical machine learning algorithms, where eligibility traces are assigned to cues and actions, as in the TD(λ) learning rule (Houk, Davis, Beiser 1995; Sutton and Barto 1998; Worgotter and Porr 2005). To make the machine learning algorithms work, the network needs to know in advance the set of all possible cues and actions. In contrast, there is combinatorially large number of possible spike-timing patterns that could trigger STDP and which could represent unspecified cues and actions of the spiking network (Izhikevich 2006). Any one of them can be tied to the reward by the environment or by the experimenter, and the network can figure out which one on its own, using a more biologically plausible way than TD(λ) or other machine learning algorithms do.

Spiking Implementation of TD

Our model shows a possible spiking network implementation of some aspects of TD reinforcement learning: the shift of DA response from US to reward-predicting CS. We stress that this property was not built-in into the model, but it appeared spontaneously when we allowed synapses onto VTA-projecting neurons to be affected by DA the same way as any other synapses in the network. Thus, the shift is a general property of DA-modulated STDP applied to synaptic circuits projecting to VTA. The mechanism of the shift is quite unexpected: It takes advantage of the sensitivity of STDP to the fine temporal structure of firing of US, CS, and VTA-projecting neurons during the presentation of CS, as we explain in Figure 4*e*.

Notice that the DA response in Figure 4 is not a true error prediction signal required by TD algorithms because the model fails to exhibit depression of firing rate (dip) in the activity of the VTA_p group when US is omitted (Montague et al. 1996; Schultz 1998). On the one hand, one would not expect the depression because the intervals between CS and US are random. However, the depression would not occur even if the intervals were fixed because there is no internal clock or anticipatory signal that tells the network when US is expected. To get the depression of firing rate, one could simulate the US anticipatory signal generated by the caudate nucleus and globus pallidus (Watanabe 1996; Suri and Schultz 2001; Lauwereyns et al. 2002) and stimulate inhibitory neurons at the moment the US is expected to arrive (modeling caudate and globus pallidus is outside the scope of this paper). Notice, also, that the DA response in Figure 4 does not exhibit a gradual shift in latency, as predicted by TD models, but jumps from US to reward-predicting CS, which is more consistent with the effects observed in in vivo experiments (Pan et al. 2005). Consistent with these recordings, the DA response to US in Figure 4 does not diminish completely but remains above a baseline level. Finally, an unexpected presentation of the US after training would result in a diminished DA response in the model because the synaptic connections $US \rightarrow VTA_p$ are depressed, that is, the association is unlearned, in contrast to in vivo recordings showing a strong response (Schultz 1998, 2002). Thus, DA-modulated STDP is sufficient to reproduce some aspects of TD reinforcement learning in biologically relevant terms of spiking

activity and synaptic plasticity, but not all aspects. To address all aspects, one needs to refine the network architecture and introduce anatomical loops similar to those of basal ganglia.

Spiking versus Mean Firing Rate Models

Our study emphasizes the importance of precise firing patterns in brain dynamics: The mechanism presented in this paper works only when reward-predicting stimuli correspond to precise firing patterns. We considered only synchronous patterns embedded into the sea of noise, but the same mechanism would work equally well for polychronous firing patterns, that is, time-locked but not synchronous (Izhikevich 2006). Interestingly, *rate-based* leaning mechanisms would fail to reinforce the patterns. Indeed, presentation of a cue, such as S_1 in Figure 2, does not increase the firing rate of any neuron; it just adds, removes, or changes the time of a single spike of each of the 50 neurons in S_1. In particular, the neurons continue to fire Poissonian-looking spike trains with 1–2 spikes per second. The information about the stimulus is contained only in the relative timings of spikes, which are seen as vertical stripes in Figure 2 and which are effective to trigger STDP. A mean firing rate description of the same network would result in neuronal activities having constant values, corresponding to constant firing rates, with no possibility to know when stimulation occurs.

Conversely, DA-modulated STDP would fail to reinforce *firing rate* patterns. Indeed, large firing rate fluctuations produce multiple coincident spikes with random pre–post order, so STDP dominated by LTD would result in the average depression of synaptic strength (Kempter et al. 1999a, 1999b; Song et al. 2000). Thus, even when coincidences are not rare, STDP can still decouple chance coincidences due to rate-based dynamics from causal pre–post relations due to spike-timing dynamics (this point was stressed to the author by Wulfram Gerstner). This is how DA-modulated STDP differs from rate-based learning rules, and this is why it is so effective to selectively reinforce precise firing patterns but insensitive to firing rate patterns.

Reward versus Punishments

One can use our approach to model not only reward but also punishments. Indeed, we can treat the variable d as concentration of extracellular DA above a certain baseline level. In this case, negative values of d, interpreted as concentrations below the baseline, result in active unlearning of firing patterns, that is, in punishments. Another way to implement punishment is to assume that DA controls only the LTP part of STDP. In this case, the absence of a DA signal results in overall depression of synaptic connections (punishment), certain intermediate values of DA result in an equilibrium between LTD and LTP parts of STDP (baseline), and strong DA signals result in potentiation of eligible synaptic connections (reward). There is anecdotal evidence that the STDP curve has a very small LTP part in the prefrontal and motor cortices (Desai NS, personal communication). The model makes a testable prediction that the STDP curve will look quite different if DA is present during or immediately after the induction of synaptic plasticity.

Conclusion

DA modulation of STDP provides an elegant solution to the distal reward/credit assignment problem: only nearly coincident spiking patterns are reinforced by reward, whereas uncorrelated spikes during the delay period to the reward do not affect the eligibility traces (variables c) and hence are ignored by the network. In contrast to previous theoretical studies, 1) the network does not have to be quiet during the waiting period to the reward and 2) reward-triggering patterns do not have to be retained by recurrent activity of neurons. If a spiking pattern out of a potentially unlimited repertoire of all possible patterns, consistently, precedes or triggers reward (even seconds later), the synapses responsible for the generation of the pattern are eligible for modification when the reward arrives and the pattern is consistently reinforced (credited). Even though the network does not know what pattern was credited, it is more likely to generate the same pattern in the same behavioral context in the future.

Notes

The author thank Gerald M. Edelman, Joseph A. Gally, Niraj S. Desai, Jeff L. Krichmar, Elisabeth C. Walcott, Anil Seth, Jason G. Fleischer, Botond Szatmary, Doug Nitz, Jeff L. McKinstry, and Wulfram Gerstner for reading the earlier draft of the manuscript and making useful suggestions. The problem of credit assignment was pointed out to the author by Olaf Sporns in 2000. This research was supported by the Neurosciences Research Foundation. This material is based upon work supported by the National Science Foundation under Grant No. 0523156 (any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation). *Conflict of Interest*: None declared.

Address correspondence to email: Eugene.Izhikevich@nsi.edu.

References

Ahissar E, Vaadia E, Ahissar M, Bergman H, Arieli A, Abeles M. 1992. Hebbian plasticity in the cortex of the adult monkey depends on the behavioral context. Science. 257:1412–1415.

Au-Young SM, Shen H, Yang CR. 1999. Medial prefrontal cortical output neurons to the ventral tegmental area VTA and their responses to burst-patterned stimulation of the VTA: neuroanatomical and in vivo electrophysiological analyses. Synapse. 34:245–255.

Barad M, Bourchouladze R, Winder DG, Golan H, Kandel ER. 1998. Rolipram, a type IV-specific phosphodiesterase inhibitor, facilitates the establishment of long-lasting long-term potentiation and improves memory. Proc Natl Acad Sci USA. 95:15020–15025.

Barto AG, Sutton RS, Anderson CW. 1983. Neuronlike elements that can solve difficult learning control problems. IEEE Trans Syst Man Cybern. 13:835–846.

Bi GQ, Poo MM. 1998. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci. 18:10464–10472.

Calabresi P, Gubellini P, Centonze D, Picconi B, Bernardi G, Chergui K, Svenningsson P, Fienberg AA, Greengard P. 2000. Dopamine and cAMP-regulated phosphoprotein 32 kDa controls both striatal long-term depression and long-term potentiation, opposing forms of synaptic plasticity. J Neurosci. 20(22):8443–8451.

Cass WA, Gerhardt GA. 1995. In vivo assessment of dopamine uptake in rat medial prefrontal cortex: comparison with dorsal striatum and nucleus accumbens. J Neurochem. 65:201–207.

Centonze D, Gubellini P, Picconi B, Giacomini P, Calabresi P, Bernardi G. 1999. Unilateral dopamine denervation blocks corticostriatal LTP. J Neurophysiol. 82:3575–3579.

Choi S, Lovinger DM. 1997. Decreased probability of neurotransmitter release underlies striatal long-term depression and postnatal development of corticostriatal synapses. Proc Natl Acad Sci USA. 94:2665–2670.

Connors BW, Gutnick MJ. 1990. Intrinsic firing patterns of diverse neocortical neurons. Trends Neurosci. 13:99–104.

Dayan P, Abbott LF. 2001. Theoretical neuroscience: computational and mathematical modeling of neural systems. Cambridge (MA): The MIT Press.

Drew PJ, Abbott LF. 2006. Extending the effects of spike-timing-dependent plasticity to behavioral timescales. Proc Natl Acad Sci USA. 103:8876–8881.
Frey U, Morris RG. 1997. Synaptic tagging and long-term potentiation. Nature. 385:533–536.
Frey U, Schroeder H, Matthies H. 1990. Dopaminergic antagonists prevent long-term maintenance of posttetanic LTP in the CA1 region of rat hippocampal slices. Brain Res. 522:69–75.
Fusi S, Drew PJ, Abbott LF. 2005. Cascade models of synaptically stored memories. Neuron. 45:599–611.
Garris PA, Ciolkowski IL, Pastore P, Wightman RM. 1994. Efflux of dopamine from the synaptic cleft in the nucleus accumbens of the rat brain. J Neurosci. 14:6084–6093.
Gerstner W, Kempter R, van Hemmen JL, Wagner H. 1996. A neuronal learning rule for sub-millisecond temporal coding. Nature. 383:76–78.
Gurden H, Takita M, Jay TM. 2000. Essential role of D1 but not D2 receptors in the NMDA receptor-dependent long-term potentiation at hippocampal-prefrontal cortex synapses in vivo. J Neurosci. 20:RC106.
Hasselmo ME. 2005. A model of prefrontal cortical mechanisms for goal-directed behavior. J Cogn Neurosci. 17:1–14.
Houk JC, Adams JL, Barto AG. 1995. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk JC, Davis JL, Beiser DG, editors. Models of information processing in the basal ganglia. Cambridge (MA): The MIT Press. p. 249–270.
Houk JC, Davis JL, Beiser DG. 1995. Models of information processing in the basal ganglia. Cambridge (MA): The MIT Press.
Hull CL. 1943. Principles of behavior. New York: Appleton-Century.
Impey S, Mark M, Villacres EC, Poser S, Chavkin C, Storm DR. 1996. Induction of CRE-mediated gene expression by stimuli that generate long-lasting LTP in area CA1 of the hippocampus. Neuron. 16:973–982.
Izhikevich EM. 2006. Polychronization: computation with spikes. Neural Comput. 18:245–282.
Izhikevich EM, Gally JA, Edelman GM. 2004. Spike-timing dynamics of neuronal groups. Cereb Cortex. 14:933–944.
Jay TM, Burette F, Laroche S. 1996. Plasticity of the hippocampal-prefrontal cortex synapses. J Physiol Paris. 90:361–366.
Kempter R, Gerstner W, van Hemmen JL. 1999a. Hebbian learning and spiking neurons. Phys Rev E. 59:4498–4514.
Kempter R, Gerstner W, van Hemmen JL. 1999b. Spike-Based Compared to Rate-Based Hebbian Learning NIPS Conference; 1998 December; Denver. In: Kearns MS, Solla SA, Cohn DA, editors. Advances in neural information processing systems 11. Cambridge (MA): The MIT Press. p. 125–131.
Klopf AH. 1982. The hedonistic neuron. Washington (DC): Hemisphere.
Koene RA, Hasselmo ME. 2005. An integrate-and-fire model of prefrontal cortex neuronal activity during performance of goal-directed decision making. Cereb Cortex. 15(12):1964–1981.
Lauwereyns J, Watanabe K, Coe B, Hikosaka O. 2002. A neural correlate of response bias in monkey caudate nucleus. Nature. 418:413–417.
Levy WB, Steward O. 1983. Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. Neurosci. 8:791–797.
Lisman J. 1989. A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. Proc Natl Acad Sci USA. 86: 9574–9578.
Ljungberg T, Apicella P, Schultz W. 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophysiol. 67:145–163.
Markram H, Lubke J, Frotscher M, Sakmann B. 1997. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science. 275:213–215.
Minsky ML. 1963. Steps toward artificial intelligence. In: Feigenbaum EA, Feldman J, editors. Computers and thought. New York: McGraw-Hill. p. 406–450.
Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 16:1936–1947.
Montague PR, McClure SM, Balfwin PR, Phillips PEM, Budygin EA, Stuber GD, Kilpatrick MR, Wightman RM. 2004. Dynamic gain control of dopamine delivery in freely moving animals. J Neurosci. 24: 1754–1759.
Otani S, Daniel H, Roisin MP, Crepel F. 2003. Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. Cereb Cortex. 13:1251–1256.
Otmakhova NA, Lisman JE. 1996. D1/D5 dopamine receptor activation increases the magnitude of early long-term potentiation at CA1 hippocampal synapses. J Neurosci. 16:7478–7486.
Otmakhova NA, Lisman JE. 1998. D1/D5 dopamine receptor inhibit depotentiation at CA1 synapses via cAMP-dependent mechanism. J Neurosci. 18:1270–1279.
Pan W-X, Schmidt R, Wickens JR, Hyland BI. 2005. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward learning network. J Neurosci. 25:6235–6242.
Pavlov IP. 1927. Conditioned reflexes. Oxford: Oxford University Press.
Pasupathy A, Miller BK. 2005. Different time courses of learning-related activity in the prefrental cortex and basal ganglia. Nature. 433: 873–876.
Rao RPN, Sejnowski TJ. 2001. Spike-timing-dependent Hebbian plasticity as temporal difference learning. Neural Comput. 12: 2221–2237.
Schultz W. 1998. Predictive reward signal of dopamine neurons. J Neurophysiol. 80:1–27.
Schultz W. 2002. Getting formal with dopamine and reward. Neuron. 36:241–263.
Schultz W. Forthcoming 2007a. Reward. Scholarpedia.
Schultz W. Forthcoming 2007b. Reward signals. Scholarpedia.
Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. Science. 275:1593–1599.
Seamans JK, Yang CR. 2004. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. Prog Neurobiol. 74:1–57.
Seung HS. 2003. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron. 40:1063–1073.
Song S, Miller KD, Abbott LF. 2000. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci. 3:919–926.
Suri RE, Schultz W. 2001. Temporal difference model reproduces anticipatory neural activity. Neural Comput. 13:841–862.
Sutton RS. 1988. Learning to predict by the methods of temporal differences. Mach Learn. 3:9–44.
Sutton RS, Barto AG. 1998. Reinforcement learning: an introduction. Cambridge (MA): The MIT Press.
Swadlow HA. 1990. Efferent neurons and suspected interneurons in S-1 forelimb representation of the awake rabbit: receptive fields and axonal properties. J Neurophysiol. 63:1477–1498.
Swadlow HA. 1994. Efferent neurons and suspected interneurons in motor cortex of the awake rabbit: axonal properties, sensory receptive fields, and subthreshold synaptic inputs. J Neurophysiol. 71:437–453.
Watanabe M. 1996. Reward expectancy in primate prefrontal neurons. Nature. 382:629–632.
Wightman RM, Zimmerman JB. 1990. Control of dopaine extracellular concentration in rat striatum by impulse flow and update. Brain Res Brain Res Rev. 15:135–144.
Worgotter F, Porr B. 2005. Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. Neural Comput. 15:245–319.