CHAPTER

23

Neuronal Circuit Computation of Choice

Xiao-Jing Wang

OUTLINE

INTRODUCTION

Behavioral experiments using different types of task paradigms have led to two broad classes of mathematical models for decision making. On the one hand, sequential-sampling models describe information accumulation that unfolds in time and determine performance accuracy and reaction times in perceptual and memory tasks. On the other hand, game-theoretical models and reinforcement learning models account for dynamic choice behavior which is based on utility maximization and interplay with the environment or other decision agents. These models are important for quantitatively describing behavioral data and assessing theoretical ideas about the cognitive processes of decision making. To truly understand the biological basis of decision behavior, however, it is critical to construct realistic neural circuit models that allow us to uncover neural machineries and collective dynamics of neural networks in the brain underlying decision making. This has recently become possible thanks to advances in animal neurophysiology, human imaging and theory. This chapter summarizes recent progress in this direction. It discusses biological mechanisms and neural circuit models of choice behavior, and offers a unifying framework for both perceptual decision making (see also Chapter 19) and value-based choice behavior (see also Chapter 20) in terms of a recurrent neural circuit model endowed with reward-dependent synaptic plasticity.

MODELS OF DECISION MAKING

Drift-Diffusion, Leaky Competing Accumulator, and Neural Circuit Models

Sequential-sampling models are based on the intuitive idea that a decision is reached when stochastic

accumulation of information about alternative choices reaches a particular threshold. For two-alternative forced choice tasks, a specific implementation of particular importance is called the drift diffusion model (DDM) which is described in Chapters 3, 8, and 19 (Ratcliff, 1978; Smith and Ratcliff, 2004). In this model, an activity variable X represents the difference between the respective amounts of accumulated information about the two alternatives, say X_A and X_B, $X = X_A - X_B$. The dynamics of X are given by the drift diffusion equation,

$$\frac{dX}{dt} = \mu + w(t) \tag{23.1}$$

where μ is the drift rate, $w(t)$ is a white noise of zero mean and finite variance. The drift rate μ represents the bias in favor of one of the two choices (and is zero if there is no net bias). For instance, in a random-dot motion direction discrimination task (see Chapter 19), μ is proportional to the strength of motion signal. This system is a perfect integrator of the input:

$$X(t) = \mu t + \int^t w(t')dt' \tag{23.2}$$

The integration process is terminated and the decision time is read out, whenever $X(t)$ reaches a positive threshold (choice A) or a negative threshold (choice B). If the drift rate μ is positive, then choice A is correct, whereas choice B is an error. Therefore, this type of models is commonly referred to as ramping-to-threshold model, with the average ramping slope given by μ.

The DDM has been widely applied to fit behavioral data of perceptual and memory experiments as described in Chapter 3 (Ratcliff, 1978; Smith and Ratcliff, 2004). This model (as written here) is the continuous-time analog of the discrete-time Sequential Probability Ratio Test (SPRT), which is the optimal procedure for making binary choices under uncertainty, in the sense that it minimizes the mean decision time among all tests for a given lower bound of error rate (Bogacz *et al.*, 2006; Wald, 1948).

Can a ramping-to-threshold mechanism be instantiated by neural circuits? One key issue in answering that question is to determine the biological basis of time integration. The drift diffusion model is an ideal, perfect integrator (with an infinite time constant), whereas neurons and synapses are in actual fact "leaky" with short time constants of tens of milliseconds (Kandel *et al.*, 2012). Usher and McClelland (2001) extended the DDM by incorporating a leak so that the integrator becomes "forgetful" with a decay time constant, an issue discussed in Chapters 3 and 4. In that model, there is a competition between the two dynamical variables X_A and X_B through mutual inhibition. What is interesting is that this "leaky" competitive accumulator model has proven to fit many behavioral datasets as well as the drift diffusion model, provided that the integration time is sufficiently long, although the biological basis of this long time constant of integration remains unspecified.

It has been proposed that a long integration time can be realized in a decision neural network through recurrent interneuronal excitation (Wang, 2002). Reverberating excitation represents a salient characteristic of cortical local circuits that has been widely observed empirically (Douglas and Martin, 2004). When this positive feedback is sufficiently strong, recurrent excitation in interplay with synaptic inhibition can create multiple stable states (known as *attractors*) in a network. Models of this type were initially proposed for working memory, which is the brain's ability to actively hold information online in the absence of direct sensory stimulus (Wang, 2001). The same model, provided that excitatory reverberation is slow, has been shown to be capable of decision-making computations (Deco *et al.*, 2009; Engel and Wang, 2011; Machens *et al.*, 2005; Miller and Wang, 2006a; Wang, 2002, 2008; Wong and Wang, 2006). Interestingly, physiological studies in behaving nonhuman primates often report neural activity correlated with decision making in cortical areas, such as the prefrontal cortex or the parietal cortex, that also exhibit mnemonic persistent activity during working memory. Hence, this model and supporting experimental data suggest a common, "cognitive-type" circuit mechanism for decision making and working memory in the brain (Wang, 2013).

What is Spiking Network Modeling?

Physiological experiments in behaving animals are critical for uncovering neural signals correlated with specific aspects of decision making. Biophysically based neural modeling can delineate circuit mechanisms that give rise to the observed neural signals, and identify key computational principles at the conceptual level. For certain questions about decision making such as those discussed below, it is important to capture neural firing of action potentials or spikes (electrical signals often described mathematically as point processes; see Chapter 5) through which neurons transmit information and communicate with each other.

To this end, single cells can be described by a *spiking neuron model*, rather than a firing-rate model of the kind that have so far been presented in this volume. A popular choice for accomplishing this alternative kind of representational model is to employ either *the*

leaky integrate-and-fire model or *the Hodgkin–Huxley model*. Such a model is calibrated by physiological measurements, such as the electrical time constant of the nerve cell membrane and the input–output function (the spike firing rate as a function of the synaptic input current), which can be different for different classes of cells like *excitatory pyramidal cells* and *inhibitory interneurons*.

It is worth emphasizing that in a biophysically based model, synapses must also be modeled accurately. Unlike *connectionist* models in which coupling between neurons is typically an instantaneous function of firing activity, synapses have their own rise-time and decay time constant, and exhibit summation properties. That is an important property in this class of model because synaptic dynamics turn out to be a crucial factor in determining the integration time of a neural circuit dedicated to decision making, as well as controlling the stability of a strongly recurrent network. Once these "building blocks" (single cells and synapses) have been constructed for a particular model, they are used to construct a network endowed with a biologically plausible architecture. A commonly assumed circuit organization is local excitation between neurons of similar selectivity combined with a more global inhibition throughout the network. Dynamic balance between synaptic excitation and inhibition is another feature of cortical microcircuit that has been increasingly recognized experimentally and incorporated in cortical network models.

A Recurrent Circuit Mechanism for Decision Making

A *neural circuit model* (NCM) for decisions with two alternative choices is schematically illustrated in Figure 23.1A (Wang, 2002, 2008; Wong and Wang, 2006). Two neural pools are selective for choice options (A or B), each consisting of a number of spiking neurons that are strongly connected with each other by excitatory synapses. The two neural pools compete with each other via shared inhibition. Conflicting and noisy evidence for two choice alternatives is described as the relative difference in the inputs (the differential input) to two neural groups, A and B, in a cortical decision circuit. Each neural group (say A) integrates input information over time, by virtue of quasi-linear stochastic ramping activity for hundreds of milliseconds, which is faster (with a larger ramping slope) when the evidence is stronger for option A. The two neural groups compete through feedback inhibition from interneurons so that, eventually, one of them wins and rises (red, Figure 23.1B), whereas the other loses and decays away (blue, Figure 23.1B). Whichever (A or B) ramps up to a particular activity level triggers an all-or-none neural signal downstream, which leads to a categorical behavioral response.

The NCM can be viewed in two different ways. In contrast to the temporal plots of neural activity (Figure 23.1B), one can portray the dynamics of a decision circuit in a so-called *state space*, where the firing rates of neural pools selective for different options are plotted against each other (Figure 23.1C). According to this view, different choices are represented by distinct attractor states. The mathematical term attractor here simply means a dynamical system state which is stable against small perturbations. An attractor does not have to be a steady state but can be a complex spatiotemporal pattern. And it is important to note that a system's attractor landscape is not necessarily rigidly fixed. Any relatively sustained input (external stimulus or top-down cognitive control signal) readily alters the attractor landscape in the state space (Figure 23.1C left versus right panels).

Neural Substrate of a Decision Threshold

Numerous monkey experiments (Chapter 19) have revealed ramping-to-threshold neural activity at the single cell level that is correlated with perceptual decision (Gold and Shadlen, 2007; Roitman and Shadlen, 2002) and action selection (Hanes and Schall, 1996; Schall, 2001). How can a decision threshold be instantiated by neurons, rather than prescribed in an *ad hoc* manner? One natural hypothesis is that, when decision neurons integrate inputs and reach a particular firing rate level, this event triggers an all-or-none response in downstream neurons and leads to the generation of a behavioral output. This idea was tested for oculomotor decision tasks in which the motor response is a rapid saccadic eye movement. In an extended, two-stage circuit model (Lo and Wang, 2006), decision neurons in the cortex (as described above) project to movement neurons in the superior colliculus (SC), an important command center for saccades (Figure 23.2A). This model also includes a direct pathway in the basal ganglia, with an input layer (caudate, CD) and an output layer (substantia nigra reticulata, SNr). As a neural pool in the cortex ramps up in time, so do their synaptic inputs to the corresponding pool of SC movement neurons. When this input exceeds a well-defined threshold level, an all-or-none burst of spikes is triggered in the movement cells, signaling a particular (A or B) motor output. In this scenario, a decision threshold (as a bound of firing rate of decision neurons) is instantiated by a hard threshold of synaptic input for downstream motor neurons. Figure 23.2B

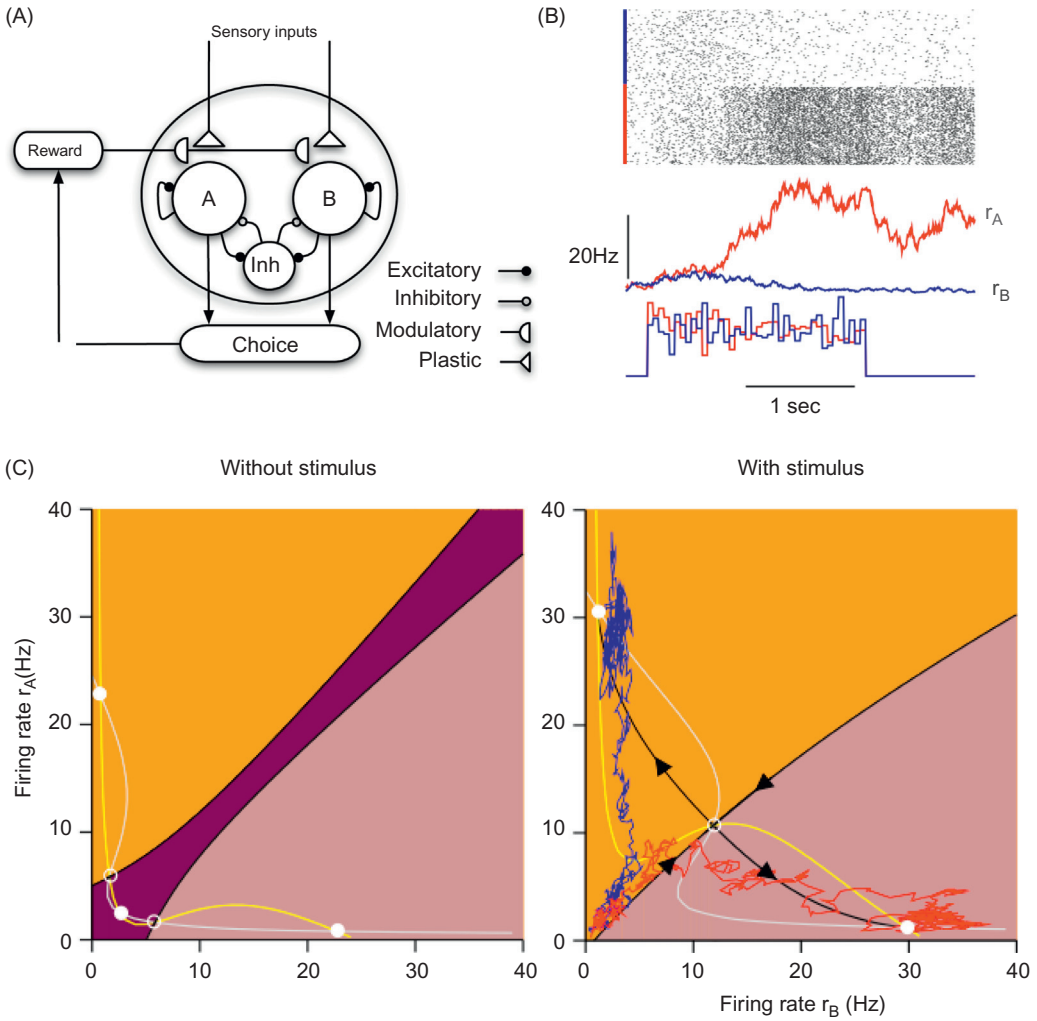

FIGURE 23.1 (A) A *neural circuit model* (NCM) for deicison making with two-alternatives. There are two pools of excitatory neurons, each of which is selective to one of the two choice options A and B. Within each pool there are strong recurrent excitatory connections that can sustain persistent activity triggered by a transient preferred stimulus. The two neural pools compete through feedback inhibition from interneurons. Depending on the task design, one of the choices may be rewarded with some probability, whereas the other may not, in any given trial. The outcome signal (reward or not) is assumed to modulate Hebbian plasticity of input synapses c_A and c_B. Since the network's decision dynamics depends on c_A and c_B, altered synaptic strengths lead to adaptive choice behavior across trials. (B) Two neural populations selective for different choices display graded ramping followed by winner-take-all competition, in a simulation of motion direction discrimination task. Top: spike trains of single neurons in the two competing neural pools A and B; middle: population firing rate r_A and r_B as a function of time; bottom: inputs to the two neural pools. (C) The population dynamics of a NCM is displayed in the state space of firing rates r_A and r_B without external input (left panel) and in the presence of a motion stimulus (right panel). Note that the attractor landscape sensitively depends on the input (left versus right panel). *Adapted with permission from Wang (2002, 2008).*

shows a sample trial of such a model simulation. The rate of ramping activity fluctuates from trial-to-trial, as a result of stochastic firing dynamics in the cortex, and is inversely related to the decision time (as defined by the time when a burst of action potentials is triggered in the SC) on a trial-by-trial basis (Figure 23.2C). When the task is more difficult, ramping activity is slower, leading to longer reaction times. However, the threshold of cortical firing activity that is read out by the downstream motion system has the same narrow distribution, regardless of the ramping speed or reaction times (Lo and Wang, 2006). Therefore, this model realizes a robust threshold detection mechanism, and the variability of reaction times is mostly attributed to the irregular ramping of neural activity itself rather than a stochastic decision bound. With this implementation of a decision threshold, the model can produce quantitative behavioral metrics such as accuracy (psychometric function) and reaction time (Figure 23.2D) that can be compared with experimental measurements.

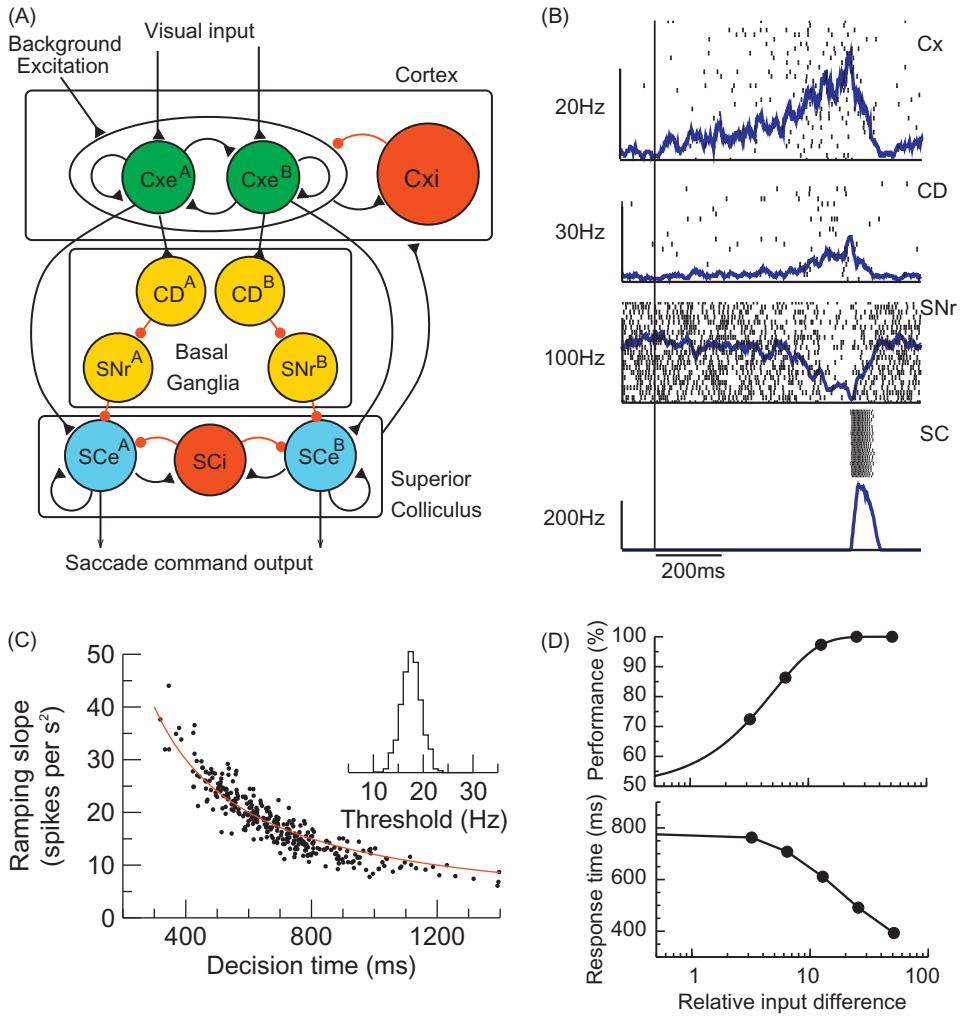

FIGURE 23.2 Decision making in a multiple-module neural circuit. (A) Schematic architecture of the model for two-alternative forced-choice oculomotor tasks. Neural pools in the cortical network integrate sensory information in favor of two choice options A and B, and compete against each other. They project to both the superior colliculus (SC) and the caudate nucleus (CD) in the basal ganglia. CD sends inhibitory projection to the substantia nigra pars reticulata (SNr) which through inhibitory synapses connect with premotor neurons in the SC. Each population consists of noisy spiking neurons. (B) A single trial simulation of the model, showing spike trains from single cells and population firing rates of Cxe, SNr, and CD, and SCe. A burst of spikes in premotor neurons (SCe) is triggered when their synaptic inputs exceed a threshold level, which results from both direct excitation by cortical neurons, and disinhibition from SNr via the cortico-striatal projection. Time zero corresponds to stimulus onset. (C) The ramping slope of Cxe firing rate is inversely related to decision time on a trial-by-trial basis (each data point corresponds to an individual trial). The red curve is 12,000/(decision time). (D) Performance (percentage of correct choices) and mean response time as a function of the differential input, the relative difference in the mean inputs to the cortical neural pools A and B. *Adapted with permission from Lo and Wang (2006).*

This model has been applied to a monkey experiment using a visual motion direction discrimination task (see Chapter 19). In that experiment, the subject was shown a display of moving random dots, a fraction of which moved coherently in one of two possible directions (say A = left, B = right), and the remaining dots moved in random directions. The task difficulty was varied from trial to trial by varying the motion coherence (0–100%). In monkeys performing this task, single neurons in the lateral intraparietal (LIP) cortex were found to exhibit slow ramping activity that is correlated with the perceptual decision about the direction (leftward or rightward) of the motion stimulus (Gold and Shadlen, 2007). At lower motion coherence, the subject's reaction time was longer, and the ramping of LIP neuronal firing rate was slower but reached the same firing activity level at the time when the behavioral response was produced, regardless of the motion coherence (Roitman and Shadlen, 2002). Thus, LIP neurons display a ramping-to-threshold process at the cellular level. Our neural circuit model successfully simulated this monkey experiment, with the

motion coherence given by the relative input strength. This model reproduces the monkeys performance and reaction times, as well as salient physiological data of LIP neurons (Lo and Wang, 2006; Wang, 2002; Wong *et al.*, 2007).

Speed–Accuracy Trade-Off

How can a decision threshold be adaptively tuned in this circuit? For instance, in a speed–accuracy trade-off, too low a threshold leads to quicker responses but more errors, whereas too high a threshold improves the accuracy but prolongs response times. Neither of these yields maximal rewards. Since in this model the decision threshold is defined as the minimum cortical firing needed to induce a burst response in the downstream SC neurons, one would expect that this threshold could be adjusted by plastic changes in the cortico-collicular pathway: the same level of cortical input to the superior colliculus could be achieved with less firing of cortical neurons, if the synapses of the cortico-collicular projection are stronger. Interestingly, this is not the case when the system is gated by the basal ganglia. This is because neurons in SNr normally fire tonically at a high rate (Figure 23.2B), and provide a sustained inhibition to SC movement neurons (Hikosaka *et al.*, 2000). This inhibition must be released, as ramping activity in the cortex activates CD neurons, which in turn suppresses the activity in the SNr, in order for SC neurons to produce a burst of action potentials as an output. This highly nonlinear disinhibition mechanism implies that the decision threshold is much more readily adjustable by tuning the synaptic strength of the cortico-striatal pathway (Lo and Wang, 2006). Indeed, such an adaptive tuning of decision threshold is expected to depend on reward signals (Reynolds *et al.*, 2001), and cortico-striatal synapses represent a major target of innervations by dopamine neurons which play a critical role in reinforcement signaling (Reynolds and Wickens, 2002). Our work suggests that the dopamine-dependent plasticity of cortico-striatal synapses is a likely neural locus for adaptive tuning of the decision threshold in the brain.

It should be noted that synaptic plasticity takes a long time to affect decision making across trials. On the other hand, we are able to adjust speed versus accuracy almost instantaneously, for example by an instruction at the beginning of each individual trial. It has been shown that, actually, a constant input readily affects speed and accuracy (Furman and Wang, 2008). This input could correspond to a top-down control signal in the brain. Interestingly, if such a control signal projects to both excitatory and inhibitory neurons in a decision circuit in a balanced way, then it can instantiate speed–accuracy trade-off by adjusting the slope of neural ramping activity (Lo and Wang, 2009). Indeed, a recent monkey experiment has shown that single neurons in the frontal cortex reduced the ramping slope when subjects traded speed in favor of accuracy (Heitz and Schall, 2012). Human studies using functional MRI suggest that both the prefrontal cortex and striatum have been implicated in speed–accuracy trade-off (Bogacz *et al.*, 2010; Forstmann *et al.*, 2008). More refined task designs could differentiate distinct brain mechanisms operating over disparate timescales for learning versus top-down control as suggested by the modeling work.

Comparison Between the Drift Diffusion Model and Neural Circuit Model

How does the neural circuit model compare with the drift diffusion model? First, one should note that they are two quite different levels of abstraction. The DDM assumes an infinite integration time; whereas NCM proposes a long but finite integration time. A possible neural basis of a long integration time is the NMDA receptor dependent recurrent synaptic excitation. Second, the functional benefit of time integration was demonstrated in the model by showing that performance improves when the system is allowed to integrate inputs over a longer time, but eventually plateaus with sufficiently long integration as the system reaches an attractor state representing a categorical choice (Wang, 2002; Figure 23.1C, right panel). This prediction was confirmed in a recent monkey experiment (Kiani *et al.*, 2008). Third, whereas in DDM evidence shown at different time points has equal weight, NCM asserts that evidence available early on has a larger impact on the ultimate choice than evidence presented later and immediately before a decision is made. This NCM prediction was supported in an experiment where a brief pulse of sensory information was introduced at different time points (Huk and Shadlen, 2005; Wong *et al.*, 2007). However, in more general situations when sensory data or attention varies continuously in time, information provided a long time ago may be forgotten, and a commitment may be reversed in the face of newly presented evidence (Resulaj *et al.*, 2009). The biological basis and possible fundamental limitation of integration time in decision making remains an outstanding subject of future research.

The NCM is a nonlinear dynamical system capable of more than one mode of operation. Indeed, in a so-called *jumping mode* neurons could show a sudden jump of firing rate instead of a smooth quasi-linear

time course, but the time at which the discrete jump occurs may vary from trial-to-trial randomly so that the trial-averaged neural activity still displays smooth ramping dynamics (Deco *et al*., 2007, 2009; Gigante *et al*., 2009; Lo and Wang, 2009; Miller and Katz, 2010; Miller and Wang, 2006c; Okamoto *et al*., 2007; Wang, 2012). The two (ramping and jumping) modes can be realized in the same model with modest variations of parameters, suggesting that they could occur in different local circuits of the brain or under different conditions in a single area.

Notably, in the jumping mode, without noise the system would remain in the resting state, therefore fluctuations are required for decision making. The sources of noise or stochasticity in a decision process have only begun to be examined experimentally (Brunton *et al*., 2013). Perceptual decisions (identification, discrimination, etc.) are often hard because sensory information is noisy, and integration of sensory data over time is computationally desirable because it improves signal-to-noise ratio (Luce, 1986). However, there is also stochasticity intrinsic to a decision circuit, and the *Fano factor* (the ratio of the variance versus mean of spike counts) of neural integrators may itself increase over time (Miller and Wang, 2006b; Churchland *et al*., 2011). This is likely to be generally true for neural circuits involved in both perceptual decisions and value-based choices, and stochastic neural dynamics of decision systems may play a critical role in indeterminacy of decision behavior (Glimcher, 2005; Wang, 2008).

ADAPTIVE VALUE-BASED CHOICE

A Decision-Making Circuit Endowed with Reward-Dependent Learning

In the NCM, decisions are made by stochastic neural dynamics in any given trial. Across many trials, the probability of choosing A (i.e., the fraction of trials when the neural pool selective for option A wins the competition through attractor dynamics) is in effect the psychometric function (Figure 23.2D, upper panel), which can be described by a softmax function of the difference in the strengths (c_A and c_B) of inputs to the two competing neural pools (Soltani and Wang, 2006):

$$P_A(c_A - c_B) = 1/(1 + \exp(-(c_A - c_B)/\sigma)) \qquad (23.3)$$

where σ expresses the amount of stochasticity due to irregular spike firing in the network and also depends on other model properties such as firing rates of input neurons. Importantly, a softmax decision criterion is widely assumed in more abstract models of choice behavior; indeed it is the same equation used in the reinforcement learning model for fitting monkey and human behavioral data (see Chapter 26 as well as Section 3 of this volume). The neural circuit modeling lends support to this general assumption, and sheds insights into its underlying stochastic recurrent neural dynamics.

In order to account for the trial-by-trial learning in adaptive choice behavior, reward-dependent learning can be incorporated into this class of model. Suppose that input synaptic connections c_A and c_B are plastic, then synaptic modifications will alter the networks future decision behavior, which in turn will lead to further changes in the synapses (Figure 23.1A). For instance, if in a trial the choice is correct (say A), a positive outcome might trigger dopamine release that leads to a potentiation of c_A. As a result, in the next trial the probability for choosing A will be enhanced.

One working hypothesis is that input synapses onto a decision circuit are up-dated according to such a reward-dependent Hebbian learning rule (see also Seung, 2003). For this purpose a number of studies have used binary synapses (Amit and Fusi, 1994; Fusi, 2002) that undergo a Hebbian learning rule, namely that synaptic plasticity depends on coactivation of presynaptic and postsynaptic neurons (Hebb, 1949). Specifically, synapses between two neurons are assumed to have two (Down and Up) states, and c_A (respectively c_B) is the fraction of synapses from an input neuron to a decision neuron in the pool A (respectively B) that are in the Up state.

In such a model it is assumed that synapses for inputs to decision neurons are potentiated only if the choice is rewarded, and depressed otherwise (Fusi *et al*. 2007; Soltani and Wang, 2006; Soltani *et al*., 2006). If A wins in a trial, implying that the firing rate is high for decision neural pool A and low for pool B, only c_A undergoes a plastic change, whereas c_B remains the same. If the choice is correct, yielding a reward, then c_A is potentiated according to

$$c_A = c_A + q_+(1 - c_A); \qquad (23.4)$$

if A is incorrect and no reward is delivered, c_A is depressed according to

$$c_A = c_A - q_- c_A \qquad (23.5)$$

where q_+ and q_- are the learning rates. Their inverses are the time constants with which the system keeps the memory trace for past reward and non reward outcomes. Note that these simple equations ensure that c_A remains positive and between 0 and 1. As a result of synaptic modifications, the input strengths for the competing neural groups of the decision network vary from trial to trial, leading to adaptive dynamics of choice behavior.

Dopamine and Synaptic Plasticity

The above reward-dependent learning rule is broadly supported by neurophysiological data. Dopamine, which plays an important role in reward-related signaling (see Chapters 15–18), can reverse the sign of plasticity (from depression to potentiation) at cortico-striatal synapses (Reynolds *et al.*, 2001) and synapses on prefrontal neurons (Matsuda *et al.*, 2006; Xu and Yao, 2010). This finding has recently been refined with the use of stimulation protocols that induce *spike-timing dependent plasticity* (STDP) (Bi and Poo, 2001; Dan and Poo, 2006). STDP refers to the fact that Hebbian synaptic modification depends on the relative timing of presynaptic and postsynaptic spikes: with positive spike timing (the presynaptic neuron of a connected pair fires first before the postsynaptic neuron by less than tens of milliseconds, therefore can contribute to the generation of postsynaptic spiking), potentiation is induced; whereas with negative timing (the presynaptic spike does not affect the postsynaptic firing) depression occurs (Figure 23.3A). The synaptic

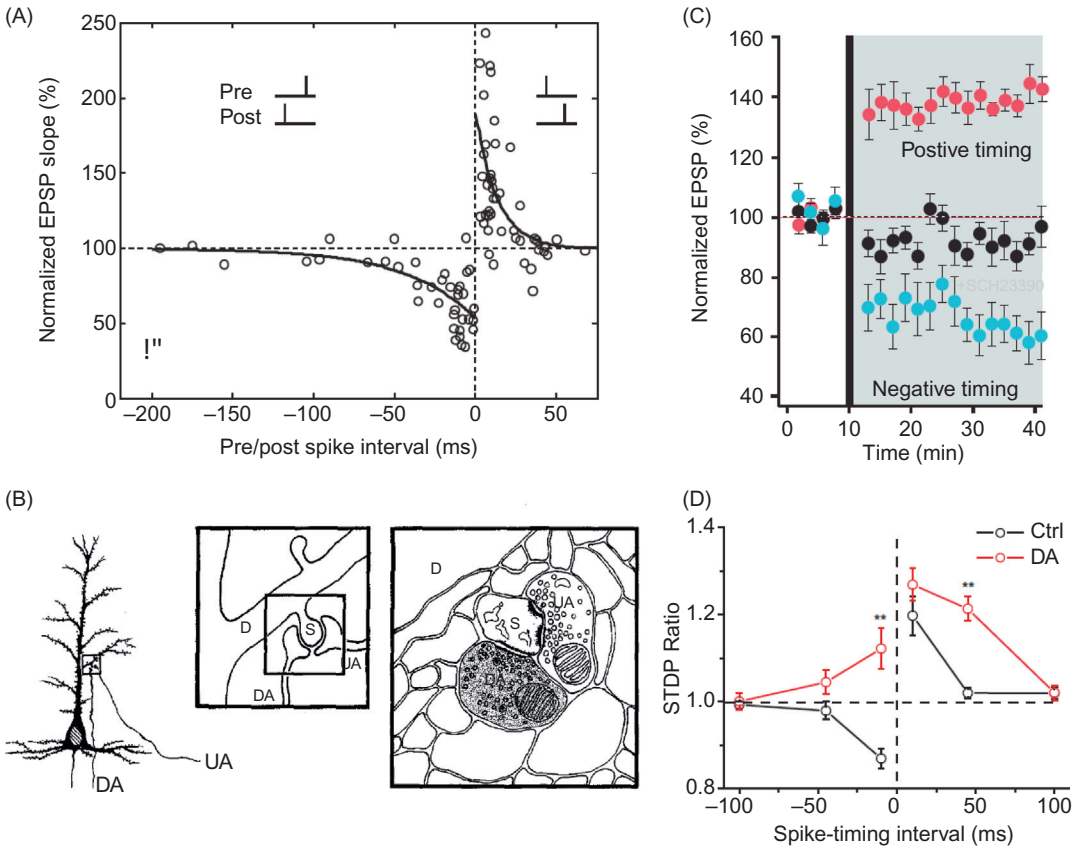

FIGURE 23.3 Dopamine modulation of synaptic plasticity. (A) Spike timing dependent plasticity (STDP). Synaptic modification induced by repetitively paired pre- and postsynaptic spikes in layer 2/3 of visual cortical slices from the rat. Each symbol represents result from one experiment. Curves are single exponential, least-squares fits of the data. Insets depict the sequence of spiking in the pre- and postsynaptic neurons. The slope of EPSP (excitatory postsynaptic potential) is a measure of synaptic strength, which is potentiated above baseline (more than 100%) when the induction protocol used positive spike timing, and depressed (less than 100%) with negative spike timing. (B) Triad arrangement involving the dopamine input to the cortex. Left: afferents labeled with a dopamine- (DA) specific antibody terminate on the spine of a pyramidal cell in the prefrontal cortex, together with an unidentified axon (UA). Middle: enlargement of axospinous synapses illustrated in the left panel. Right: diagram of ultrastructural features of the axospinous synapses illustrated in middle panel; the dopamine terminal (darkened profile representing DA immunoreactivity) forms a symmetrical synapse; the unidentified profile forms an asymmetrical synapse with the postsynaptic membrane. *Adapted with permission from Goldman-Rakic (1995) with data published in Goldman-Rakic* et al. *(1989).* (C) Dopamine gates the sign of plasticity for cortical synapses on D1 receptor-expressing medium spiny neurons in the striatum. Positive spike timing produces long-term potentiation (red), whereas negative timing does not induce plastic changes (black). When D1 receptors are blocked by SCH23390, negative timing induced long-term depression is unmasked (blue). *Adapted with permission from Surmeier* et al. *(2010) with data published in Shen* et al. *(2008).* (D) Dopamine alters the STDP window in hippocampal neurons. STDP window in control conditions (black circles) and when dopamine was present during the STDP induction protocol (red circles). With positive spike timing, dopamine allowed for longer intervals between spike and synaptic activation to induce potentiation of synaptic strength. With negative spike timing, dopamine enabled potentiation induction with a protocol that induced depression under control conditions. *Adapted with permission from Zhang* et al. *(2009).*

triad arrangement (synapse colocalizing with dopamine input) at cortico-striatal synapses (Freund *et al.*, 1984; Surmeier *et al.*, 2010) and excitatory synapses onto prefrontal neurons (Figure 23.3B; Goldman-Rakic, 1995) suggest that dopamine can potently modulate synaptic plasticity. Indeed, it has been found (Shen *et al.*, 2008) that at cortico-striatal synapses, in the presence of dopamine D1 receptors, positive timing leads to potentiation (Figure 23.3C, red) and negative timing does not induce synaptic change (Figure 23.3C, black). When D1 receptors are blocked pharmacologically, however, negative spike timing yields depression (Figure 23.3C, blue). In hippocampal neurons, bath application of dopamine enlarged the temporal window for potentiation with positive spike timing and converted depression to potentiation with negative spike timing (Figure 23.3D; Zhang *et al.*, 2009). Taken together, these experimental results are consistent with the modeling proposal that dopamine activation (presumably mediated by D1 receptors) can reserve the signal of synaptic modification. This is worth noting that other neuromodulators (such as noradrenaline; Seol *et al.*, 2007) can also alter synaptic modification or reverse its sign (reviewed in Pawlak *et al.*, 2010).

Are there general mathematical models for reward-dependent learning rules? The aforementioned learning rule is simple and turns out to be validated by its applications to a number of adaptive processes (see below). Various reward-dependent learning rules have been proposed, where plasticity is gated by either reward or *reward prediction error* (RPE) (Frémaux *et al.*, 2010; Izhikevich, 2007; Legenstein *et al.*, 2010; Loewenstein and Seung, 2006; Pfeiffer *et al.*, 2010). RPE, of course, plays a key role in reinforcement learning theory (see Chapters 15 and 16; Dayan and Abbott, 2001; Rutledge *et al.*, 2010; Sutton and Barto, 1998), and phasic spiking activity of dopamine neurons is known to resemble an RPE signal (Bayer and Glimcher, 2005; Montague *et al.*, 1996; Schultz, 1998; Schultz *et al.*, 1997). It is presently unclear whether the existing models are fundamentally different, or they are essentially similar under different mathematical forms. For instance, it has been shown that the covariance of neural activity and reward is a common denominator of several reward dependent learning rules (Frémaux *et al.*, 2010; Loewenstein and Seung, 2006). Also, theoretical work suggests that RPE must be distinct for each task and stimulus (Frémaux *et al.*, 2010), whether that holds true and how that might be realized by the dopamine system remain unclear. A related question is exactly what dopamine neurons compute and how their computations depend on subcortical (Bromberg-Martin *et al.*, 2010) and prefrontal (Takahashi *et al.*, 2011) inputs, and intrinsic circuit properties within the ventral segmental area (Cohen *et al.*, 2012). Furthermore, a major open issue is concerned with the so-called *eligibility trace* of Chapter 16 linking an action and its reward outcome that are temporally separated (Dayan and Abbott, 2001; Izhikevich, 2007; Sutton and Barto, 1998). The biological substrate of such eligibility trace remains uncertain.

Computation of Returns by Synapses: Matching Law Through Melioration

The NCM described here endowed with such three-factor synaptic plasticity is a general one rather than designed for a particular task. This model has been further tested by applying it to a foraging task, in which a subject makes successive choices adaptively in a stochastic environment (Lau and Glimcher, 2005, 2007; Sugrue *et al.*, 2004). In these tasks, whether a subject's choice yields reward or not depends on the stochastic environment. In either case, the model simulates a decision maker whose choice outcomes lead to synaptic plasticity that in turn influences future choices, thereby learning to forage adaptively.

In foraging tasks commonly used in laboratories, rewards are delivered to two response options stochastically at baiting rates λ_A and λ_B, respectively, according to a particular concurrent reinforcement schedule (Herrnstein *et al.*, 1997; Lau and Glimcher, 2005; Sugrue *et al.*, 2004). Behavioral studies using this task have led to Herrnstein's matching law, which states that a subject allocates her or his choices in a proportion which matches the relative reinforcement obtained from these choices (Herrnstein *et al.*, 1997). Moreover, the spiking activity of neurons in the lateral intraparietal cortex (LIP) is modulated by a representation of value that was defined as fractional income (Sugrue *et al.*, 2004). To explore a cortical circuit mechanism of matching behavior, one can endow this neural circuit model of decision with reward-dependent synaptic plasticity. As shown in Figure 23.4A, B, the model applied to the foraging task reproduces the matching behavior observed in the monkey experiment. As the reward rate λ_A/λ_B varies from one block of trials to the next block, the choice behavior of the model changes quickly, so that the probability of choosing A versus B matches approximately λ_A/λ_B. It has been shown analytically that the synaptic strengths (c_A and c_B) are proportional to the returns (reward per choice) rather than income (the amount of reward per unit time) of the two targets, namely $c_A \simeq R_A$ and $c_B \simeq R_B$.

Figure 23.4C shows the probability of choosing option A (P_A) along with the input synaptic strengths (c_A and c_B) across six blocks of trials. The process of synaptic plasticity is stochastic, and there is

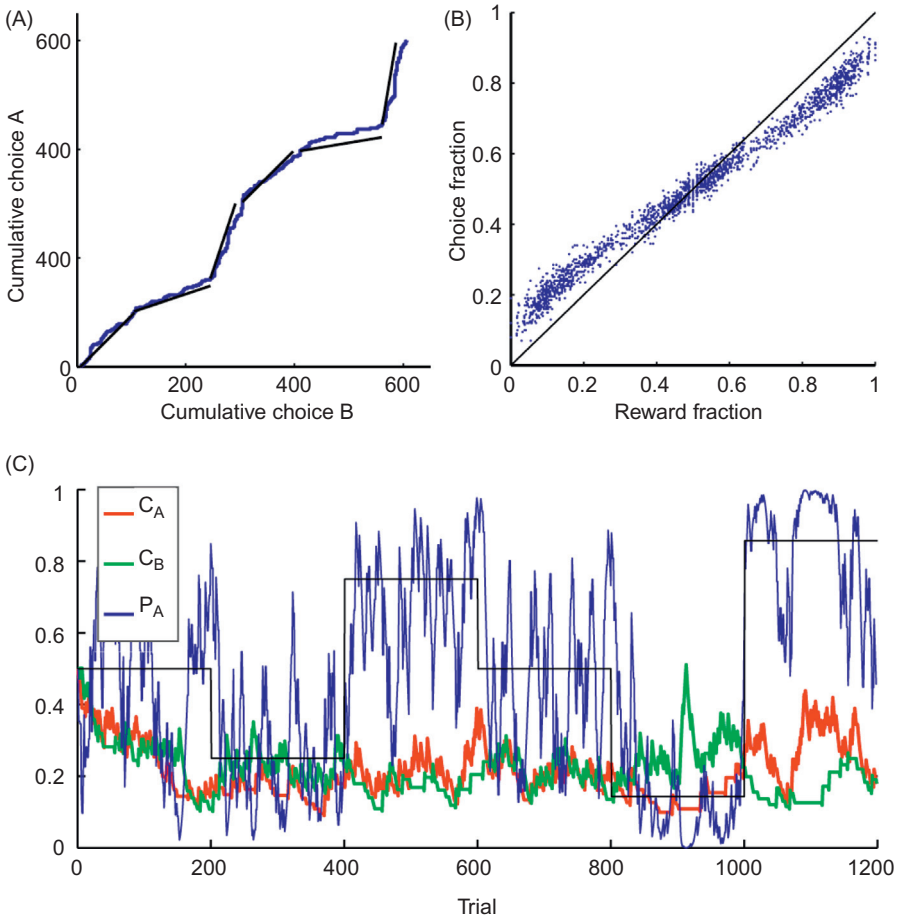

FIGURE 23.4 A neural circuit model shows matching behavior in a dynamic environment. (A) For one session of the simulated matching experiment, the cumulative choice probability for target A is plotted against the cumulative choice probability for target B. Black straight lines show the baiting probability ratio in each block. The slope of the cumulative plot (the choice ratio) is approximately equal to the baiting probability ratio. In this session the following baiting probability ratios are used in sequence [1:1, 1:3, 3:1, 1:1, 1:6, 6:1]. (B) Each point shows the block-wise choice fraction as a function of the block-wise reward fraction for a block of trials on which the baiting probabilities are held constant. Most of the points fall close to the diagonal line (perfect matching), but the choice fraction is slightly lower than the reward fraction when the latter is larger than 1/2 (under-matching). (C) The synaptic strengths, c_A (red) and c_B (green), and the choice probability (blue) plotted as a function of time. The thin black line indicates the baiting probability ratio in each block. In each block the synaptic strengths fluctuate according to the returns from the two choices (not shown). *Adapted with permission from Soltani and Wang (2006).*

considerable variability within each block of 200 trials. However, on average (indicated by the blue line for P_A), the choice probability ratio matches that of rates at which rewards are delivered to the two targets, and this matching behavior is learned through plastic synapses. For instance, if in a block of trials, the reward probability λ_A is larger than λ_B then c_A is more likely to be potentiated than c_B through the successive decisions of the network across trials because the return from choosing A is higher, leading to a larger P_A. The converse occurs in a block of trials where λ_B is larger than λ_A.

Note that synaptic modifications take place on a trial-by-trial basis, locally in time. Moreover, synapses are forgetful and behave like a leaky integrator of past choice outcomes. In our model, synaptic integration of past rewards has a time constant of a few trials, and therefore the decision behavior is influenced only by rewards harvested locally in time, in agreement with behavioral (Lau and Glimcher, 2005; Sugrue *et al.*, 2004, 2005) and neurophysiological (Seo and Lee, 2007; Seo *et al.*, 2007) observations. There is no prescription in the model for global optimization (Bogacz and Larsen, 2011; Sakai and Fukai, 2008). The models performance is close to the matching behavior, which is achieved dynamically through a so-called *melioration process*, i.e., the model chooses the alternative with a higher return, so that the interplay between decision behavior and synaptic plasticity iteratively improves the total income (reward per unit time) to the maximum possible, given the constraints of the stochastic neuronal and synaptic dynamics. The model also reproduces the observation that in the monkey experiment, matching is not perfect, and the relative probability of choosing the more rewarding option is slightly smaller than the relative reward rate (*under-matching*) (Figure 23.4B). A model analysis explained this finding, revealing that under-matching is a natural consequence of stochasticity in neural activity (Soltani and Wang, 2006).

Furthermore, because neural activity depends on input strengths, the model naturally reproduces the experimental observation that neural activity in LIP is parametrically modulated by the values of the choice options (Figure 23.5; Soltani and Wang, 2006). The implication is that, although activity of LIP neurons depends on values of response options, valuation may occur elsewhere, possibly at the synaptic level and in

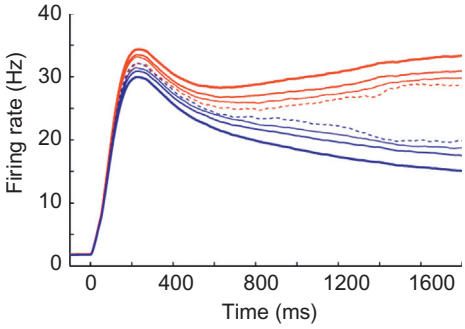

FIGURE 23.5 Graded activity of model neurons as a function of the input synaptic strengths which encode the values (returns) of choice options in a matching task. The activity of decision neurons shows a graded pattern, if single-trial firing rates are sorted and averaged according to the network's choice and the difference between synaptic strengths. Activity is aligned by the onset of two targets and it is shown separately for the choices corresponding to the neurons' preferred (red) or non-preferred (blue) target. In addition, trials are subdivided into four groups according to the difference between the strength of synapses to the two competing neural populations [$c_A - c_B = -0.16$ to -0.05 (dashed), -0.05 to 0 (thin), 0 to 0.05 (medium), 0.05 to 0.16 (thick)]. *Reproduced with permission from Soltani and Wang (2006).*

the form of returns. For the sake of simplicity, Soltani and colleagues considered a local network model, but importantly they remained agnostic about the actual site of synaptic plasticity that is critically involved with valuation. Candidate loci include the cortico-striatal connections in the basal ganglia (Lo and Wang, 2006), or synaptic pathways within the orbitofrontal cortex (see Chapter 13).

Random Choice Behavior in Matching Pennies Game

This class of models has also been extended to decision making in competitive games between multiple agents (introduced in Chapter 2). The idea captured by this line of research is that several such models, each simulating a "decision maker," can interact according to a payoff matrix. This class of model can thus be used to simulate monkey experiments using game-theoretic tasks (Chapter 26; Barraclough *et al.*, 2004; Dorris and Glimcher, 2004), in which monkeys play matching pennies with a computer opponent that uses three different algorithms (0, 1 and 2, see Chapter 26). The model reproduces many salient behavioral observations (Soltani *et al.*, 2006). If the opponent is not interactive (using Algorithm 0), the model decision behavior is idiosyncratic, and might, for instance, choose one of the targets exclusively. When the opponent uses algorithm 1, the model exhibits prominent *win-stay-lose-switch* (WSLS) behavior, as observed in monkeys. Finally, when the opponent uses algorithm 2 and is fully interactive according to the rules of matching pennies, the model behavior becomes quasi-random. This is shown in Figure 23.6, with several different sets of initial values for the synaptic variables c_A and c_B (Figure 23.6, left panel). Different c_A and c_B values yield different initial probability P_A of choosing response A versus B (Figure 23.6, right panel). Competitive interaction with the opponent, however, quickly equalizes the synaptic variables (Figure 23.6, left panel), and the choice probability becomes very close to 0.5 (Figure 23.6, right panel), regardless of the initial state of the system. For instance, if initially the system chooses target A more frequently because c_A is larger than c_B, it would be exploited by the opponent, and the unrewarded outcomes from choosing A induce depression of c_A of the synapses to the neural pool A, so that the difference $c_A - c_B$ decreases over time, and the system gradually chooses B more frequently.

Interestingly, the model, with a reinforcement learning rule that changes only synapses onto neurons selective for the chosen option, does not capture all the details of the monkeys behavior. In particular, it shows a probability of WSLS, P(WSLS), below a limited value (about 0.65), whereas P(WLSL) can be nearly 1 in monkeys with algorithm 1. Further studies have revealed that $P(WLSL) \simeq 1$ can be realized in this model with a different learning rule, according to which synapses onto both neural populations (selective for the chosen and unchosen targets) are modified in each trial. This is akin to a *belief-dependent learning rule* (discussed in Chapter 25; Camerer, 2003; Lee *et al.*, 2005).

Although the model can reproduce monkey behavior obtained with different opponent-algorithms, different model parameters are required for each algorithm. How can these model parameters be tuned adaptively, as the opponents algorithm is changed? To address this question, Soltani and colleagues (2006) incorporated a meta-learning rule proposed by Schweighofer and Doya (2003) that maximizes long-term rewards. They found that the enhanced model captures the very slow changes of the monkey's behavior, as the opponents algorithm changes from session to session.

A general insight that can be drawn from this work is that a decision circuit produces random choice behavior, not necessarily because the system has a prescribed "random number generator," but because the trial-to-trial choice dynamics forces the decision agent to play randomly. This is well demonstrated in the model, because the same model produces either stereotypical responses or random responses, depending on the behavior of its opponent. The model decision maker thus does not have a goal to play randomly, but simply tries to play at its best, given the environment and other decision agent(s) involved in the game. This conclusion

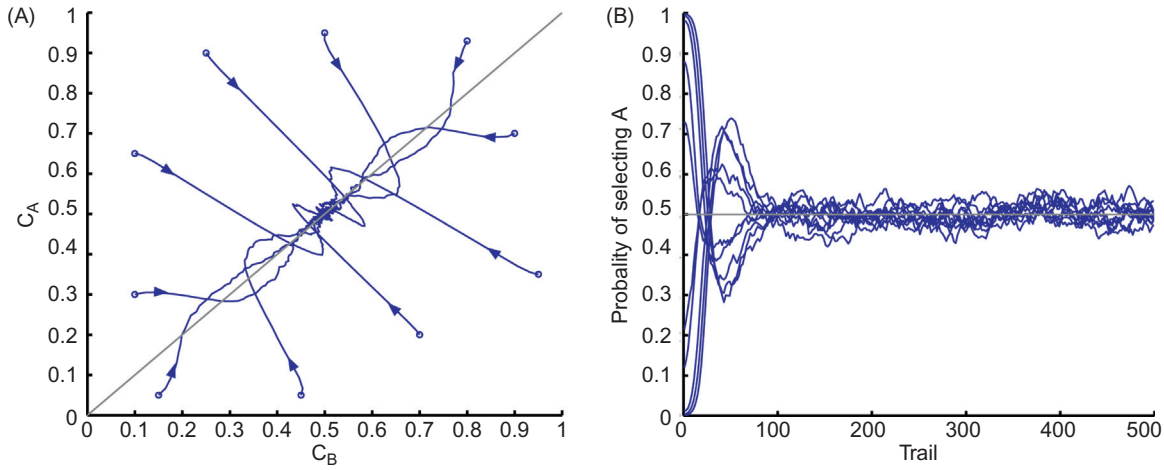

FIGURE 23.6 Model simulation of the dynamic choice behavior during the matching pennies game. (A) The synaptic strengths c_A and c_B plotted against each other show the time evolution of adaptive learning. (B) The corresponding choice probability for A, which is a softmax function of $c_A - c_B$. Regardless of the initial state of the neural circuit (different c_A and c_B values), the network quickly evolves towards the Nash equilibrium state of random mixed strategy, with $c_A \simeq c_B$ (the diagonal line in A), and the choice probability becomes chance level (0.5 in B).

is consistent with previous behavioral studies and models, emphasizing the critical importance of feedback in the production of quasi-random behavior (Camerer, 2003; Rapoport and Budescu, 1992). Moreover, the model suggests that irregular neural firing that gives rise to sigmoid decision criterion, and the stochastic nature of synaptic learning, contribute to the generation of random choice behavior, which can be desirable and even optimal in interactive decision tasks. Thus, this model sheds insights into neural processes in the brain underlying the randomness observed at the psychological level (Glimcher, 2005; Wang, 2008). In this way, neurobiologically based neural modeling helps to bridge the gap between cognitive behavior and its underlying neural network mechanisms.

Reward Memory and Reinforcement Learning on Multiple Time Scales

As mentioned above, the synaptic learning rule described above assumes certain time constant(s) with which the system integrates past reward events, and the memory trace decays away in the absence of reward delivery. This is generally the case for reinforcement learning models. In a simple model, a variable V represents the value of certain action, which is updated as

$$V(t+1) = V(t) + \alpha\delta \tag{23.6}$$

where α is a learning rate, and $\delta = r - V$ is RPE (the difference between the actual reward r and the expected reward V). The inverse of α is a time constant τ. For instance, in the absence of reward delivery, $V(t)$ decays over time exponentially as $exp(-t/\tau)$.

Intuitively, the learning rate α should be dynamically adjustable: if the environment is stochastic but stable, then it is desirable to deploy a long integration time in order to learn about and exploit the statistics of the environment; whereas if the environment is highly uncertain, one should use a short time constant and high learning rate to explore different options quickly. Indeed, in a human experiment where the volatility of reward delivery statistics is systematically varied and the learning rate of human subjects was estimated by fitting behavioral data with a mathematical model, it was found that the estimated learning rate is higher when the environment is more unpredictable (Behrens *et al.*, 2007).

Can such a time constant for integrating past reward events be extracted from single cells in decision making? By developing a novel data analysis, Bernacchia and colleagues (2011) analyzed how rewards in previous trials affect the firing activity of neurons in the dorsolateral prefrontal cortex, anterior cingulate cortex and intraparietal cortex from monkeys performing the matching pennies task. Surprisingly, they was found that the histogram of time constant (τ) extracted from about 800 individual neurons display a power law like $\sim 1/\tau^2$, whereas the history of the memory trace amplitude (A) behaves lawfully as $exp(-A)$ (Figure 23.7, upper panels). The power law tail of the time constant distribution means that very long time constants have a much higher probability than if the distribution is Gaussian or exponential.

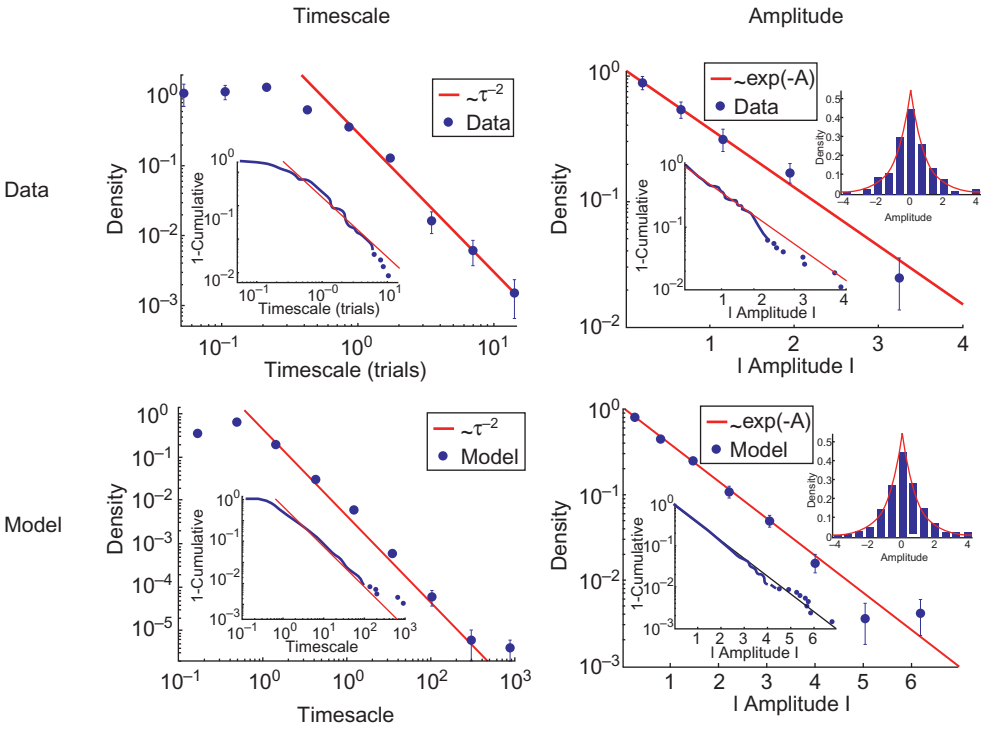

FIGURE 23.7 Neuronal representation of reward memory traces on a wide range of timescales. Upper panel: the histograms of time constants (left) and memory amplitudes (right) extracted from single neurons in the prefrontal and parietal cortices in monkeys performing a decision task. Insert: cumulative histograms. The time constant distribution exhibits a power law tail, and the amplitude distribution is exponential. Lower panel: the histograms of time constants (left) and memory amplitudes (right) in a high-dimensional and recurrent linear network of neurons. *Adapted with permission from Bernacchia* et al. *(2011).*

The implication is that a "reservoir" of time constants are heterogeneously distributed in prefrontal and parietal neurons, and that the reward memory system operates in a high dimensional space. Indeed, Bernacchia and colleagues showed that a simple linear neural network model in an infinite dimensional space, under certain conditions (such as "at the edge of chaos"), reproduces the same power law distribution of time constant and exponential distribution of memory trace (Figure 23.7, lower panels). Further experimental and modeling work is needed in order to assess the validity of this model. Regardless, these results suggest that reinforcement learning modeling should be extended to more than one dimension, in order to allow for reward-dependent learning over many timescales. It is worth mentioning that reinforcement learning has been extended to a hierarchical structure with an inherently high dimensionality (Botvinick *et al.*, 2009). We propose that such a system is dynamical and endowed with a very broad range of time constants. Therefore, in principle, a readout system could deploy short or long time constants (high or low learning rates) from this reservoir, exibly, depending on the degree of volatility of the environment.

Probabilistic Inference

The same framework of a decision circuit endowed with reward-dependent learning has been applied to other decision processes, such as arbitrary sensorimotor mapping (Asaad *et al.*, 1998; Fusi *et al.*, 2007; Wise and Murray, 2000) and pattern matching decisions (Engel and Wang, 2011). Unexpectedly, the model was also found to be capable of statistical calculations at the core of probabilistic inference. In a *weather prediction task*, several (for example four) sensory cues (s_i, $i = 1, 2, 3, 4$) are shown, each is associated with a *weight of evidence* (WOEs), defined by *log likelihood ratios* (LRs) log $P(s_i|A)/log\ P(s_i|B)$, that one of the two outcomes A (rain) and B (shine) is true (Gluck and Bower, 1988; Knowlton *et al.*, 1994). When the prior $p(A) = p(B)$, this is also log posterior ratio, but real-life situations involve unequal priors, or "base rates." The subject is required to make a decision ("rain" or "shine") based on the combined evidence, the sum of WOEs of four cues presented in a single trial (Gluck *et al.*, 2002). How can such a quantity as *summated log likelihood ratio* or *log posterior ratio* be actually computed in the brain? Using the reward-dependent learning rule described above, it was found that summing log posterior odds can be

readily realized, through approximations, by plastic synapses (provided that synapses are bounded) in a decision circuit (Soltani and Wang, 2010). Specifically, one can show that, according to our three-factor learning rule, synaptic strength c_A and c_B from sensory neurons encoding stimulus s to decision neurons A and B compute posteriors $p(A|s)$ and $p(B|s)$, respectively. Moreover, recall that the decision circuit generates choices in such a way that the probability of selecting A is a softmax function of the differential input, $c_A - c_B = p(A|s) - p(B|s) = 2p(A|s) - 1$ (with $p(B|s) = 1 - p(A|s)$). Mathematically, $x - (1-x) \simeq \log(x/(1-x))$, if $0.2 \leq x \leq 0.8$. It follows that, for the intermediate range of posteriors where the model's choice behavior is stochastic, the difference in the synaptic strengths is linearly proportional to the log posterior ratio. For smaller or larger values of posteriors, the choice behavior is deterministic (the probability of choosing A is close to 0 or 1). As a result, decision making is based on log posterior ratio, as required by probabilistic inference.

When several cues are presented to inform a decision, log posterior ratios for the presented stimuli are readily added by virtue of convergence of cue-encoding neurons to decision neurons (Figure 23.8A). Therefore, a decision circuit endowed with such synapses makes choices on the basis of the summed log posterior ratios and performs near-optimal cue combination. This model was validated by reproducing not only behavioral performance of monkeys of the Yang and Shadlen experiment (Yang and Shadlen, 2007; Figure 23.8B), but also single-neuron physiological data recorded from behaving monkeys (Figure 23.8C-D).

Another study (Pfeiffer *et al.*, 2010) considered an ideal three-factor Bayesian–Hebb rule that was designed to yield synaptic weights w_i equal to the log likelihood ratio. They found that updating w_i requires an exponential function of w_i. However, when the exponential function is approximately linearized, the learning rule becomes precisely the same as that of (Soltani and Wang, 2010). Moreover, it was shown that the linearized Bayesian–Hebb rule performs nearly as well as the theoretical optimum in a number of benchmark tasks, including the weather prediction task.

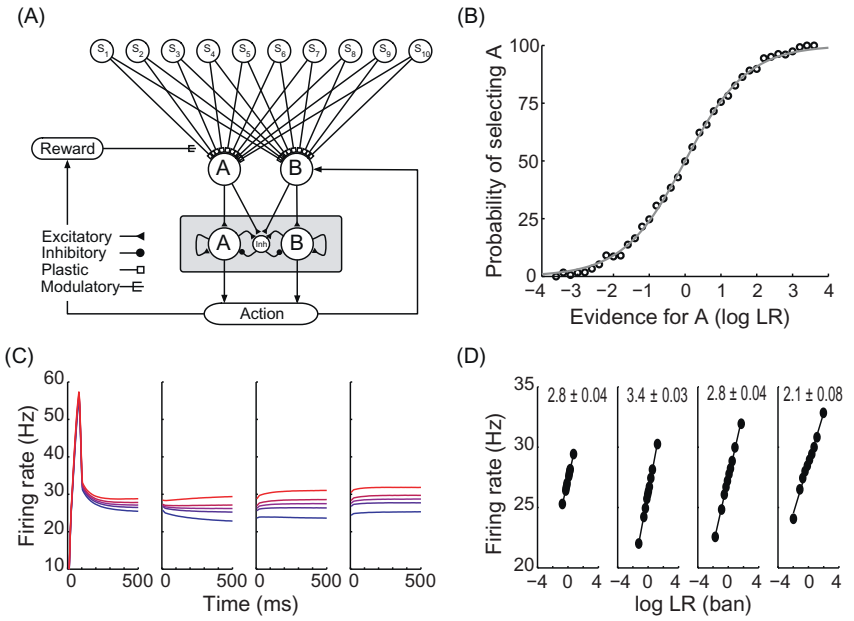

FIGURE 23.8 Probabilistic inference in a decision-making circuit endowed with reinforcement learning. (A) Schematic of the three-layer model for a weather prediction task. The first layer consists of cue-selective neural populations, each is activated upon the presentation of a cue. The sensory cue-selective neurons provide, through synapses that undergo reward-dependent Hebbian plasticity, inputs to two neural populations in an intermediate layer that encode reward values of two choice alternatives (action values). Combination of cues is accomplished through convergence of cue-selective neurons onto action value encoding neurons. The latter project to a decision-making circuit (gray box, same as the cortical circuit in Figure 23.1A). The choice (*A* or *B*) is determined by which of the two decision neural populations wins competition on a trial. Depending on the reward schedule, a chosen action may be rewarded or not. The presence (respectively absence) of a modulatory reward signal leads to potentiation (respectively depression) of plastic synapses. (B) Choice behavior of the model in the weather prediction task. Probability of choosing *A* as a function of the sensory evidence favoring this option, defined by the sum of log LR of four cues. (C) Effect of the log LR on the firing rate of model neurons. Five traces are plotted for five quintiles of the log LR in that epoch (more red means larger log LR favoring alternative A). The log LR in each epoch is equal to the sum of the log LR of shapes that are presented before and during that epoch. (D) Average population firing rate as a function of the log LR (in base 10) for four epochs. *Adapted with permission from Soltani and Wang (2010).*

Furthermore, when the choice alternatives have unequal priors, the model predicts deviations from the Bayes' decision rule that are akin to an effect called "base-rate neglect" commonly observed in human studies, namely a cue that is equally predictive of each outcome is perceived to be more predictive of the less probable outcome (Soltani and Wang, 2010). Therefore, our model might be sufficiently general to describe more complex probabilistic problem solving.

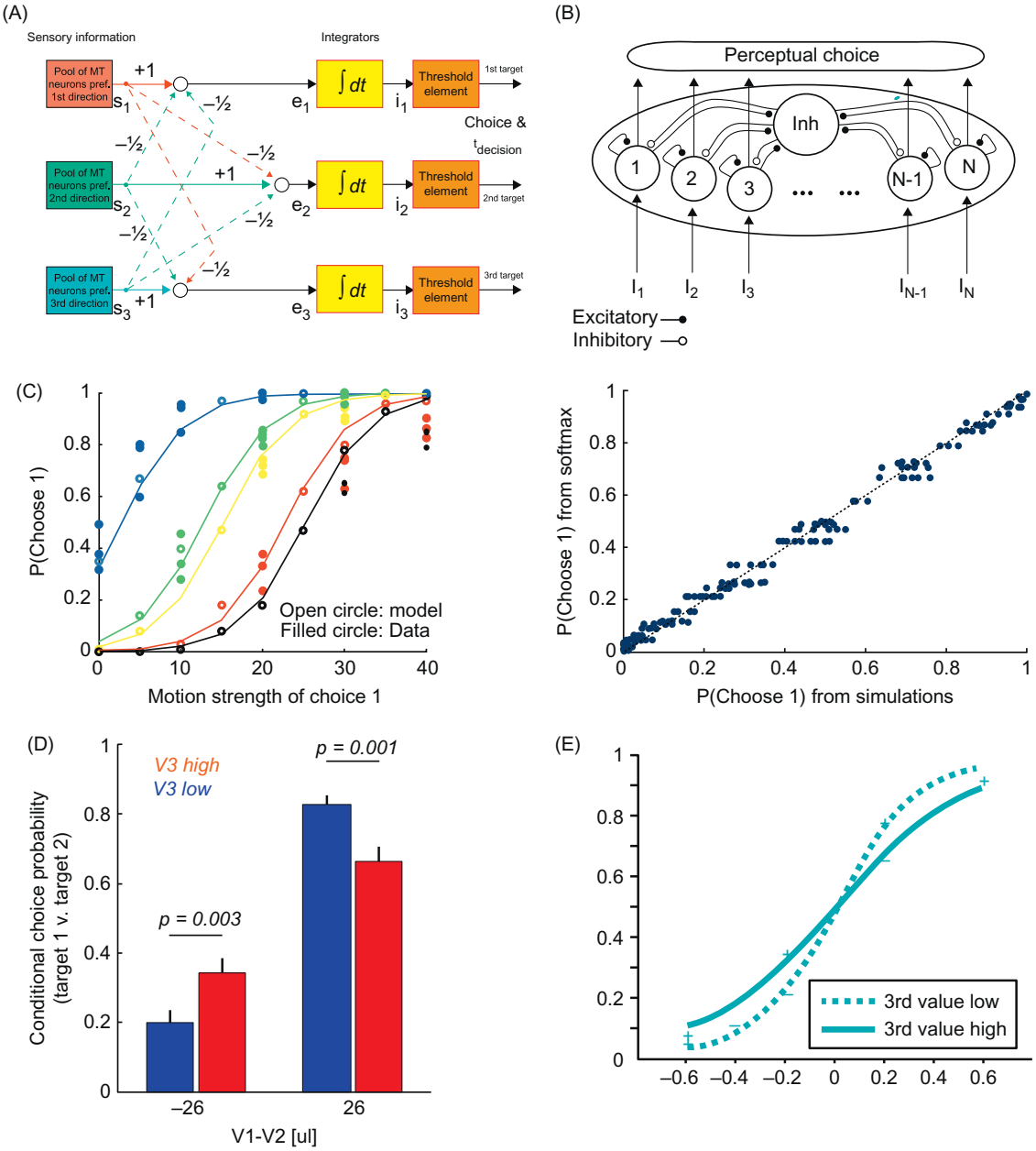

FIGURE 23.9 Deviation from theory of rationality in three-choice decision making. (A) A proposed generalization of DDM to three-choice. *Adapted from Niwa and Ditterich (2008) with permission.* (B) NCM for multiple choice. (C) Left: simulation results (open circles) of a three-choice NCM are well fitted with a softmax function (filled circle). See also right panel for comparison. (D) In a value-based choice task (Louie and Glimcher, 2011), three options are offered in the order of values (1: best, 2: second best, 3: worst). According to normative decision theory, option 3 should be irrelevant and changing its value should not influence the relative probability of choosing option 1 among the first two options P(1)/(P(1) + P(2)). In contrast to this ideal optimality, in the monkey experiment a higher value for option 3 reduces the relative probability for choosing the best of the two better options, which is inconsistent with the softmax decision criterion. *Figure kindly provided by K. Louie and P. Glimcher.* (E) Similar finding as in (D) in another monkey experiment, when medial orbitofrontal cortex was lesioned. *Adapted from Noonan et al. (2010) with permission.*

Deviation from Rational Behavior: An Example

The notion of rational behavior is linked to that of optimality, but often times what constitutes an optimal strategy for a given decision task is unclear. Whereas SPRT is optimal for two-alternative forced choice tasks, optimal tests for three or more options are not known. There are multiple ways to generalize the DDM to multi-alternative choice (Churchland *et al.*, 2008; Krajbich and Rangel, 2011; McMillen and Holmes, 2006; Niwa and Ditterich, 2008), one of them is shown in Figure 23.9A compared to a generalized NCM to muti-alternative decision making (Albantakis and Deco, 2009; Furman and Wang, 2008; Smith and Wang, 2008; Usher and McClelland, 2001) shown in Figure 23.9B. As shown in Figure 23.9C, the decision behavior of a three-choice version of the attractor network model can be well described by a softmax function, $P(1) = \exp(\sigma V_1)/(\exp(\sigma V_1) + \exp(\sigma V_2) + \exp(\sigma V_3))$, where V1, V2, and V3 are the values or strengths of evidence for the three options, and σ is a parameter that quantifies the amount of stochasticity. This model (Smith and Wang, 2008) fits well with human performance data from a 3-choice visual motion direction discrimination experiment (Niwa and Ditterich, 2008).

One prediction of the softmax decision criterion is that the relative probability of choosing one of two options (say 1 and 2), $P(1 \mid 1 + 2) = P(1)/(P(1) + P(2))$, is independent of the strength of the third option (V_3). This prediction has been shown to be contradicted by observed economic choice behavior in a surprising way. In one experiment using both monkeys and humans, three choice options were associated with different reward values (1: best, 2: second best, 3; Louie and Glimcher, 2011; Louie *et al.*, 2012). The third option has a lower value than both the first and second options, thus is irrelevant and should be ignored. Yet, when the value for the worst option 3 was increased (while remaining lower than options 1 and 2), subjects reduced the relative probability for choosing the best of the two better options, contrary to normative models of rational behavior (Figure 23.9D). Similar findings were reported in another monkey experiment, but only when the medial orbitofrontal cortex was lesioned (Figure 23.9E; Noonan *et al.*, 2010). Why this deviation from rational behavior was found in normal subjects in one experiment, yet only in animals with brain damage in another experiment, needs to be elucidated in future studies.

Interestingly, deviations from rational behavior in the monkey experiment of (Louie and Glimcher, 2011; Louie *et al.*, 2012) can be concisely accounted for with the assumption that the neural circuit is endowed with divisive normalization, namely the activity of a neuron is divided by the sum of its neighboring neurons. Divisive normalization has been widely observed in a number of cortical circuits (Carandini and Heeger, 2011). Therefore, this combined approach using monkey behavior, physiology and model demonstrated how a neural circuit mechanism predicts behavioral trends that are not anticipated nor easily explained by optimality-based theories.

CONCLUSION

Much of the research in behavioral economics focuses on how the decision makers choose among various options when the information about the uncertain future prospects are provided explicitly. For example, in studies on decision making under risk, the decision makers are given specific information about the magnitudes and probabilities of possible payoffs from each choice. Given the knowledge, one should devise a behavioral strategy to strive for global optimality. In real life, however, information about the magnitude, likelihood, and temporal delay of reward and punishment resulting from a particular choice often has to be estimated through experience by trial and error. Furthermore, such reward contingencies often change over time, and this happens frequently when multiple agents interact. Recent findings summarized here and in other chapters suggest that adaptive choice behavior is more dynamical, through choice-by-choice melioration; and that memory of past reward events is leaky both at the behavioral and neuronal levels. Nevertheless, the brain is endowed with a reservoir of disparate time constants for reward memory hence potentially reinforcement learning, which is functionally desirable for dynamical exploitation–exploration trade-off.

Neurophysiological experiments with behaving animals and computational work have begun to establish an empirically well tested core neural circuit model, that is characterized by strongly recurrent or attractor dynamics and endowed with reward-dependent Hebbian synaptic plasticity. This model has been successfully applied to perceptual decision making, foraging, flexible sensori-motor mapping, competitive game, and probabilistic causal learning. These studies provide important clues as to how adaptive stochastic decision making, such as matching behavior in a foraging task, approximate Nash equilibrium in a competitive game or probabilistic inference, result from a dynamic interplay between a decision-making network and its environment. The model will need to be extended to investigate how a neural network or a system of networks can suitably combine the information about various aspects of reward and punishment, such as their magnitude, uncertainty, and temporal delay.

Also, the biophysical basis of reward-dependent plasticity in the brain remains to be fully elucidated. Recent work illustrated how deviations from optimality might

be naturally explained by known neural mechanisms, future research along these lines will shed fundamental insights into the discrepancy between the behaviors of humans and animals and the theory of rational choice.

Acknowledgments

I would like to thank Alireza Soltani, Kong-Fatt Wong, Stefano Fusi, Chung-Chuan Lo, Alberto Bernacchia, Nathaniel Smith, Daeyeol Lee and Hyojung Seo for their contributions; Kenway Louie and Paul Glimcher for providing Figure 23.9D. This work was supported by the NIH grant MH062349.

References

Albantakis, L., Deco, G., 2009. The encoding of alternatives in multiple-choice decision making. Proc. Natl. Acad. Sci. U.S.A. 106, 10308–10313.

Amit, D.J., Fusi, S., 1994. Dynamic learning in neural networks with material synapses. Neural Comp. 6, 957–982.

Asaad, W.F., Rainer, G., Miller, E.K., 1998. Neural activity in the primate prefrontal cortex during associative learning. Neuron. 21, 1399–1407.

Barraclough, D.J., Conroy, M.L., Lee, D., 2004. Prefrontal cortex and decision making in a mixed-strategy game. Nat. Neurosci. 7, 404–410.

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

Behrens, T., Woolrich, M., Walton, M., Rushworth, M., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–1221.

Bernacchia, A., Seo, H., Lee, D., Wang, X.-J., 2011. A reservoir of time constants for memory traces in cortical neurons. Nat. Neurosci. 14, 366–372.

Bi, G., Poo, M., 2001. Synaptic modification by correlated activity: Hebb's postulate revisited. Annu. Rev. Neurosci. 24, 139–166.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., Cohen, J.D., 2006. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychol. Rev. 113, 700–765.

Bogacz, R., Larsen, T., 2011. Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. Neural Comput. 23, 817–851.

Bogacz, R., Wagenmakers, E.J., Forstmann, B.U., Nieuwenhuis, S., 2010. The neural basis of the speed-accuracy tradeoff. Trends Neurosci. 33, 10–16.

Botvinick, M.M., Niv, Y., Barto, A.C., 2009. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition. 113, 262–280.

Bromberg-Martin, E.S., Matsumoto, M., Hikosaka, O., 2010. Distinct tonic and phasic anticipatory activity in lateral habenula and dopamine neurons. Neuron. 67, 144–155.

Brunton, B.W., Botvinick, M.M., Brody, C.D., 2013. Rats and humans can optimally accumulate evidence for decision-making. Science. 340, 95–98.

Camerer, C.F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton Unversity Press, Princeton, NJ.

Carandini, M., Heeger, D.J., 2011. Normalization as a canonical neural computation. Nat. Rev. Neurosci. 13, 51–62.

Churchland, A.K., Kiani, R., Chaudhuri, R., Wang, X.-J., Pouget, A., Shadlen, M.N., 2011. Variance as a signature of neural computations during decision making. Neuron. 69, 818–831.

Churchland, A., Kiani, R., Shadlen, M., 2008. Decision-making with multiple alternatives. Nat. Neurosci. 11, 693–702.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., Uchida, N., 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature. 482, 85–88.

Dan, Y., Poo, M.M., 2006. Spike timing-dependent plasticity: from synapse to perception. Physiol. Rev. 86, 1033–1048.

Dayan, P., Abbott, L.F., 2001. Theoretical Neuroscience. MIT Press, Cambridge MA.

Deco, G., Péerez-Sanagustín, M., de Lafuente, V., Romo, R., 2007. Perceptual detection as a dynamical bistability phenomenon: a neurocomputational correlate of sensation. Proc. Natl. Acad. Sci. U.S.A. 104, 20073–20077.

Deco, G., Rolls, E.T., Romo, R., 2009. Stochastic dynamics as a principle of brain function. Prog. Neurobiol. 88, 1–16.

Dorris, M.C., Glimcher, P.W., 2004. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron. 44, 365–378.

Douglas, R.J., Martin, K.A.C., 2004. Neuronal circuits of the neocortex. Annu. Rev. Neurosci. 27, 419–451.

Engel, T.A., Wang, X.-J., 2011. Same or different? A neural circuit mechanism of similarity-based pattern match decision making. J. Neurosci. 31, 6982–6996.

Forstmann, B.U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K.R., et al., 2008. Striatum and pre-SMA facilitate decision-making under time pressure. Proc. Natl. Acad. Sci. U.S.A. 105, 17538–17542.

Frémaux, N., Sprekeler, H., Gerstner, W., 2010. Functional requirements for reward-modulated spike-timing-dependent plasticity. J. Neurosci. 30, 13326–13337.

Freund, T.F., Powell, J.F., Smith, A.D., 1984. Tyrosine hydroxylase-immunoreactive boutons in synaptic contact with identified striatonigral neurons, with particular reference to dendritic spines. Neuroscience. 13, 1189–1215.

Furman, M., Wang, X.-J., 2008. Similarity effect and optimal control of multiple-choice decision making. Neuron. 60, 1153–1168.

Fusi, S., 2002. Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. Biol. Cybern. 87, 459–470.

Fusi, S., Asaad, W.F., Miller, E.K., Wang, X.-J., 2007. A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. Neuron. 54, 319–333.

Gigante, G., Mattia, M., Braun, J., Del Giudice, P., 2009. Bistable perception modeled as competing stochastic integrations at two levels. PLoS Comput. Biol. 5, e1000430.

Glimcher, P.W., 2005. Indeterminacy in brain and behavior. Annu. Rev. Psychol. 56, 25–56.

Gluck, M.A., Bower, G.H., 1988. From conditioning to category learning: an adaptive network model. J. Exp. Psychol. Gen. 117 (3), 227–247.

Gluck, M.A., Shohamy, D., Myers, C., 2002. How do people solve the "weather prediction" task? Individual variability in strategies for probabilistic category learning. Learn. Mem. 9, 408–418.

Gold, J.I., Shadlen, M.N., 2007. The neural basis of decision making. Annu. Rev. Neurosci. 30, 535–574.

Goldman-Rakic, P.S., 1995. Cellular basis of working memory. Neuron. 14, 477–485.

Goldman-Rakic, P.S., Leranth, C., Williams, S.M., Mons, N., Geffard, M., 1989. Dopamine synaptic complex with pyramidal neurons in primate cerebral cortex. Proc. Natl. Acad. Sci. U.S.A. 86, 9015–9019.

Hanes, D.P., Schall, J.D., 1996. Neural control of voluntary movement initiation. Science. 274, 427–430.

Hebb, D.O., 1949. Organization of Behavior. Wiley, New York.

Heitz, R.P., Schall, J.D., 2012. Neural mechanisms of speed–accuracy trade-off. Neuron. 76, 616–628.

Herrnstein, R.J., Rachlin, H., Laibson, D.I., 1997. The Matching Law: Papers in Psychology and Economics. Harvard Univ. Press, Cambridge.
Hikosaka, O., Takikawa, Y., Kawagoe, R., 2000. Role of the basal ganglia in the control of purposive saccadic eye movements. Physiol. Rev. 80, 953–978.
Huk, A.C., Shadlen, M.N., 2005. Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. J. Neurosci. 25, 10420–10436.
Izhikevich, E., 2007. Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb. Cortex. 17, 2443–2452.
Kandel, E.R., Schwartz, J.H., Jessell, T.M., Siegelbaum, S.A., Hudspeth, A.J., 2012. Principles of Neural Science, fifth ed. McGraw-Hill, New York.
Kiani, R., Hanks, T., Shadlen, M., 2008. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. J. Neurosci. 28, 3017–3029.
Knowlton, B.J., Squire, L.R., Gluck, M.A., 1994. Probabilistic classification learning in amnesia. Learn. Mem. 1, 106–120.
Krajbich, I., Rangel, A., 2011. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. Proc. Natl. Acad. Sci. U.S.A. 108, 13852–13857.
Lau, B., Glimcher, P., 2007. Action and outcome encoding in the primate caudate nucleus. J. Neurosci. 27, 14502–14514.
Lau, B., Glimcher, P.W., 2005. Dynamic response-by-response models of matching behavior in rhesus monkeys. J. Exp. Anal. Behav. 84, 555–579.
Lee, D., McGreevy, B., Barraclough, D., 2005. Learning and decision making in monkeys during a rock-paper-scissors game. Brain Res. Cogn. Brain Res. 25, 416–430.
Legenstein, R., Chase, S.M., Schwartz, A.B., Maass, W., 2010. A reward-modulated hebbian learning rule can explain experimentally observed network reorganization in a brain control task. J. Neurosci. 30, 8400–8410.
Lo, C.C., Wang, X.-J., 2006. Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. Nat. Neurosci. 9, 956–963.
Lo, C.C., Wang, X.-J., 2009. Functional tuning of a decision neural network by top-down balanced synaptic input produces skewed reaction time distributions with a long tail. In Society for Neuroscience Abstracts. 803.2.
Loewenstein, Y., Seung, H., 2006. Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. Proc. Natl. Acad. Sci. U.S.A. 103, 15224–15229.
Louie, K., Glimcher, P., 2011. Cortical normalization predicts stochastic choice behavior in value-guided decision-making. Soc. Neurosci. Abstr. 515.11.
Louie, K., Grattan, L.E., Glimcher, P.W., 2011. Reward value-based gain control: divisive normalization in parietal cortex. J. Neurosci. 31, 10627–10639.
Luce, R.D., 1986. Response Time: Their Role in Inferring Elementary Mental Organization. Oxford University Press, New York.
Machens, C.K., Romo, R., Brody, C.D., 2005. Flexible control of mutual inhibition: a neural model of two-interval discrimination. Science. 18, 1121–1124.
Matsuda, Y., Marzo, A., Otani, S., 2006. The presence of background dopamine signal converts long-term synaptic depression to potentiation in rat prefrontal cortex. J. Neurosci. 26, 4803–4810.
McMillen, T., Holmes, P., 2006. The dynamics of choice among multiple alternatives. J. Math. Psych. 50, 30–57.
Miller, P., Katz, D.B., 2010. Stochastic transitions between neural states in taste processing and decision-making. J. Neurosci. 30, 2559–2570.
Miller, P., Wang, X.-J., 2006a. Inhibitory control by an integral feedback signal in prefrontal cortex: a model of discrimination between sequential stimuli. Proc. Natl. Acad. Sci. U.S.A. 103, 201–206.
Miller, P., Wang, X.-J., 2006b. Power-law neuronal fluctuations in a recurrent network model of parametric working memory. J. Neurophysiol. 95, 1099–1114.
Miller, P., Wang, X.-J., 2006c. Stability of discrete memory states to stochastic fluctuations in neuronal systems. Chaos. 16, 026109.
Montague, P., Dayan, P., Sejnowski, T., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16, 1936–1947.
Niwa, M., Ditterich, J., 2008. Perceptual decisions between multiple directions of visual motion. J. Neurosci. 28, 4435–4445.
Noonan, M.P., Walton, M.E., Behrens, T.E., Sallet, J., Buckley, M.J., Rushworth, M.F., 2010. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. Proc. Natl. Acad. Sci. U.S.A. 107, 20547–20552.
Okamoto, H., Isomura, Y., Takada, M., Fukai, T., 2007. Temporal integration by stochastic recurrent network dynamics with bimodal neurons. J. Neurophysiol. 97, 3859–3867.
Pawlak, V., Wickens, J.R., Kirkwood, A., Kerr, J.N., 2010. Timing is not everything: neuromodulation opens the STDP gate. Front. Synaptic. Neurosci. 2, 146.
Pfeiffer, M., Nessler, B., Douglas, R.J., Maass, W., 2010. Reward-modulated Hebbian learning of decision making. Neural Comput. 22, 1399–1444.
Rapoport, A., Budescu, D.V., 1992. Generation of random series in two-person strictly competitive games. J. Exp. Psychol. Gen. 121, 352–363.
Ratcliff, R., 1978. A theory of memory retrieval. Psychol. Rev. 85, 59–108.
Resulaj, A., Kiani, R., Wolpert, D.M., Shadlen, M.N., 2009. Changes of mind in decision-making. Nature. 461, 263–266.
Reynolds, J.N., Wickens, J.R., 2002. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw. 15, 507–521.
Reynolds, J.N.J., Hyland, B.I., Wickens, J.R., 2001. A cellular mechanism of reward-related learning. Nature. 413, 67–70.
Roitman, J.D., Shadlen, M.N., 2002. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J. Neurosci. 22, 9475–9489.
Rutledge, R.B., Dean, M., Caplin, A., Glimcher, P.W., 2010. Testing the reward prediction error hypothesis with an axiomatic model. J. Neurosci. 30, 13525–13536.
Sakai, Y., Fukai, T., 2008. The actor-critic learning is behind the matching law: matching versus optimal behaviors. Neural Comput. 20, 227–251.
Schall, J.D., 2001. Neural basis of deciding, choosing and acting. Nature Neurosci. 2, 33–42.
Schultz, W., 1998. Predictive reward signal of dopamine neurons. J. Neurophysiol. 80, 1–27.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science. 275, 1593–1599.
Schweighofer, N., Doya, K., 2003. Meta-learning in reinforcement learning. Neural Netw. 16, 5–9.
Seo, H., Barraclough, D., Lee, D., 2007. Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. Cereb. Cortex. 17 (Suppl. 1), i110–117.
Seo, H., Lee, D., 2007. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J. Neurosci. 27, 8366–8377.
Seol, G.H., Ziburkus, J., Huang, S., Song, L., Kim, I.T., Takamiya, K., et al., 2007. Neuromodulators control the polarity of spike-timing-dependent synaptic plasticity. Neuron. 55, 919–929.
Seung, H., 2003. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron. 40, 1063–1073.

Shen, W., Flajolet, M., Greengard, P., Surmeier, D., 2008. Dichotomous dopaminergic control of striatal synaptic plasticity. Science. 321, 848–851.

Smith, N.J., Wang, X.-J., 2008. A reduced-variable biophysical model for multiple alternative decision-making. In Society for Neuroscience Abstracts. 871.23.

Smith, P.L., Ratcliff, R., 2004. Psychology and neurobiology of simple decisions. Trends Neurosci. 27, 161–168.

Soltani, A., Wang, X.-J., 2006. A biophysically based neural model of matching law behavior: melioration by stochastic synapses. J. Neurosci. 26, 3731–3744.

Soltani, A., Wang, X.-J, 2010. Synaptic computation underlying probabilistic inference. Nat. Neurosci. 13, 112–119.

Soltani, A., Lee, D., Wang, X.-J., 2006. Neural mechanism for stochastic behaviour during a competitive game. Neural Netw. 19, 1075–1090.

Sugrue, L.P., Corrado, G.C., Newsome, W.T., 2004. Matching behavior and representation of value in parietal cortex. Science. 304, 1782–1787.

Sugrue, L.P., Corrado, G.S., Newsome, W.T., 2005. Choosing the greater of two goods: neural currencies for valuation and decision making. Nat. Rev. Neurosci. 6, 363–375.

Surmeier, D.J., Shen, W., Day, M., Gertler, T., Chan, S., Tian, X., et al., 2010. The role of dopamine in modulating the structure and function of striatal circuits. Prog. Brain Res. 183, 149–167.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: an Introduction. MIT Press, Cambridge, MA.

Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., et al., 2011. Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat. Neurosci. 14, 1590–1597.

Usher, M., McClelland, J., 2001. On the time course of perceptual choice: the leaky competing accumulator model. Psychol. Rev. 108, 550–592.

Wald, A., 1948. On cumulative sums of random variables. Ann. Math. Stat. 15, 283–296.

Wang, X.-J., 2001. Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci. 24, 455–463.

Wang, X.-J., 2002. Probabilistic decision making by slow reverberation in cortical circuits. Neuron. 36, 955–968.

Wang, X.-J., 2008. Decision making in recurrent neuronal circuits. Neuron. 60, 215–234.

Wang, X.-J., 2012. Neural dynamics and circuit mechanisms of decision-making. Curr. Opin. Neurobiol. 22, 1039–1046.

Wang, X.-J., 2013. The prefrontal cortex as a quintessential 'cognitive-type' neural circuit: Working memory and decision making. In: Stuss, D.T., Knight, R.T. (Eds.), Principles of Frontal Lobe Function, second ed. Cambridge University Press, New York, pp. 226–248.

Wise, S.P., Murray, E.A., 2000. Arbitrary associations between antecedents and actions. Trends Neurosci. 23 (6), 271–276.

Wong, K.F., Huk, A.C., Shadlen, M.N., Wang, X.-J., 2007. Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision-making. Front. Comput. Neurosci. 1, 6. 10.3389/neuro.10/006.2007.

Wong, K.F., Wang, X.-J., 2006. A recurrent network mechanism of time integration in perceptual decisions. J. Neurosci. 26, 1314–1328.

Xu, T.X., Yao, W.D., 2010. D1 and D2 dopamine receptors in separate circuits cooperate to drive associative long-term potentiation in the prefrontal cortex. Proc. Natl. Acad. Sci. U.S.A. 107, 16366–16371.

Yang, T., Shadlen, M., 2007. Probabilistic reasoning by neurons. Nature. 447, 1075–1080.

Zhang, J.C., Lau, P.M., Bi, G.Q., 2009. Gain in sensitivity and loss in temporal contrast of STDP by dopaminergic modulation at hippocampal synapses. Proc. Natl. Acad. Sci. U.S.A. 106, 13028–13033.