CHAPTER

13

Neural Circuit Mechanisms of Value-Based Decision-Making and Reinforcement Learning

A. Soltani[1], W. Chaisangmongkon[2,3], X.-J. Wang[3,4]

[1]Dartmouth College, Hanover, NH, United States; [2]King Mongkut's University of Technology Thonburi, Bangkok, Thailand; [3]New York University, New York, NY, United States; [4]NYU Shanghai, Shanghai, China

Abstract

Despite groundbreaking progress, currently we still know preciously little about the biophysical and circuit mechanisms of valuation and reward-dependent plasticity underlying adaptive choice behavior. For instance, whereas phasic firing of dopamine neurons has long been ascribed to represent reward-prediction error (RPE), only recently has research begun to uncover the mechanism of how such a signal is computed at the circuit level. In this chapter, we will briefly review neuroscience experiments and mathematical models on reward-dependent adaptive choice behavior and then focus on a biologically plausible, reward-modulated Hebbian synaptic plasticity rule. We will show that a decision-making neural circuit endowed with this learning rule is capable of accounting for behavioral and neurophysiological observations in a variety of value-based decision-making tasks, including foraging, competitive games, and probabilistic inference. Looking forward, an outstanding challenge is to elucidate the distributed nature of reward-dependent processes across a large-scale brain system.

INTRODUCTION

Animals survive in complex environments by learning to make decisions that avoid punishments and lead to rewards. In machine learning, a number of reinforcement learning (RL) algorithms have been developed to accomplish various tasks in terms of reward-optimization problems, ranging from sequential decision-making to strategic games to training multiagent systems [1]. In neuroscience, models inspired by RL have been used both as normative descriptions of animals' behaviors and as mechanistic explanations of value-based learning processes (see also Chapters 2 and 4). Physiologically, such learning processes are thought to be implemented by changes in synaptic connections between pairs of neurons, which are regulated by reward-related signals. Over the course of learning, this synaptic mechanism results in reconfiguration of the neural network to increase the likelihood of making a rewarding choice based on sensory stimuli. The algorithmic computations of certain reinforcement models have often been translated to synaptic plasticity rules that rely on the reward-signaling neurotransmitter dopamine (DA).

There are two main theoretical approaches to derive plasticity rules that foster rewarding behaviors. The first utilizes gradient-descent methods to directly maximize expected reward—an idea known as policy gradient methods in machine learning [2,3]. Because neurons possess stochastic behaviors, many of these learning rules exploit the covariation between neural activity fluctuation and reward to approximate the gradient of reward with respect to synaptic weights. This class of learning rules has been implemented in both spiking models [4—7] and firing rate models [8—10] and has been shown to replicate operant-matching behavior in a foraging task [9], learn temporal spiking patterns [7], and generate an associative representation that facilitates perceptual decision [8,10]. The second approach is prediction-based, in which the agent estimates the values of encountered environmental states and learns decision policy that maximizes reward [1]. This idea encompasses several related algorithms, such as actor—critic models, which issue separate updates for value and policy, and temporal difference (TD) models that use a continuous prediction error signal over time [1,11]. Early work using prediction-based models aimed to explain the patterns of firing in DA neurons [12,13], whereas later work has expanded the framework to explain complex animal behaviors with more realistic spiking models [14,15].

Decision Neuroscience
http://dx.doi.org/10.1016/B978-0-12-805308-9.00013-0

In both gradient-based and prediction-based approaches, an essential element of successful learning is the reward-prediction error (RPE) signal—the difference between expected and actual reward that the agent receives. Such "teaching" signal, which derives from the Rescorla–Wagner model [16], is thought to be encoded by midbrain DA neurons. Different modeling studies interpret the DA modulatory signal differently, but the dominant idea is that DA signal dynamics mirror that of the TD error. In a series of classic experiments, Schultz et al. have shown that the firing patterns of DA neurons resembled TD error in a classical conditioning paradigm [17]. More recent experimental findings successfully used the TD framework to model DA neuron behaviors in more complex dynamic value-based choice tasks [18–21]. Experiments using optogenetics have begun to uncover the circuit logic in the ventral tegmental area that gives rise to the computation of an RPE signal [22,23] (see Chapter 5). On the other hand, some investigations into the response of DA neurons have revealed aspects of DA neural activity that do not match the theoretically defined RPE [24,25], leading others to provide alternative explanations for the observed patterns of DA response [26].

To examine the link between adaptive, value-based choice behavior and DA-dependent plasticity (Fig. 13.1), we first review various aspects of reward value representations in the brain and then discuss how such representations could be acquired through DA-dependent plasticity. The chapter will focus on a framework for reward-dependent learning in which the reinforcement signal is binary (DA release or no release) and learning does not involve any explicit optimization process. We show how this model can explain observed patterns of value representation and can account for some of the observations previously attributed to learning based on an error-driven signal and the RPE [27].

REPRESENTATIONS OF REWARD VALUE

To make good decisions, organisms must be able to assess environmental states and choose actions that are likely to yield reward. Accumulating evidence indicates that the primate brain achieves this goal by engaging multiple interconnected networks, especially in prefrontal cortex and basal ganglia, for computing and storing value estimates for sensory and motor events already experienced [28–30] (see Chapters 10, 14, and 16). In RL models, the notions of state value and action value are well distinguished. State value refers to the average reward that can be obtained in the future from the environmental state that the animal encounters at the present

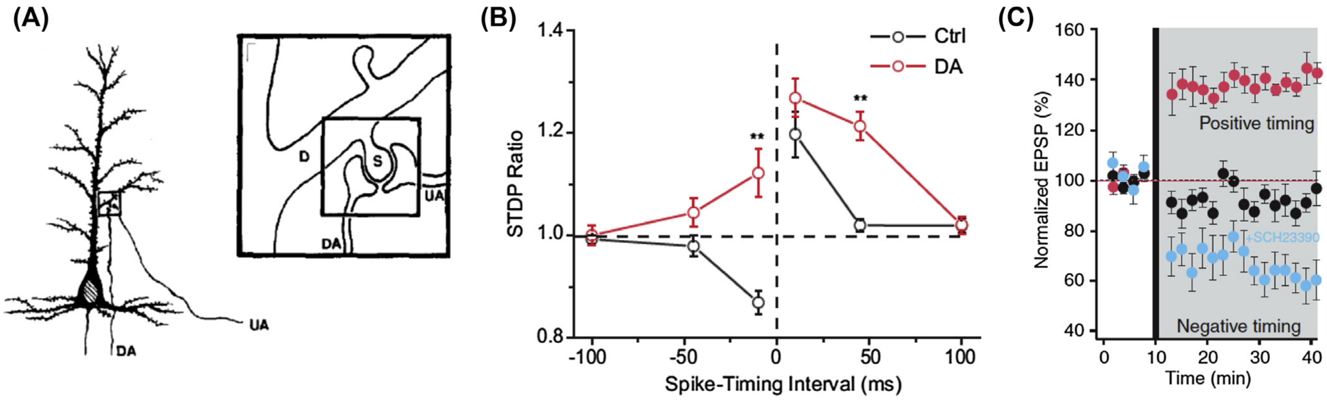

FIGURE 13.1 Modulation of synaptic plasticity by dopamine. (A) Triad configuration of dopamine (DA) and other synaptic inputs to pyramidal cells. Putative DA afferents terminate on the spine of a prefrontal pyramidal cell in the vicinity of an unidentified axon (UA). Inset shows an enlargement of axospinous synapses illustrated on the left. *(Adapted with permission from Goldman-Rakic PS. Cellular basis of working memory. Neuron 1995;14:477–85 with data published in Goldman-Rakic PS, Leranth C, Williams SM, Mons N, Geffard M. Dopamine synaptic complex with pyramidal neurons in primate cerebral cortex. Proc Natl Acad Sci USA 1989;86:9015–19.)* (B) Alteration of the spike-timing-dependent plasticity (STDP) window in hippocampal neurons with DA. *Black circles* show the STDP window under control (Ctrl) conditions and *red circles* show the same window when dopamine was present. In the presence of DA, positive spike timing includes a wider window for potentiation. With negative spike timing, dopamine reversed the direction of plasticity from depression to potentiation. *(Adapted with permission from Zhang JC, Lau PM, Bi GQ. Gain in sensitivity and loss in temporal contrast of STDP by dopaminergic modulation at hippocampal synapses. Proc Natl Acad Sci USA 2009;106(31):13028–33.)* (C) Dependence of the direction of plasticity on dopamine for cortical synapses on D1 receptor-expressing striatal, medium spiny neurons. During the STDP paradigm, when presynaptic spikes precede postsynaptic spikes (positive spike timing) long-term potentiation occurs (red). On the other hand, when presynaptic spikes follow postsynaptic spikes (negative spike timing) no changes in plasticity occur (black). When D1 receptors are blocked by SCH23390, negative timing results in long-term depression (blue). *(Adapted with permission from Surmeier DJ, Plotkin J, Shen W. Dopamine and synaptic plasticity in dorsal striatal circuits controlling action selection. Curr Opin Neurobiol 2009;19(6):621–8 with data published in Shen W, Flajolet M, Greengard P, Surmeier DJ. Dichotomous dopaminergic control of striatal synaptic plasticity. Science 2008;321(5890):848–51.) D*, dendrite; *EPSP*, excitatory postsynaptic potential; *S*, synapse.

time, whereas action value refers to the future reward expected from an action taken at a specific environmental state. Neurophysiological studies have shown that both state and action values are represented in a distributed fashion across the brain.

Signals related to reward expectancy have been abundantly found in striatum [31,32] and many subregions of prefrontal cortex [33]. More specifically, neurons in several areas of the brain encode the expected reward of the state environment, independent of actions, in accordance with the theoretical definition of state values. For example, the activity of some neurons in ventral striatum [34], dorsal anterior cingulate cortex [35], and amygdala [36] is correlated with the sum of values associated with different upcoming choice alternatives. Another type of action-independent reward expectation is termed "good-based value," meaning that the agent assigns different prices to different goods based on subjective preferences (e.g., one orange is worth more than two apples if the agent prefers orange). A series of neurophysiological studies in primates and rodents found that neurons in the orbitofrontal cortex play a prominent role in encoding the economic value of a chosen option in the unit of common currency [37,38]. The orbitofrontal cortex is known to be a part of the interconnected network that guides reward-related behaviors [39]. The network encompasses other areas that demonstrate similar chosen value encoding, such as striatum [40,41], medial prefrontal cortex [38], and ventromedial prefrontal cortex [42].

Signals related to values of specific actions are also prevalently observed across different brain areas. To name a few, the representation of action values was found in striatum [34,40,41,43,44] and posterior parietal cortex [45,46] (see also Chapter 14). These action-value neurons may exhibit correlations between firing activity and magnitude or probability of reward associated with different choices before the action is executed. The signals related to choice values were detected even in areas that directly regulate execution of movements, such as frontal eye field [47] and superior colliculus [48]. Furthermore, subsets of neurons in these areas are shown to encode the difference in values between two competing choice alternatives, which might reflect the underlying decision mechanism [34,38,46].

Overall, there is a diverse and heterogeneous representation of reward value throughout the brain. The implications of such diverse representation remain to be delineated. It is conceivable, for instance, that neurons in one brain area display value-related signal simply as a result of inputs from another area where such signal is computed. However, it is unclear what reward signal and DA-dependent learning mechanisms result in the formation of such diverse representation.

LEARNING REWARD VALUES

The prevalence of value representation across the brain begs the question of how the neurons learn to encode such properties in the first place. In RL models, a subnetwork responsible for value estimation is often referred to as the "critic" (Fig. 13.2A). An adaptive critic can be implemented simply as a single neuron (or a single pool of neurons) that receives synaptic connections from the sensory circuits, which represent environmental states [49]. The synapses are updated by a learning rule so that the activity of the critic neuron approximates the expected reward in each environmental state. The derivation of such learning rules can be accomplished by either an error-driven learning method [49] or a gradient-descent method [50]. The learning rule minimizes the difference between the actual reward and the system's reward prediction; hence, the error term is equivalent to the RPE. The idea of an adaptive critic forms the basis for TD learning, in which the expected reinforcement signal is defined by temporally discounted reward from all future time steps [49].

Early attempts in converting the TD learning algorithm into biophysical models aimed to understand the patterns of dopaminergic neuron firing [12,13,51,52]. Despite the partial success of these models in explanation of these patterns based on TD learning, there is experimental evidence indicating that other mechanisms are required to capture the exact pattern of experimental data. For example, a 2005 experiment suggested that the use of a long-lasting eligibility trace is necessary to replicate the dynamic behaviors of DA neurons as the agent is learning [53]. Interestingly, such memory traces are required to model human decisions in a sequential economic decision game [54]. These results indicate that in the neural circuits, RL may take place on a slower timescale (in the order of trial events), which is inconsistent with the rapid timescale proposed in previous theoretical studies [28,55].

Recent models have integrated more refined knowledge in machine learning to tackle difficult learning tasks, such as spatial navigation to solve a water maze [15,56] by expanding the actor−critic model for the case of continuous space and time [57]. In one spiking model, a population of critic neurons was trained by a reward-dependent Hebbian rule, which is a product of TD error and filtered pre- and postsynaptic spike trains. The synaptic kernel acts as an eligibility trace with a timescale of hundreds of milliseconds. As a result, the average firing rate of the critic population approximates the dynamic value function as the animal moves toward the target platform [15]. This is one example of many successful models that implement reward-dependent learning in spiking networks [4,6,14].

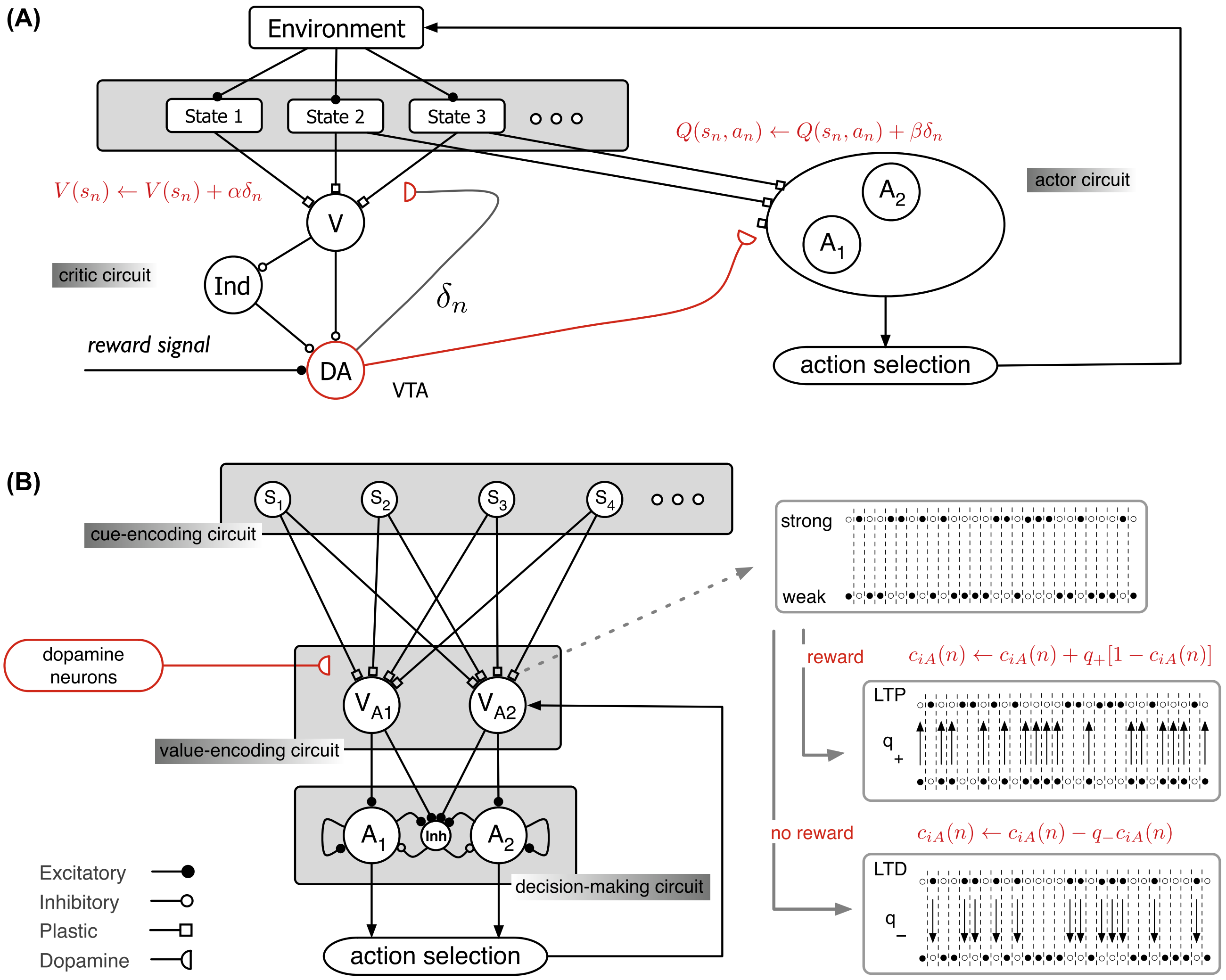

FIGURE 13.2 Comparison of two alternative models for value-based decision-making. (A) The actor–critic architecture of the temporal difference learning and use of the reward-prediction error (RPE). In this model, the critic circuit learns the reward value associated with different states of the environment, $V(s_n)$, and the actor circuit learns reward value associated with taking different actions (e.g., A_1 and A_2) in a given state, $Q(s_n, a_n)$. The critic circuit combines the reward signal and the output of state-value neurons (V pool) via a direct and indirect pathways (Ind pool) to compute the RPE. The RPE is then used to update value functions in the critic circuit and the action value functions in the actor circuit. (B) The architecture of our proposed network model with three distinct circuits: the cue-encoding, value-encoding, and decision-making circuits. The cue-encoding circuit represents sensory information pertinent for valuation and contains pools of sensory neurons that are selective for individual cues (shapes S_1 to S_4). They project to the value-encoding circuit, where neurons learn to encode reward values of the two alternative responses (action values V_{A1} and V_{A2}) through plastic synapses. At the end of the trial, only plastic synapses from sensory neurons selective for the presented cues onto action value-coding neurons selective for the chosen alternative are updated. The direction of update (long-term potentiation or depression, LTP and LTD) depends on the choice and a binary reward signal. If the choice of the model is rewarded, synapses in the depressed state transition to the potentiated state with a probability q_+; otherwise, the transition in the reverse direction occurs with a probability q_-. The decision-making circuit uses the output of the value-encoding circuit and, therefore, combines evidence from different sources to determine the model's choice on each trial. This circuit has two competing neural pools that are selective for two alternative choices (A_1 and A_2). Importantly, the finite number of synaptic states and the stochastic nature of plasticity allow synapses to transform reward signal into quantities important for value-based choice, but this transformation depends on the state of the synapses encoding reward and, therefore, on the existing representation of reward value in a given neuron. *DA*, dopamine; *VTA*, ventral tegmental area.

Nevertheless, there are a few serious caveats when applying an RPE framework in learning value-based decision-making. First, although much effort has been devoted to modeling the dynamics of DA neurons, little is understood about the temporal properties of DA release and the resulting intracellular signaling that determines synaptic plasticity in target neurons. One study suggests that the dynamics of DA release and reuptake are slow and imprecise relative to the time resolution required for RPE signals in TD learning [58]. Second, RPE signals can be both positive and negative values, whereas the neural response of DA neurons can be only positive. Although depression in the baseline activity of DA neurons could act as a negative value, the baseline activity of DA is very low (about 2 Hz), which, in turn, limits the representation of negative errors. Therefore, the decreased activity might not result in a reversal in the direction of DA-dependent plasticity (Fig. 13.1) required by TD learning unless other assumptions are included [59]. Finally, the machinery required for the computation of RPE is expensive, as it requires representation of all possible actions at all time points (with relatively high time resolution) during a given trial [12].

One could ask whether the diverse representations of reward value described in the previous section could arise only from an RPE-based mechanism or whether other types of reward signal suffice to develop such representations. One possibility would be that reward signal is not sensitive to states and actions per se, and the recipient brain areas have a crucial role in translating a more general reward signal to diverse representation of reward value.

STOCHASTIC DOPAMINE-DEPENDENT PLASTICITY FOR LEARNING REWARD VALUES

To explain various aspects of reward-based choice behavior, we have proposed an alternative learning rule based on DA-dependent synaptic plasticity. In this model, we assume plastic synapses to be bounded, such that there are a limited number of synaptic states, each with a finite value of synaptic efficacy. There are three main assumptions underlying our learning rule [60]. First, learning is reward-dependent and Hebbian, meaning that it depends on the firing rates of presynaptic and postsynaptic neurons, but is modulated by the reward signal according to the experimental data on DA-dependent synaptic plasticity [61]. Second, the reward signal mediated by DA is binary, such that presence/absence of reward is signaled by release or no release of DA. This assumption can be extended to a graded reward signal reflecting size or uncertainty of reward. Third, DA-dependent plasticity is stochastic, such that synaptic modifications occur probabilistically when conditions for plasticity are met [62−64].

In a specific formulation of our model, plastic synapses are assumed to be binary [65,66] with two discrete states: potentiated ("strong") and depressed ("weak") (Fig. 13.2B). The Hebbian characteristic of learning implies that learning depends on the firing rates of pre- and postsynaptic neurons. The presynaptic neurons encode visual targets or, more generally, stimuli that precede or predict alternative responses. On the other hand, the postsynaptic neurons represent the value of alternative actions and could be the selective neurons in a putative decision-making network. Moreover, the presence of reward is signaled by a global DA release and results in long-term potentiation, whereas the absence of reward is signaled by the lack of DA release, which reverses the direction of synaptic plasticity from potentiation to depression. Because synaptic plasticity is stochastic, in potentiation instances, each synapse in the weak state has a probability q_+ to transition to the strong state. Similarly, in depression instances, each synapse in the strong state has a transition probability q_- to move to the weak state.

The information stored at a specific set of synapses between neurons encoding stimulus s and neurons encoding action A on trial n can be quantified as the fraction of these synapses in the strong state, $c_{sA}(n)$ (or equivalently, the remaining fraction, $1 - c_{sA}(n)$, in the weak state). Based on the aforementioned assumptions, at the end of each trial when reward feedback is received, the fraction of synapses in the strong state is updated as follows:

$$\begin{aligned} c_{sA}(n+1) &= c_{sA}(n) + q_+(r, \nu_s, \nu_A)[1 - c_{sA}(n)] \\ &\quad \text{in case of LTP} \\ c_{sA}(n+1) &= c_{sA}(n) - q_-(r, \nu_s, \nu_A)c_{sA}(n) \\ &\quad \text{in case of LTD} \end{aligned} \tag{13.1}$$

where potentiation (q_+) and depression (q_-) rates, collectively called learning rates, are assumed to depend on the firing rates of pre- and postsynaptic neurons (ν_s and ν_A, respectively) and on the reward outcome ($r = 0$ or 1). More generally, the learning rates could also depend on the concentration of DA at the site of plasticity, allowing modifications of plasticity based on more graded response of DA neurons.

In a series of works, we have shown that a learning rule based on these assumptions enables the neural network to compute and estimate quantities that are crucial for performing various value-based decision-making tasks that require learning from reward feedback [55,60,67−69]. Importantly, in some cases, reward value computed by plastic synapses in our model using a binary reward signal resembles those computed by RL models using the RPE [55]. However, in our model, changes in reward value mainly depend on the state of plastic synapses receiving the reward signal, rather than the reward signal itself as in RL models.

FORAGING WITH PLASTIC SYNAPSES

In the first application of our learning rule described in Eq. (13.1), we show how reward-dependent and stochastic modifications in binary synapses allow these synapses to estimate reward value for possible choice alternatives in a value-based, decision-making task known as the "matching" task. In this task, which has been extensively used to study animals' response to reinforcement [70−73], the subject chooses between two options (color targets A and B), which are assigned different reward probabilities. Specifically, each target is baited with reward stochastically and independently of the other target, but if a reward is assigned to a target, it stays until harvested [46]. Moreover, the probability of baiting rewards on the two targets (i.e., reward schedule) changes between blocks of trials without any signal to the subject. The baiting probability ratios are randomly chosen from a set of ratios, whereas the overall baiting probability is fixed. Therefore, performing this task requires continuous estimation of reward value of the two choice options and dynamic shift of choice toward the better option.

We assumed that the neural circuit for solving this task should have a few components: sensory representation of choice options (color targets), plastic synapses that learn reward expected from choosing each target (action values), a set of neural populations for color to location remapping (because the color targets could appear on either side), and a decision-making (DM) network for choosing between the two options [60] (Fig. 13.2B). Because the two targets are identical except for the reward values, which have to be learned through reward feedback, it is reasonable to assume that target presentation leads to identical activation of presynaptic sensory neurons that project to the DM network. Therefore, the only difference in the inputs to the two competing neural populations (A and B) in the DM network is due to the efficacies of the plastic synapses. The average currents through plastic synapses selective for option A depend on the states of synapses, c_A (i.e., fraction of these synapses in the strong state), the number of plastic synapses (N_p), the presynaptic firing rate (r_{st}), and the peak conductance of the strong and weak states (g_+ and g_-) and their corresponding decay (τ_{syn}):

$$I_A \propto N_p r_{st}(c_A g_+ + (1 - c_A) g_-)\tau_{syn} \tag{13.2}$$

Therefore, the difference in the average currents into neurons selective for the two options is equal to

$$I_A - I_B \propto (c_A - c_B) N_p r_{st}(g_+ - g_-)\tau_{syn} \tag{13.3}$$

Importantly, the choice behavior of the DM network is mainly a sigmoid function of the difference in the average input currents; therefore, the probability of choosing option A can be written as

$$P_A = \frac{1}{1 + e^{-\left(\frac{c_A - c_B}{\sigma}\right)}} \tag{13.4}$$

where σ determines the inverse sensitivity to the differential current. As a result, any of the factors in the right side of Eq. (13.3) can affect the difference in the overall synaptic currents and consequently change σ and, therefore, the choice behavior. Although σ in this model is equivalent to the temperature in RL models, it can be related to biophysical properties of the neural circuit involved in value-based DM.

Based on the learning rule defined in Eq. (13.1), the level of activity (high or low) in pre- and postsynaptic neurons determines the condition for synaptic plasticity, whereas the presence or absence of reward determines its direction. However, because both targets are present on each trial of the matching task, the presynaptic neurons are always active and therefore learning becomes independent of presynaptic activity. Moreover, we assume that the postsynaptic firing rate is low (respectively, high) for the neurons selective for the unchosen (respectively, chosen) target. Finally, assuming that learning happens only if the postsynaptic neurons are highly active, the learning rule in Eq. (13.1) can be simplified as

$$\begin{aligned} &c_A(n+1) = c_A(n) + q_+[1 - c_A(n)], \\ &\text{A selected and rewarded} \\ &c_A(n+1) = c_A(n) - q_- c_A(n), \\ &\text{A selected but not rewarded} \end{aligned} \tag{13.5}$$

whereas synapses onto neurons selective for the unchosen target (B in this case) are not modified. We found that such a reward-dependent learning rule enables plastic synapses to estimate quantities crucial for matching behavior. More specifically, the steady state of synaptic strengths of the two sets of plastic synapses is a monotonic function of the returns from the two choices:

$$c_A^{ss} = \frac{q_+ R_A}{(q_+ - q_-)R_A + q_-} \tag{13.6}$$

where R_A is the return (i.e., average reward harvested per choice of A). In the special case in which $q_+ = q_-$, the steady state of c_A is equal to R_A. The steady state indicates what can be computed by plastic synapses over a long period of time. However, because these synapses have a limited number of states and bounded efficacy, the information stored in them is constantly rewritten [28]. Therefore, plastic synapses provide a local (in time) estimate of return (or a monotonic function of it). In other words, plastic synapses integrate reward feedback over time, but in a leaky fashion.

In order to see how the proposed learning rule enables the model to perform matching behavior we need to consider the interplay between learning and

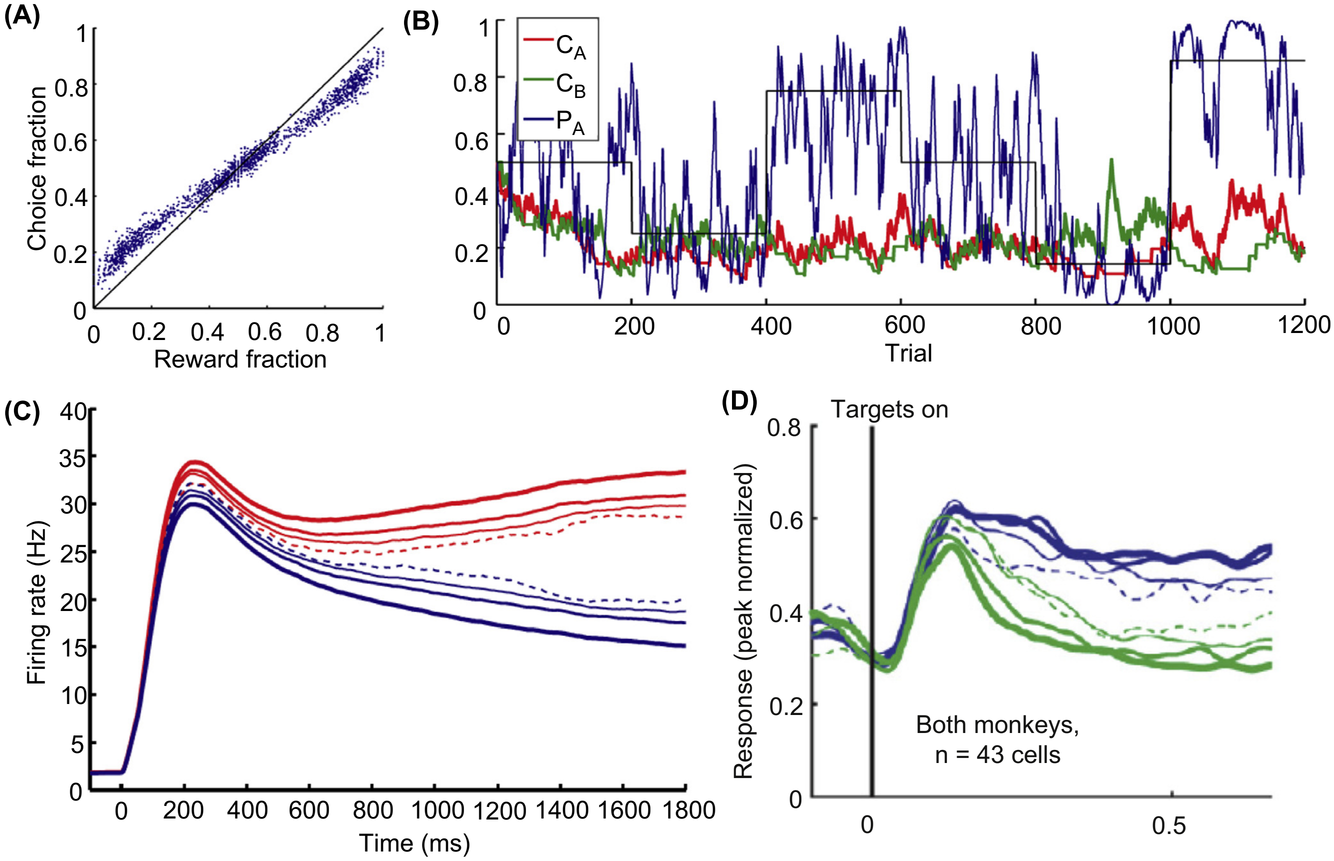

FIGURE 13.3 Behavioral and neural response of the model during the matching task. (A) The overall choice behavior of the model. Each point shows the fraction of trials within a given block on which a choice alternative is selected as a function of the fraction of reward on that choice. Matching law corresponds to the diagonal line, when the choice fraction matches the reward fraction. (B) Changes in synaptic strengths and the outcome choice behavior. Continuous update of plastic synapses selective to choices A and B, measured by synaptic strengths c_A and c_B, allows the model to track the reward fraction. P_A is the probability of choosing A and the *black line* shows the reward fraction in each block of the experiment. (C) The graded activity in the DM network during the matching task. Plotted is the average response of neurons in the two decision-making populations selective for the two targets. Activity is aligned on target onset and is shown separately for the choice that is the preferred (red) or nonpreferred (blue) target of the neurons. Moreover, trials are subdivided into four groups according to the difference between the strength of synapses onto the two pools of neurons selective to choices A and B: [$(c_A - c_B) = -0.05$ to -0.14 (*dashed*), 0 to -0.05 (*thin*), 0 to 0.05 (*normal*), 0.05 to 0.14 (*thick*)]. *((A)—(C) are adapted from Soltani A, Wang X-J. A biophysically-based neural model of matching law behavior: melioration by stochastic synapses. J Neurosci 2006;26:3731—44.)* (D) Average peak-normalized firing rates of lateral intraparietal area neurons over time. Activity is aligned on target onset and different colors correspond to trials when choice was into (blue) and out of (green) the neuron's receptive field. Trials are subdivided into four groups according to the local fractional income of the chosen target: *solid thick lines*, 0.75 to 1.0; *solid medium lines*, 0.5 to 0.75; *solid thin lines*, 0.25 to 0.5; *dotted thin lines*, 0 to 0.25. *(Adapted from Sugrue LP, Corrado GC, Newsome WT. Matching behavior and representation of value in parietal cortex. Science 2004;304:1782—7.)*

decision-making. On one hand, plastic synapses approximate return from each choice option (Eq. (13.6)). On the other hand, the difference between synaptic strengths, $c_A - c_B$ (Eq. (13.4)) determines the choice probability, which in turn modifies the returns (Fig. 13.3). This interplay between the synaptic strengths (or equivalently returns) and the choice probability gives rise to the dynamics underlying matching behavior. At each moment, the model selects the choice option with a larger return (say $R_A > R_B$) with a higher probability ($P_A > .5$). This then reduces the return on that option (without changing the return on the unselected option) because not every selection of A is accompanied by a reward. This continues until the return on A falls below the return on the other options ($R_A < R_B$), which then causes the model to select choice B more often.

Because of the probabilistic nature of learning and DM in our model, there is always a limit for approaching perfect matching. The dynamic bias of choice toward the more rewarding option makes the model reach a choice probability that is generally smaller but close to the prediction of matching (i.e., undermatching). The extent of undermatching depends on the value of σ so that for a smaller value of σ, the steady state of the model is closer to the prediction of matching. Moreover, the model's estimate of reward values and ensuing choice behavior always fluctuate around a steady state due to ongoing learning. Therefore, perfect matching can be achieved by reducing the learning rates to zero however, such a solution hinders the model from adapting to changes in the reward schedule. Because the model selects the better option (in terms of return), our model acts according to the melioration principle [74,75], which states that the choice behavior should be biased toward the option with the higher return. However, decision is not deterministic in our model and, therefore, our model achieves matching through "probabilistic" melioration.

The model's choice and reward sequences can be used to quantify the dependence of choice on the history of reward by employing different methods, such as the "choice-triggered average of rewards" (CTA) [76]. Such analysis reveals that similar to experimental data and RL models based on RPE, the choice behavior of the model depends on previous reward outcomes in an exponential fashion or a variant of it (e.g., the sum of two exponentials). Interestingly, the time constant of the CTA is a function of the overall baiting probability (or the maximal reward rate) in the environment, such that it decreases as the overall baiting probability increases. In other words, the abundance of reward reduces the time constant of reward influence on choice. This happens because more rewarding instances increases the frequency of potentiation of plastic synapses, reducing the impact of more distant reward. The dependence of the time constant of reward integration on the overall reward rate in the environment enables the model to achieve matching over a wide range of learning rates.

Finally, the pattern of neural activity in the DM circuit of the model matches the graded neural response observed in the lateral intraparietal cortex during this task [46] (Fig. 13.3). This result supports the idea that DA-dependent learning based on a binary reward signal could account for the formation of value signal, as well as the resulting choice behavior.

RANDOM CHOICE AND COMPETITIVE GAMES

Many social interactions, often studied using games [77], require learning and predicting the behavior of other agents while behaving unpredictably to avoid getting exploited by other agents. Because game theory provides only a normative account of the average behavior (or equilibrium) in a given game, RL models are often used to describe the dynamic choice behavior. But how does the dynamics of choice depend on the underlying learning? To answer this question and explore the neural mechanisms of dynamic choice behavior during competitive games, we incorporated our DA-dependent plasticity into a DM network and simulated the choice behavior during the game of matching pennies [55].

During the game of matching pennies, monkeys were required to freely choose between one of two visual targets (on the left and right sides of the fixation point) and were rewarded only if their choice matched the computer opponent's choice on a given trial (see also Chapter 21). The optimal strategy during this game was to choose the two targets randomly with equal probability [78,79]. However, the strategy or algorithm used by the computer opponent became more complex in three successive stages. In the first stage, the computer selected one of the two targets randomly, each with $p = .5$ (algorithm 0). In the second stage, the computer used the entire history of the animal's previous choices in a given session to predict the monkey's next choice (algorithm 1). The maximum reward could be obtained in algorithm 1 if the animal selected the two targets with equal probability and independently of its previous choices. In the final stage, the computer used the entire history of the animal's choice and reward in a given session to predict the monkey's choice on the next trial (algorithm 2). Therefore, the maximum performance could be obtained if the animal selected its targets, not only with equal probability and independently of its previous choices, but also independently of the combination of its previous choices and their outcomes.

Interestingly, animals showed specific patterns of choice during these three stages. During algorithm 0 when the computer opponent selected between two targets randomly, monkeys displayed a strong bias (choice bias) toward one of the two targets. These choice biases quickly diminished when the computer started to exploit the animal's preference for one of the targets in algorithm 1, but a new bias emerged. Specifically, the animal tended to repeat its decision if the choice was rewarded on the previous trial (win−stay strategy) and switch to the alternative target if the previous choice was not rewarded (lose−switch). Interestingly, the use of the win−stay−lose−switch (WSLS) strategy slowly increased over the period of many days. However, following the introduction of algorithm 2, the probability of WSLS strategy declined toward .5.

In many competitive games, it is important to select between choices stochastically to outsmart the opponent. Our model exhibits such probabilistic behavior because neural spike discharges are intrinsically stochastic. The network's choice varies from trial to trial but is more frequently biased toward the target with a stronger input determined by the recent reward history on the two targets. Similar to the matching game, the choice probability in this task is also a sigmoid function of the difference between the synaptic strengths as in

Eq. (13.4), where the value of σ determines the randomness of the choice behavior. Therefore, any of the biophysical factors appearing on the right side of Eq. (13.3) can change the value of σ and therefore, the desired stochastic behavior.

The specific form of the learning rule used for simulating foraging behavior (Eq. (13.5)) can also be applied to the game of matching pennies. However, for foraging we assumed that only synapses projecting to neurons selective for the chosen target are modified, making the learning rule "choice specific." Such choice-specific learning rule is equivalent to stateless Q-learning [1]. Alternatively, the general learning rule described by Eq. (13.1) can be simplified by assuming that synapses projecting to neurons selective to the unchosen target are also modified by the same amount as the synapses projecting to neurons selective for the chosen target, but in the opposite direction:

$$\begin{gathered}
\text{Right is selected and rewarded:} \\
\begin{cases} c_R(t+1) = c_R(t) + (1 - c_R(t))q_r \\ c_L(t+1) = c_L(t) - c_L(t)q_r \end{cases} \\
\text{Right is selected but not rewarded:} \\
\begin{cases} c_R(t+1) = c_R(t) - c_R(t)q_n \\ c_L(t+1) = c_L(t) + (1 - c_L(t))q_n \end{cases}
\end{gathered} \tag{13.7}$$

where q_r and q_n are the learning rates in the rewarded and unrewarded trials, respectively. In game theory, learning rules that indiscriminately modify the value functions for chosen and unchosen actions are referred to as belief learning [80,81]. These results show the flexibility of our general learning rule in producing different update rules used in the RL, but with the advantage that our model connects learning rates to transition probabilities at the synaptic level.

The strong choice bias observed during algorithm 0 could be attributed to lack of incentive for the subject to adopt the equilibrium strategy (i.e., choosing the two targets with equal probability). An alternative explanation would be that reward-dependent learning causes the equilibrium strategy to be unstable and, therefore, unattainable. Analyzing the steady state of the model's choice behavior, we found that for specific values of learning rates, the equilibrium strategy becomes unstable, resulting in a strong bias toward one of the two choices.

This happens if $\frac{\sigma(q_r+q_n)}{2(q_r-q_n)} < 0.25$ when learning rates for rewarded trials are larger than unrewarded trials [i.e., $(q_r - q_n) > 0$]. On the other hand, when $(q_r - q_n) < 0$, an unstable equilibrium occurs when $\frac{\sigma(2-q_r-q_n)}{2(q_n-q_r)} < 0.25$. Therefore, a biased choice behavior does not necessarily reflect insensitivity to reinforcement or feedback, but instead may result from a reward-dependent learning mechanism.

These results were obtained based on the assumption that the choice behavior is not intrinsically biased toward one of the choices. However, the DM network may have an intrinsic bias due to differences in pathways that drive the two competing pools in the DM circuit, resulting in preference for one target and shifting the steady state of choice behavior to a point rather than $p = .5$. Although we found that choice bias could emerge from learning when the computer opponent selects randomly, the question remains as to whether the same learning can mitigate an intrinsic bias in the network if the choice bias is penalized by the opponent (e.g., in algorithm 1). Interestingly, we found that such an intrinsic bias can be compensated through DA-dependent learning that modifies the strength of plastic synapses based on reward feedback during algorithm 1.

Finally, an important aspect of behavioral data described above was a slow change in the probability of WSLS over the course of many days [79]. Could such a change stem from an ongoing learning mechanism that tries to adjust the learning rate over time? To answer this question, we implemented a modified version of the metalearning algorithm [82]. The goal of a metalearning model is to maximize the long-term average of rewards, by comparing the medium-term and long-term running averages of the reward rate. We found that metalearning can account for the gradual change in the animal's strategy. Moreover, we found that, in addition to the initial condition and metalearning parameters, the probabilistic nature of this task contributes to the time course of the choice behavior. Therefore, metalearning and its stochastic nature can provide a mechanism for generating a diverse repertoire of choice behavior observed in various competitive games (see also Chapters 16 and 21).

PROBABILISTIC INFERENCE WITH STOCHASTIC SYNAPSES

In the third application of our DA-dependent plasticity rule (Eq. (13.1)), we show how such a learning mechanism enables plastic synapses to perform probabilistic inference. Probabilistic inference is the ability to combine information from multiple sources that are only partially predictive of alternative outcomes, as well as making inferences about the predictive power of individual sources. These tasks are challenging in naturalistic situations because often only a single action

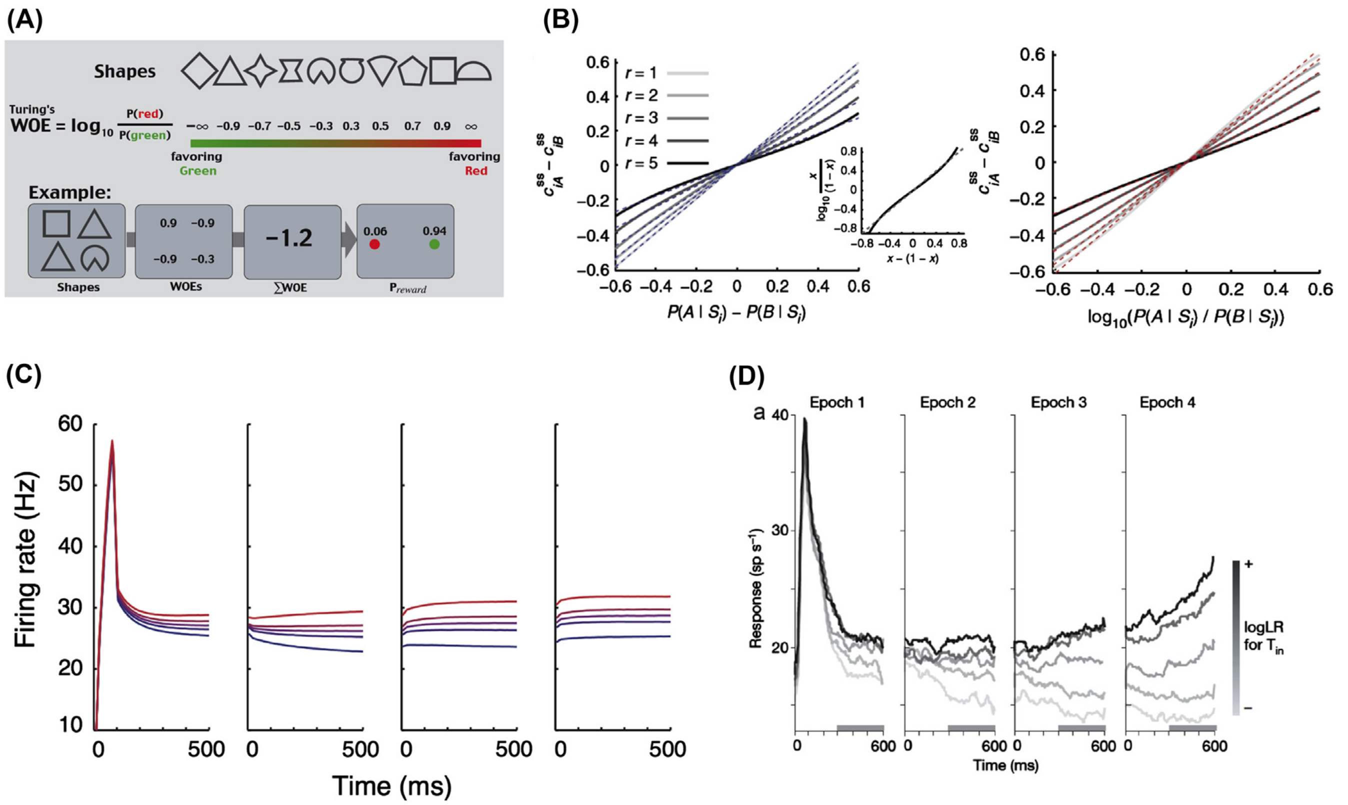

FIGURE 13.4 Behavioral and neural response of the model during the weather prediction task. (A) Reward assignment on each trial of the experiment. Reward is assigned based on the sum of the weight of evidence (WOE) associated with the presented shapes. (B) Difference in the steady state of the synaptic strengths as a function of the difference in the posteriors (left) and of the log posterior odds (right), for different learning rate ratios. *Dashed lines* show linear fits for the values of posterior between 0.2 and 0.8. The inset shows the relationship between $\log_{10}\left(\frac{x}{1-x}\right)$ and $x-(1-x)$ over the same range, where $x=P(A|S_i)$ and $P(B|S_i)=1-P(A|S_i)=1-x$. (C) Effect of the log likelihood ratio (LR) on the firing rate of model neurons in the DM network. Plotted is the average population activity over many trials, computed for five quintiles of the log LR in each epoch (more red means larger log LR). The response is aligned to the onset of each shape in a given epoch. *((B) and (C) are adapted from Soltani A, Wang X-J. Synaptic computation underlying probabilistic inference. Nat Neurosci 2010;13:112−9.)* (D) Effect of the log LR on the population average firing rate in a monkey's lateral intraparietal area. Average responses are aligned to the onset of the shapes and extend 100 ms into the subsequent epoch. The averages were computed for five quintiles of log LR in each epoch (indicated by *shading*). *(Adapted from Yang T, Shadlen MN. Probabilistic reasoning by neurons. Nature 2007;447:1075−80.)*

outcome such as a binary reward feedback follows the presentation of multiple sources of information (or cues), and so it is unclear how presented cues should be associated with the final outcome.

An example of such a naturalistic situation is simulated in the so-called weather prediction task, in which a categorical choice (rain or sunshine) is predicted based on a number of given cues [83]. A 2007 study using a variant of this probabilistic categorization task suggested that even monkeys are capable of some forms of probabilistic inference and revealed neural correlates of this ability at the single-cell level [84]. In this task, four shapes precede a selection between two color targets on each trial [84] (Fig. 13.4A). These shapes were selected randomly from a set of 10 distinguishable shapes (S_i, $i=1, 2, \ldots, 10$), each of which was allocated a unique weight of evidence (WOE) about the probability of reward assignment on one of the two choice alternatives $\left(\mathrm{WOE}=\log_{10}\frac{P(A|S_i)}{P(B|S_i)}\right)$. The computer assigned a reward to one of the two alternatives with a probability that depended on the sum of the WOEs from all the shapes presented on a given trial.

We used the three-circuit network model (Fig. 13.2B) to simulate the choice behavior and neural activity during the weather prediction task to test whether our proposed learning rule (Eq. (13.1)) enables probabilistic inference. First, we found that the specific form of the learning rule allows plastic synapses to estimate variants of posteriors depending on the number of shapes presented simultaneously. When one shape is presented alone, the steady state of the synaptic strength for synapses selective for shape i and alternative A is

$$c_{iA}^{ss} = \frac{rP(A|S_i)}{1 + (r-1)P(A|S_i)} \qquad (13.8)$$

where $P(A|S_i)$ is the probability that A is assigned with reward, given shape i is presented, and r is the learning rate ratio ($r \equiv q_+/q_-$). Thus, when each cue is presented alone, the steady state is proportional to posteriors. When several shapes are presented together, the reward feedback is determined by all presented shapes, and synapses selective for all these shapes are updated. As a result, plastic synapses estimate the naïve posterior probability, $\widetilde{P}(A|S_i)$, or the posterior probability that a choice alternative is assigned with reward given that a cue is presented in any combination of cues.

Importantly, the decision circuit stochastically generates a categorical choice with a probability, which is a sigmoid function of the difference in the overall inputs (differential input, ΔI) to its selective pools [55,60,67,85]. We assume that cue-selective neurons fire at a similar rate, and, therefore, the differential input is determined by the sum of the difference in the synaptic strengths $\left(\Delta c_i^{ss} \equiv c_{iA}^{ss} - c_{iB}^{ss}\right)$ from the action value-coding neurons onto the decision neurons (Fig. 13.4B). Using Eq. (13.8) and replacing posteriors for naïve posteriors, we can compute the differential input:

$$\Delta I \propto \sum\nolimits_i \Delta c_i^{ss} \propto \sum\nolimits_i \frac{r\left(\widetilde{P}(A|S_i) - \widetilde{P}(B|S_i)\right)}{r + (r-1)^2 \widetilde{P}(A|S_i)\widetilde{P}(B|S_i)} \qquad (13.9)$$

This formula can be simplified by noting that $r \gg (r-1)^2 \widetilde{P}(A|S_i)\widetilde{P}(B|S_i)(\equiv k)$, which happens when $r \approx 1$ and the values of posterior probabilities are in an intermediate range $\left(2 \leq \left(\widetilde{P}(A|S_i)\right) \leq .8\right)$:

$$\Delta I \propto \sum\nolimits_i \left(1 - \frac{k}{r}\right)\left(\widetilde{P}(A|S_i) - \widetilde{P}(B|S_i)\right) \qquad (13.10)$$

Using another simplification, $x - (1-x) \simeq \log_{10}\left(\frac{x}{1-x}\right)$ (true if $0.2 \leq x \leq 0.8$) we get:

$$\Delta I \propto \sum\nolimits_i \left(1 - \frac{k}{r}\right) \log_{10} \frac{\widetilde{P}(A|S_i)}{\widetilde{P}(B|S_i)} \qquad (13.11)$$

Note that for smaller or larger values of posteriors, the choice behavior is deterministic and so our simplification does not affect the final calculation. Therefore, because of the convergence of sensory neurons onto action value-coding neurons, the latter naturally summate the currents through sets of plastic synapses related to presented cues. Subsequently, the outputs of action value neurons drive the decision circuit, and the resulting choice is a function of the sum of log naïve posterior odds. Thus, summation of currents through plastic synapses provides a natural mechanism for integrating evidence from different cues in terms of log posterior odds.

Overall, our model provides a solution to probabilistic inference based on a mixture of synaptic and neural mechanisms. On the one hand, modifications of synapses selective for the presented shapes enables plastic synapses to directly estimate naïve posteriors (but not posteriors). A consequence of such a learning rule is that the evidence a model assigns to a given cue is smaller than the WOE assigned to that cue. On the other hand, summation of currents evoked by presented cues causes the model to integrate evidence in terms of log posterior odds. This feature enables the model to perform cue combination near optimality (i.e., according to Bayes' rule), but only given equal priors. When priors are unequal, plastic synapses carry information about priors as well, and therefore, our model provides an answer different from what is expected by adding log prior odds to the summed log likelihood ratios. More specifically, the model accounts for a cognitive bias known as base-rate neglect, that is, a cue that is equally predictive of each outcome is perceived to be more predictive of the less probable outcome [86]. Moreover, it predicts a bias in making inference about a combination of cues which depends on the number of cues used for making inference. Interestingly, a recent study has provided evidence for both predictions of the model [86a].

Finally, neural activity in our model reproduces the main physiological observations from the lateral intraparietal area in the monkey experiment [84] (Fig. 13.4C and D). Overall, our results demonstrate that empirically observed neural correlates of probabilistic inference can rely on synaptic, rather than neuronal, computations. Despite the complexity of the weather prediction task, stochastic DA-dependent plasticity enables the model to perform the task using a binary "teaching" signal.

CONCLUDING REMARKS

Reward is one of the determining factors for choice behavior, and the neural correlates of reward value are observed in many brain areas. Consequently, linking the influence of reward on behavior and the measured neural response (i.e., representation of reward value) is one of the critical goals in the study of DM. It is generally believed that reward is signaled throughout the brain via DA [87–89], and DA-dependent plasticity provides the neural substrates for learning reward values; however, the neural mechanisms underlying these processes are not fully understood.

There are reasons for why identifying neural mechanisms underlying value-based learning and representation of reward is challenging. Dopaminergic neurons project to many brain areas, where DA influences both synaptic plasticity and neural excitability; the diverse effects remain poorly explored (see also Chapter 2). Furthermore, neural correlates of value-based behavior are often measured in one area at the time, not allowing for the exploration of the role of interactions between different areas in value-based DM. Similarly, computational models of value-based DM and learning often are not concerned with the plausibility of the learning rules, ignore important effects of DA on neuronal processes such as short-term synaptic plasticity and changes in neural excitability [69], overlook different types of DA receptors involved in dopaminergic modulations, and do not consider interactions between brain areas.

To move forward, the focus needs to shift to biological plausibility of proposed learning rules and future models of value-based DM. Learning should go beyond a local circuit and include interactions between various circuits in determining the choice behavior. Understanding the role of interactions between different brain areas in value-based DM may elucidate the origin and function of various types of value representation throughout the brain (see Chapters 10, 14, 16, and 24). The diverse representation of reward value throughout the brain also begs the question of investigating DA-dependent plasticity in those areas, and the role of local circuits in translating a “global” reward signal into meaningful information.

Acknowledgments

This work was supported by NIH Grant R01MH062349 to Xiao-Jing Wang.

References

[1] Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge (MA): MIT Press; 1998.

[2] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 1992;8:229–56.

[3] Baxter J, Bartlett PL. Infinite-horizon policy-gradient estimation. Artif Intell 2001;15:319–50.

[4] Vasilaki E, Fremaux N, Urbanczik R, Senn W, Gerstner W. Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. PLoS Comput Biol 2009;5:e1000586.

[5] Xie X, Seung HS. Learning in neural networks by reinforcement of irregular spiking. Phys Rev E 2004;69:1–10.

[6] Florian RV. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. Neural Comput 2007;19: 1468–502.

[7] Pfister JP, Toyoizumi T, Barber D, Gerstner W. Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. Neural Comput 2006;18:1318–48.

[8] Engel TA, Chaisangmongkon W, Freedman DJ, Wang X-J. Choice-correlated activity fluctuations underlie learning of neuronal category representations. Nat Commun 2015;6. http://dx.doi.org/10.1038/ncomms7454 [Article No: 6454].

[9] Loewenstein Y, Seung HS. Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. Proc Natl Acad Sci USA 2006;103:15224–9.

[10] Roelfsema PR, van Ooyen A. Attention-gated reinforcement learning of internal representations for classification. Neural Comput 2005;17:2176–214.

[11] Sutton RS. Learning to predict by the methods of temporal differences. Mach Learn 1988;3:9–44.

[12] Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 1996;16:1936–47.

[13] Suri RE, Schultz W. Temporal difference model reproduces anticipatory neural activity. Neural Comput 2001;13:841–62.

[14] Potjans W, Morrison A. A spiking neural network model of an actor-critic learning agent. Neural Comput 2009;339:301–39.

[15] Frémaux N, Sprekeler H, Gerstner W. Reinforcement learning using a continuous time actor-critic framework with spiking neurons. PLoS Comput Biol 2013;9:e1003024.

[16] Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning II: current research and theory; 1972. p. 64–9.

[17] Schultz W. Neuronal reward and decision signals: from theories to data. Physiol Rev 2015;95:853–951.

[18] Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 2005;47: 129–41.

[19] Nomoto K, Schultz W, Watanabe T, Sakagami M. Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. J Neurosci 2010;30:10692–702.

[20] Glimcher PW. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci USA 2011;108(Suppl. 3):15647–54.

[21] Enomoto K, et al. Dopamine neurons learn to encode the long-term value of multiple future rewards. Proc Natl Acad Sci USA 2011;108:15462–7.

[22] Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 2012;482:85–8.

[23] Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N. Arithmetic and local circuitry underlying dopamine prediction errors. Nature 2015;525:243–6.

[24] Fiorillo CD, Tobler PN, Schultz W. Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors. Behav Brain Funct 2005;1: 1–5.

[25] Fiorillo CD, Newsome WT, Schultz W. The temporal precision of reward prediction in dopamine neurons. Nat Neurosci 2008;11: 966–73.

[26] Redgrave P, Gurney K. The short-latency dopamine signal: a role in discovering novel actions? Nat Rev Neurosci 2006;7:967–75.

[27] Daw ND, Tobler PN. Value learning through reinforcement: the basics of dopamine and reinforcement learning. In: Neuroeconomics. Academic Press; 2013. p. 283–98.

[28] Soltani A, Wang X-J. From biophysics to cognition: reward-dependent adaptive choice behavior. Curr Opin Neurobiol 2008; 18:209–16.

[29] Lee D, Seo H, Jung MW. Neural basis of reinforcement learning and decision making. Annu Rev Neurosci 2012;35:287–308.

[30] Louie K, Glimcher PW. Efficient coding and the neural representation of value. Ann NY Acad Sci 2012;1251:13–32.

[31] Cromwell HC, Schultz W. Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. J Neurophysiol 2003;89:2823–38.

[32] Hollerman JR, Tremblay L, Schultz W. Influence of reward expectation on behavior-related neuronal activity in primate striatum. J Neurophysiol 1998;80:947−63.

[33] Wallis JD, Kennerley SW. Heterogeneous reward signals in prefrontal cortex. Curr Opin Neurobiol 2010;20:191−8.

[34] Cai X, Kim S, Lee D. Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. Neuron 2011;69:170−82.

[35] Seo H, Lee D. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J Neurosci 2007;27:8366−77.

[36] Belova MA, Paton JJ, Salzman CD. Moment-to-moment tracking of state value in the amygdala. J Neurosci 2008;28:10023−30.

[37] Padoa-Schioppa C, Assad JA. Neurons in the orbitofrontal cortex encode economic value. Nature 2006;441:223−6.

[38] Sul JH, Kim H, Huh N, Lee D, Jung MW. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron 2010;66:449−60.

[39] Öngür D, Price J. The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. Cereb Cortex 2000;10:206−19.

[40] Lau B, Glimcher PW. Value representations in the primate striatum during matching behavior. Neuron 2008;58:451−63.

[41] Kim H, Sul JH, Huh N, Lee D, Jung MW. Role of striatum in updating values of chosen actions. J Neurosci 2009;29:14701−12.

[42] Kable JW, Glimcher PW. The neurobiology of decision: consensus and controversy. Neuron 2009;63:733−45.

[43] Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science 2005;310:1337−40.

[44] Lau B, Glimcher PW. Action and outcome encoding in the primate caudate nucleus. J Neurosci 2007;27:14502−14.

[45] Platt ML, Glimcher PW. Neural correlates of decision variables in parietal cortex. Nature 1999;400:233−8.

[46] Sugrue LP, Corrado GC, Newsome WT. Matching behavior and representation of value in parietal cortex. Science 2004;304:1782−7.

[47] Ding L, Hikosaka O. Comparison of reward modulation in the frontal eye field and caudate of the macaque. J Neurosci 2006;26:6695−703.

[48] Ikeda T, Hikosaka O. Reward-dependent gain and bias of visual responses in primate superior colliculus. Neuron 2003;39:693−700.

[49] Houk J, Davis J. Models of information processing in the basal ganglia. The MIT Press; 1994.

[50] Si J, Barto AG, Powell WB, Wunsch D. Handbook of learning and approximate dynamic programming (IEEE press series on computational intelligence). Wiley-IEEE Press; 2004.

[51] Montague P, Dayan P, Person C, Sejnowski T, et al. Bee foraging in uncertain environments using predictive Hebbian learning. Nature 1995;377:725−8.

[52] Suri RE, Schultz W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience 1999;91:871−90.

[53] Pan W-X, Schmidt R, Wickens JR, Hyland BI. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J Neurosci 2005;25:6235−42.

[54] Bogacz R, McClure SM, Li J, Cohen JD, Montague PR. Short-term memory traces for action bias in human reinforcement learning. Brain Res 2007;1153:111−21.

[55] Soltani A, Lee D, Wang X-J. Neural mechanism for stochastic behavior during a competitive game. Neural Netw 2006;19:1075−90.

[56] Foster D, Morris R, Dayan P, et al. A model of hippocampally dependent navigation, using the temporal difference learning rule. Hippocampus 2000;10:1−16.

[57] Doya K. Reinforcement learning in continuous time and space. Neural Comput 2000;12:219−45.

[58] Montague PR, et al. Dynamic gain control of dopamine delivery in freely moving animals. J Neurosci 2004;24:1754−9.

[59] Potjans W, Diesmann M, Morrison A. An imperfect dopaminergic error signal can drive temporal-difference learning. PLoS Comput Biol 2011;7:e1001133.

[60] Soltani A, Wang X-J. A biophysically-based neural model of matching law behavior: melioration by stochastic synapses. J Neurosci 2006;26:3731−44.

[61] Reynolds JN, Wickens JR. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw 2002;15:507−21.

[62] Amit DJ, Fusi S. Dynamic learning in neural networks with material synapses. Neural Comput 1994;6:957−82.

[63] Fusi S. Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. Biol Cybern 2002;87:459−70.

[64] Fusi S, Drew PJ, Abbott LF. Cascade models of synaptically stored memories. Neuron 2005;45:599−611.

[65] Petersen CC, Malenka RC, Nicoll RA, Hopfield JJ. All-or-none potentiation at CA3-CA1 synapses. Proc Natl Acad Sci USA 1998;95:4732−7.

[66] O'Connor DH, Wittenberg GM, Wang SS-H. Graded bidirectional synaptic plasticity is composed of switch-like unitary events. Proc Natl Acad Sci USA 2005;102:9679−84.

[67] Fusi S, Asaad WF, Miller EK, Wang X-J. A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. Neuron 2007;54:319−33.

[68] Soltani A, Wang X-J. Synaptic computation underlying probabilistic inference. Nat Neurosci 2010;13:112−9.

[69] Soltani A, Noudoost B, Moore T. Dissociable dopaminergic control of saccadic target selection and its implications for reward modulation. Proc Natl Acad Sci USA 2013;110:3579−84.

[70] Herrnstein RJ. Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav 1961;4:267−72.

[71] Williams BA. Reinforcement, choice, and response strength. In: Atkison RC, Herrnstein RJ, Lindzey G, Luce RD, editors. Steven's handbook of experimental psychology. 2nd ed., vol. 2. New York: Wiley; 1988. p. 167−244.

[72] Gallistel CR. Foraging for brain stimulation: toward a neurobiology of computation. Cognition 1994;50:151−70.

[73] Herrnstein RJ, Rachlin H, Laibson DI. The matching law: papers in psychology and economics. Harvard UP; 1997.

[74] Herrnstein RJ, Vaughan WJ. Melioration and behavioral allocation. In: Staddon JER, editor. Limits to action: the allocation of individual behavior. New York: Academic; 1980. p. 143−76.

[75] Herrnstein RJ, Prelec D. Melioration: a theory of distributed choice. J Econ Perspect 1991;5:137−56.

[76] Corrado GS, Sugrue LP, Seung HS, Newsome WT. Linear-nonlinear-Poisson models of primate choice dynamics. J Exp Anal Behav 2005;84:581−617.

[77] Lee D. Game theory and neural basis of social decision making. Nat Neurosci 2008;11:404−9.

[78] Barraclough DJ, Conroy ML, Lee D. Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci 2004;7:404−10.

[79] Lee D, Conroy ML, McGreevy BP, Barraclough DJ. Reinforcement learning and decision making in monkeys during a competitive game. Brain Res Cogn Brain Res 2004;22:45−58.

[80] Camerer CF. Behavioral game theory: experiments in strategic interaction. Princeton: Princeton Univ. Press; 2003.

[81] Lee D, McGreevy BP, Barraclough DJ. Learning and decision making in monkeys during a rock-paper-scissors game. Brain Res Cogn Brain Res 2005;25:416−30.

[82] Schweighofer N, Doya K. Meta-learning in reinforcement learning. Neural Netw 2003;16:5−9.

[83] Knowlton BJ, Squire LR, Gluck MA. Probabilistic classification learning in amnesia. Learn Mem 1994;1:106—20.

[84] Yang T, Shadlen MN. Probabilistic reasoning by neurons. Nature 2007;447:1075—80.

[85] Wang X-J. Probabilistic decision making by slow reverberation in cortical circuits. Neuron 2002;36:955—68.

[86] Gluck MA, Bower GH. From conditioning to category learning: an adaptive network model. J Exp Psychol Gen 1988;117:227—47.

[86a] Soltani A, Khorsand P, Guo CZ, Farashahi S, Liu J. Neural substrates of cognitive biases during probabilistic inference. Nat Commun 2016;7:11393.

[87] Schultz W. Multiple reward signals in the brain. Nat Rev Neurosci 2000;1:199—207.

[88] Vickery TJ, Chun MM, Lee D. Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron 2011;72: 166—77.

[89] Clark AM. Reward processing: a global brain phenomenon? J Neurophysiol 2013;109:1—4.

[90] Goldman-Rakic PS. Cellular basis of working memory. Neuron 1995;14:477—85.

[91] Goldman-Rakic PS, Leranth C, Williams SM, Mons N, Geffard M. Dopamine synaptic complex with pyramidal neurons in primate cerebral cortex. Proc Natl Acad Sci USA 1989;86:9015—9.

[92] Zhang JC, Lau PM, Bi GQ. Gain in sensitivity and loss in temporal contrast of STDP by dopaminergic modulation at hippocampal synapses. Proc Natl Acad Sci USA 2009;106(31):13028—33.

[93] Surmeier DJ, Plotkin J, Shen W. Dopamine and synaptic plasticity in dorsal striatal circuits controlling action selection. Curr Opin Neurobiol 2009;19(6):621—8.

[94] Shen W, Flajolet M, Greengard P, Surmeier DJ. Dichotomous dopaminergic control of striatal synaptic plasticity. Science 2008; 321(5890):848—51.