Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Neurobiology

From biophysics to cognition: reward-dependent adaptive choice behavior

Alireza Soltani[1] and Xiao-Jing Wang[2]

In neurobiological studies of various cognitive abilities, neuroscientists use mathematical models to fit behavioral data from well-controlled experiments and look for neural activities that are correlated with parameters in those models. The pinpointed neural correlates are often taken as evidence that a given task is performed according to the prescription of the applied model, and the relevant brain areas encode parameters of such a model. However, to go beyond correlations toward causal understanding, it is necessary to elucidate at multiple levels the neural circuit mechanisms of cognitive processes. This review focuses on recent studies of reward-based decision-making that have begun to tackle this challenge.

Addresses
[1] Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA
[2] Department of Neurobiology and Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, CT 06520, USA

Corresponding author: Soltani, Alireza (soltani@caltech.edu) and Wang, Xiao-Jing (xjwang@yale.edu)

Current Opinion in Neurobiology 2008, **18**:209–216

This review comes from a themed issue on
Cognitive neuroscience
Edited by Read Montague and John Assad

Available online 21st August 2008

0959-4388/$ – see front matter
Published by Elsevier Ltd.

DOI 10.1016/j.conb.2008.07.003

Introduction

As it has become increasingly common to study high-level cognitive processes using neural recording from behaving animals and human brain imaging, perhaps we are entering a new era of consilience between the science of the brain and the science of the mind. This exciting trend has also raised many new challenges, among which is the issue of bridging the language of cellular neurophysiology and the language of cognitive psychology: how should major questions in these fields be redefined in common terms so that they can be rigorously investigated at both levels. Consider reward-based decision-making, the process of choosing from a set of alternatives that is informed by their respective expected rewards. This topic is at the forefront of neurobiology, cognitive science, and neuroeconomics [1,2•,3,4]. One common approach in the field is reinforcement learning (RL), of which a key concept is reward values. Unlike signals in sensory systems such as vision or olfaction, reward value signals are still not well defined in neuronal terms, and valuation processes involve many different systems (see other articles in this issue). Another approach is Bayesian inference that plays a key role in psychological decision theories and provides inspiration for neurobiological studies of decision-making [5]. Yet, it is notoriously difficult to experimentally demonstrate that humans or animals behave optimally or that neural populations represent probability densities, as predicted by Bayesianists.

To make progress, in our view, obstacles must be overcome to bridge gaps between levels of description, from behavior to neural systems, microcircuits, and cellular mechanisms. In this endeavor computational models play a critical role because they can be simultaneously studied at multiple levels in order to link observations at these different levels. This article aims at illustrating how such an approach is applied to reward-based decision-making. Specifically, we focus on two issues. First, inasmuch as a choice entails selection of an option among possible alternatives (e.g. behavioral responses), how are expected reward value and prior probability of choice alternatives learnt through a subject's encounters with the environment? Second, how are these different types of information represented and subsequently combined with sensory data to subserve a perceptual decision or action selection? This article will not cover human studies; instead our emphasis is on electrophysiological and modeling studies on the biophysical mechanisms of reward-based decision processes. By focusing on electrophysiological studies that quantitatively assess behavior and report activity of single neurons in behaving animals, biologically realistic modeling strives to elucidate neural circuit mechanisms that explain how the recorded neural activity is generated and how they give rise to the observed decision behavior.

Representation of reward value

At a given time, our deliberation on choices presented to us is influenced by our past experiences about the outcomes of similar choices that may have resulted in reward, no reward, or punishment. Therefore the important questions are how the values of previous rewards are encoded and integrated over time, and how this reward value representation precisely affects decision-making. According to RL models [6], each potential action is associated with an 'action value,' a measure of the reward expected from taking such an action. Action values are updated by a reward prediction error, the difference between the actual and the expected reward.

It was first discovered in primates that phasic activity of midbrain dopamine (DA) neurons represents a reward prediction error signal at the time an outcome is revealed [7]. Since then similar observations have also been made in rodents [8•]. Presently, the precise neural metrics of the expected reward is unknown, as is the nature of neural inputs received by the DA neurons that enable them to compute reward prediction error signals. By estimating value as a weighted average of rewards in previous trials, it was found that DA neurons predominantly encode the positive prediction error, that is, when the outcome is better than the expected value [9]. At the same time, the activity of DA neurons signals negative errors as well, albeit with a smaller change of neural activity. A recent work reported the finding that neurons in the lateral habenula signal negative reward prediction errors by an increased activity, and microstimulation of habenula inhibits DA neurons [10], suggesting that the lateral habenula may be a source of negative prediction error signals.

It has also been shown that DA neurons signal prediction errors in a context-dependent manner [11•] and adapt their gain according to the variance of rewards [12]. Moreover, neural activity correlated with both positive and negative prediction errors has also been observed in the medial frontal cortex, a cortical area involved in outcome monitoring [13•]. Unlike DA cells, both positive and negative prediction errors in the medial prefrontal cortex are signaled by an increase in firing activity, in two different neural populations.

A number of electrophysiological studies have investigated the neural representation of reward values [1,14]. One important issue is to differentiate between neural activities that reflect action values (specific for choice options), or their overall value (summated over all options and hence, insensitive to response choice). For example, in an experiment in which monkeys played an interactive game with a computer opponent, it was found that activity of neurons in the dorsolateral prefrontal cortex was modulated by choice, reward, and their combination in the previous and current trials [15]; whereas activity of neurons in the dorsal anterior cingulate cortex (ACCd) was modulated mostly by the sum of the values of alternative choices [16•].

Robust action value signals have been consistently reported in the striatum. In a monkey study using a stochastic choice task design, a subset of neurons in the striatum were found to encode the values of possible actions but were not selective for the chosen responses, for example, a neural firing activity was high if the value associated with action A was large, regardless of whether this option was actually chosen or not [17•]. In another stochastic decision task, during the delay period before the monkeys make choices, many neurons in the striatum displayed activity that was correlated with the value of each action [18]. Different types of neural activity in the striatum covaried with either action values before movement execution or the value of the chosen action after a behavioral response occurred [19].

Another important issue is how the brain assigns in a common currency, values to potential reward outcomes. The orbitofrontal cortex (OFC), which receives information from all sensory modalities and is extensively connected with the limbic system, appears to play a central role in this process [20]. In an economic choice task in which there was no 'correct response' and choices were based on subjective preferences, a population of neurons in the OFC was found to represent the economic value of the chosen option [21]. In addition to the reward magnitude, OFC neural activity also encodes the probability and the amount of delay in time of reward delivery; whereas decision costs such as effort appear to be represented elsewhere such as in the ACC [14]. Understanding how these various factors are computed by neural circuits and combined in a valuation process represents a major challenge for future research.

Finally, several studies have estimated reward value as a leaky integrator of past rewards, and concluded that the time constant of this integration is on the order of a few trials [22–24,16•], suggesting a characteristic time of tens of seconds.

Modulation by reward signal

How are reward values in brain areas such as the striatum, OFC, and ACC, computed based on feedback signals through reward? One possibility is that reward signals mediated by DA, gate synaptic plasticity thereby implementing reward-dependent learning. As we argue in the next section, such a learning mechanism can be used to compute and store reward values. Indeed, there is evidence that long-term potentiation (LTP) and long-term depression (LTD) are modulated by the activation of DA signals in many brain areas such as the hippocampus, prefrontal cortex, and striatum [25,26].

The cortico-striatal synapses are the best-studied system for the modulatory effect of DA on plasticity. In an early study on the role of DA in reward-dependent learning, it was shown that the induction of LTP in cortico-striatal synapses required the presence of DA [27]. In this *in vivo* experiment, an extracellular stimulation protocol induced LTD at cortico-striatal synapses; the stimulation of DA neurons in the substantia nigra pars compacta with optimal frequency resulted in DA release in the striatum, and led to a switch from LTD to LTP. This and other experiments gave rise to the idea that Hebbian plasticity in cortico-striatal synapses can be modulated by the presence or the absence of DA signaling [25].

A recent study challenged this view by showing that without manipulation of DA signaling, both LTP and

LTD can be induced at the cortico-striatal synapses by different temporal patterns of activity in striatal neurons [28]. In a more recent study, it was shown that DA activation was necessary for spike-timing-dependent plasticity (STDP) at cortico-striatal synapses [29•]. Specifically, blocking the D1/D5 receptors prevented the induction of both LTD and LTP in cortico-striatal synapses, whereas blocking D2 receptors had differential effects on LTP and LTD induction. Whereas Fino *et al.* [28] found that LTP (respectively LTD) was induced when the postsynaptic spikes occurred before (after) the presynaptic spikes, Pawlak and Kerr [29•] found the opposite. A possible explanation of this discrepancy is that GABAergic synaptic transmission, which could intervene in network dynamics, was blocked in [29•] but not in [28]. In addition, a recent study demonstrated that cholinergic interneurons have an important role in DA-dependent plasticity in the striatum [30].

Several studies showed that DA can also transform LTD to LTP in the rat prefrontal cortex, and this effect requires activation of both D1 and D2 receptors [31]. Moreover, using a mouse preparation, it was shown that activation of the dopamine D1 receptor facilitated the maintenance of LTP, whereas the D1 antagonist blocked the maintenance (but not the induction) of LTP in the prefrontal cortex [32]. DA application had no effect on the LTD induction in heterozygous mice that lacked D1 receptors. Similar findings were also obtained in the rat hippocampus, where the activation of D1/D5 receptors reversed LTD induced by low-frequency stimulation [33]. These results may indicate a role of D1/D5 receptors in maintaining activity-dependent LTP and reducing LTD at hippocampal synapses.

Overall, these findings provide growing evidence for important modulatory effects of DA on LTP and LTD in the striatum and prefrontal cortex. However, most of these studies were carried out *in vitro*, with bath application of DA or DA receptor blockers. It remains to be determined whether such manipulation of baseline DA is meaningful to understanding the impact of phasic activity of DA neurons *in vivo*, or whether it is more akin to tonic firing [34] or slower changes of activity in DA neurons [35•].

Learning reward value

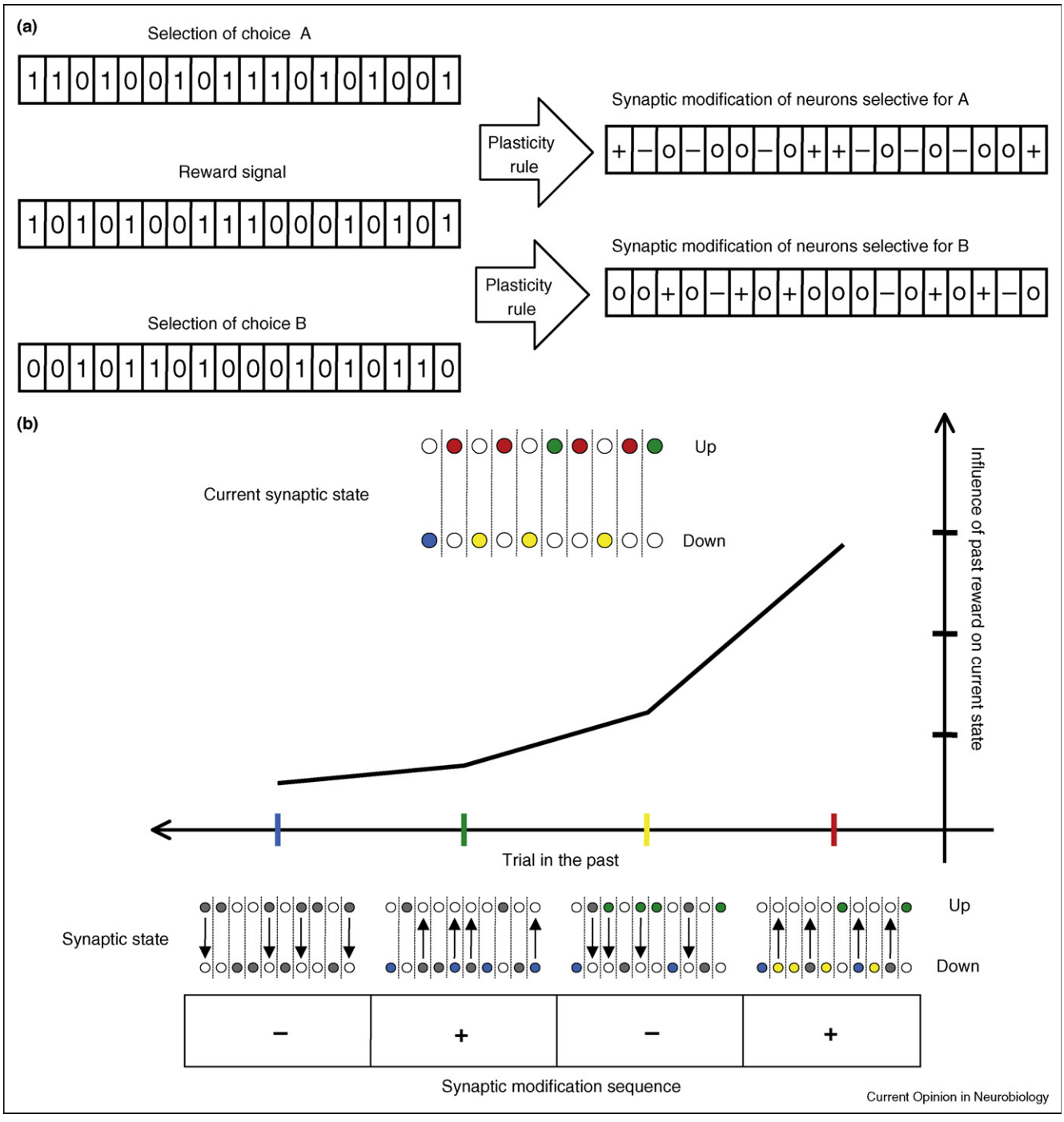
With a growing body of work on representations of reward value, the question of how reward value is learnt and updated over time at the cellular level has gained urgency. This is the topic of several recent computational studies, which implement specific forms of reward-dependent synaptic plasticity into spiking neural networks that perform various types of behavioral tasks [36,37•,38,39•].

One proposed scenario for reward-dependent learning assumes that a global reward signal modifies the probability p of release of transmitter vesicles at 'hedonistic' synapses [36]. It was shown that a network with such hedonistic synapses could perform gradient learning that results in optimal reward harvest. The stochastic nature of transmitter release is a crucial component for optimization in this model. Because usually there is a temporal gap between the reward signal and the activity that results in the reward (i.e. distal reward problem), a signal known as the eligibility trace is required to 'remember' synaptic activities before the arrival of the reward signal. The eligibility trace in this model tracks the success and failure of transmitter release over time. In this way only synapses that contribute to the rewarded outcomes are strengthened. In this model, learning occurs on a spike-by-spike basis (i.e. if a presynaptic spike leads to transmitter release, the synapse is 'rewarded' with an increase in p; otherwise it is 'punished' with a decrease in p), which seems to be at odds with the much slower time course of reward-related DA signaling [40]. Moreover, it has been shown that the behavior of human subjects in a sequential economic decision game can be explained by adding an eligibility trace, which persists over actions in time, to the temporal difference learning model [41].

Another study examined the distal reward problem using an eligibility trace for synaptic changes that follow a STDP rule and are multiplicatively modulated by a DA signal [39•]. It was shown that a spiking network model endowed with this learning rule could simulate classical (Pavlovian) and instrumental (operant) conditioning. Moreover, by assuming that synapses onto DA neurons follow the same learning rule, the model replicates the observation that during the Pavlovian conditioning, the neural response of DA neurons shifts backward in time, from the unconditioned stimulus to the conditioned stimulus.

The aforementioned studies focused on reward-dependent learning rules, rather than the neural network that produces behavior. In contrast, in our recent work [37•,38,42], a stochastic reward-dependent Hebbian plasticity rule was incorporated in a biophysically-based recurrent (attractor) network model of decision-making [43]. This same model was applied to several monkey experiments and shown to account for both behavioral and neurophysiological observations, such as stochastic foraging [37•], playing a competitive matching-pennies game [38], and arbitrary sensorimotor mapping [42]. The learning rule is a stochastic Hebbian rule [44] gated by reward signal, based on the experimental findings that the presence or absence of DA signaling can alter the direction of synaptic plasticity. For instance, after the network generates a categorical decision, LTP occurs if the choice is rewarded; otherwise LTD occurs. Interestingly, it was realized that according to this learning rule, the rewarding value of each choice is learnt and stored in a set of plastic synapses in the form of return (i.e. the amount of reward

Figure 1



A mechanism for leaky integration of rewards over trials. **(a)** In each trial, one of the choices is selected by recurrent dynamics of a decision-making network consisting of neural pools that are selective for choice alternatives and compete against each other. Each decision leads to a feedback signal, which indicates whether the choice is rewarded or not. Based on a plasticity rule which depends on the choice and reward outcome, synapses onto decision neurons undergo stochastic modification. Namely, if a neural pool fires at a high rate and wins the competition, and the choice is rewarded (respectively, not rewarded), synapses onto these neurons are potentiated (+) (depressed (−)). If firing activity is low (for those neurons that loses the competition), no synaptic modification occurs (o). **(b)** Synapses are binary, Down and Up states corresponding to depressed and potentiated states, respectively. In every trial in which the condition for modification is met for a set of synapses, synapses in this set are updated probabilistically: for potentiation a fraction of synapses switch from the Down state to the Up state (and *vice versa* for depression). Here for illustration purposes we show a set of 10 synapses onto one of the decision neural pool, and how their states change in four trials for a given synaptic modification sequence (shown at the bottom). The state of each synapse is shown by a filled circle. Different colors indicate the trial number in the past when the synapse was updated: red (last trial), yellow (second trial in the past), green (third trial in the past), blue (fourth trial in the past), gray (more than four trials in the past). The current state of the set of plastic synapses shows the sum of updates in past trials but more synapses have been updated by reward in most recent trials. Therefore, synapses integrate reward while more recent rewards have a stronger influence on the current state of synapses and the choice behavior. The integration of past rewards is leaky because there is a finite number of synapses with discrete states, so that more recent modifications override previous ones.

per choice selection), rather than in the form of income (i.e. the amount of reward per trial).

In this model valuation occurs at the synapse level, whereas selection is carried out by the decision neural circuit. In a single trial, decision-making takes place through stochastic attractor dynamics that integrates inputs and generates winner-take-all competition between neural pools selective for different alternatives (say target A versus target B in a binary choice). Across trials, the decision behavior is statistically described by the probability of choosing one of the options, as a sigmoid function of the difference in synaptic strength of inputs to the two competitive neural pools. These synapses are dynamically updated in each trial, leading to adaptive choice behavior.

Stochastic choice behavior in our model originates from irregular spiking activity of decision neurons. This variability enables the model to explore both alternatives and avoid repetitive selection of only one option even if it is more rewarding. This randomness is also crucial for performance in interactive games when it is desirable for decision makers to be unpredictable to opponents [38].

It can be shown that this model is similar to the implementation of a certain type of RL model known as the state-less Q-learning, but the underlying learning mechanism is different [38]. Ongoing LTP and LTD lead to leaky integration of past rewards, with a time constant that depends on not only the learning rate but also the reward statistics in the environment (i.e. the probability of reward delivery assigned to each choice) [37•]. As illustrated in Figure 1, plastic synapses act as leaky integrators of past rewards associated with specific actions. Thus, this model provides a possible synaptic explanation for reward integration on single [16•,22,23] or multiple timescales [42]. Moreover, this model supports a local (in time) mechanism for producing the observed global matching behavior (i.e. the proportional allocation of choices matches the relative reinforcement obtained on those choices), namely melioration through probabilistic selection of the more valuable option in individual trials. Alternatively, it has been proposed that matching behavior can result from a synaptic plasticity rule that is driven by the covariance between reward and neural activity [45].

Therefore, existing models suggest that reward value can be computed through plastic synapses onto neural populations that instantiate choice selection. Hence, these neurons are modulated by action values. The experimental implication is that neurons selective for action-specific rewards may actually underlie action selection, rather than encoding action values separately from the decision process itself, unless they are explicitly shown to be insensitive to choices [18,19,22]. Another issue highlighted by modeling work is the need for a mechanism that bridges the temporal gap between an action and its outcome in typical situations when outcomes are revealed only long after actions take place. Most proposed mechanisms for the distal reward problem require a form of eligibility traces, of which the cellular basis remains elusive [46].

Representation of prior information

Our behavior is often cued by prior information about possible outcomes. In perceptual or economic decision-making, prior probability is often instructed in terms of identification of possible choice alternatives or is learned through experience of the number of times that each alternative choice is rewarded. In general, it is difficult to design an experiment that examines the effect of prior information on behavior independent of reward information. This is because in most paradigms prior information is instructed by reward.

In a seminal work on saccadic movement toward one of many visual targets, Basso and Wurtz showed that buildup neurons in the superior colliculus (SC) decreased their activity as the probability of the saccade to their response field (RF) was diminished (either by varying the number of possible targets or through learning) [47•,48]. This modulation of SC buildup neurons may partly be caused by an enhanced inhibition from the substantia nigra pars reticulata, where neural activity was found to increase with the number of target alternatives [49]. This finding was supported by a slightly different experiment reporting that, as the probability of saccade to the RF of SC neurons was increased, the pretarget activity of these neurons increased, and this modulation was negatively correlated with the saccadic reaction time (RT) [50]. Moreover, SC neurons showed anticipatory activity when a reward was expected in their RF [51].

In contrast, modulation by the number of targets is more complex in the frontal eye field (FEF), a cortical command center for saccadic eye movements within the frontal lobe. Lee and Keller showed that, as the number of alternatives was increased, the activity of visual neurons (primarily responsive to visual stimuli) in FEF decreased, whereas the activity of visuo-motor neurons (whose activity is selective for both visual inputs and saccade responses) increased [52•]. Taken together, prior information influences preparatory neural activity of neural circuits (such as SC and FEF) involved in action selection and execution.

In value-based choice tasks, many studies have shown that prior knowledge about reward probability of response alternatives modulates neural activity in multiple brain regions, including midbrain DA neurons [53], caudate [54], posterior parietal cortex (area LIP) [55], and various

parts of the prefrontal cortex [14]. For instance, in a cued saccade task in which two alternatives were rewarded with different probabilities or magnitudes, the activity of LIP neurons was positively modulated by the probability and the magnitude of reward of the choice into the neuron's RF [55]. Similarly, the precue activity of caudate neurons increased or decreased as the monkey learned whether the preferred saccade direction of the recorded neuron was rewarding or not in a block of trials [54]. These observations, however, can be interpreted in terms of modulation by reward values of which reward probability is an integral part, rather than prior *per se*.

Overall, these findings indicate that preparatory activity in different brain areas is correlated with the number of alternative choices and the probability that an outcome is rewarded. The output of these neurons can be used in different ways to influence decision: either by providing an extra input to bias selection toward a more probable alternative or by adjusting the decision threshold for choice selection. Future experimental and modeling work is needed to differentiate these two scenarios.

Combination of information from multiple sources

As we discussed, if reward values and prior information are learnt through reward-dependent synaptic plasticity in a decision circuit, this information should be combined with sensory data to guide behavior. The combination of different sources of information has been the subject of numerous behavioral studies, which are strongly influenced by Bayesian inference theories [56,57,5]. This line of work has been mostly concerned with integrating sensory information from different modalities, or accumulation of information over time. Little is known about how sensory information is combined with reward values and probabilities in a decision process.

A recent experiment was specifically designed to explore neural mechanisms for the combination of prior probability and sensory information (ME Mazurek *et al.*, Soc for Neurosci Abstr 2005, 621.3). Monkeys chose between two color targets corresponding to two alternatives for the net direction of a random dot motion stimulus. The prior probability that the motion direction was toward one of the color targets was kept constant in a block of trials, and was changed from one block to the next (for instance, 1:1, 1:3, 3:2, etc.). It was found that the monkeys' choice behavior was biased toward the more probable alternative, and the RT for selection of this alternative was shorter. LIP neurons recorded from behaving monkeys showed ramping activity correlated with the integration of sensory data over time; the slope of ramping activity of a given neuron increased (respectively decreased) as the prior probability that the motion was toward the target in the neuron's RF had increased (decreased). Such modulation of neural activity may be useful to influence perceptual decisions at the behavioral level.

Interestingly, we found that the same recurrent decision circuit model discussed above, with two additive sets of synaptic inputs, can capture both behavioral and neural observations in this experiment (A Soltani, PhD thesis, Brandeis University, 2006). One input pathway conveys sensory (visual motion) inputs onto the two competing neural pools. Color target inputs arrive onto the decision circuit in another input pathway, passing through synapses endowed with reward-dependent plasticity as in [37•]. The outcome of each trial provides a feedback signal (reward or no reward), by which the plastic synapses for each color target dynamically learn to estimate a function of prior probability that the target corresponds to the correct choice. We found that the difference in the strengths of plastic synapses provides an extra input to decision neurons that biases the choice behavior toward the more probable alternative and shortens RT for selection of this alternative. In addition, this extra signal increases the rate of ramping activity in the decision neurons, as observed in the LIP. It is note worthy that a change in the ramping slope is in contrast to the Bayesian prescription that the optimal strategy for incorporating priors in a ramping-to-threshold decision process is to modify the starting point (or equivalently the decision bound) [2].

There is now a sizable body of work documenting that LIP neurons are modulated by the time integration of sensory stimuli in perceptual decisions [2], the magnitude and probability of reward in free-choice tasks [55,22], and the combination of evidence from different shapes in a probabilistic categorization task [58]. Therefore, LIP may play an important role in the integration of different types of information in oculomotor decision-making. The output signal of this integration can be used to guide eye movement or shift attention. However, it is currently unknown whether neural signals observed in LIP are generated locally, or if they emerge as the collective behavior of a larger circuit encompassing interconnected parietal and frontal areas. Furthermore, it remains to be determined whether LIP is a more general-purpose system of information integration, even in decisions that do not involve saccadic eye movements.

Conclusions

The articles reviewed here illustrate recent works that aim at relating RL theories and cellular mechanisms of reward-based decision-making. Presently, there still exists a wide gap between these different levels of description of adaptive choice behavior, but our knowledge has greatly benefited from exchanges and collaboration between disciplines. Future research in this direction will help us

understand the neural representations of reward value, prior information, and how they are combined in the brain.

Acknowledgements

This work was supported by NIH grants MH073246. We thank Zahra Ayubi and Daeyeol Lee for comments on the manuscript.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest

1. Sugrue LP, Corrado GC, Newsome WT: **Choosing the greater of two goods: neural currencies for valuation and decision making**. *Nat Rev Neurosci* 2005, **6**:363-375.

2. • Gold JI, Shadlen MN: **The neural basis of decision making**. *Annu Rev Neurosci* 2007, **30**:535-574.
This is an authoritative and comprehensive overview of electrophysiological findings in different types of decision-making tasks, as well as theoretical approaches based on the signal detection theory and sequential analysis.

3. Heekeren HR, Marrett S, Ungerleider LG: **The neural systems that mediate human perceptual decision making**. *Nat Rev Neurosci* 2008, **9**:467-479.

4. Loewenstein G, Rick S, Cohen JD: **Neuroeconomics**. *Annu Rev Psychol* 2008, **59**:647-672.

5. Doya K, Ishii S, Pouget A, Rao RPN (Eds): *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press; 2007.

6. Sutton RS, Barto AG: **Reinforcement Learning: An Introduction**MIT Press; 1998.

7. Schultz W, Dayan P, Montague PR: **A neural substrate of prediction and reward**. *Science* 1997, **275**:1593-1599.

8. • Roesch MR, Calu DJ, Schoenbaum G: **Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards**. *Nat Neurosci* 2007, **10**:1615-1624.
This is an important paper showing that, while rats perform a time discounting task, dopamine neurons in the ventral tegmental area encode the reward prediction error, as well as the size and delay of reward.

9. Bayer HM, Glimcher PW: **Midbrain dopamine neurons encode a quantitative reward prediction error signal**. *Neuron* 2005, **47**:129-141.

10. Matsumoto M, Hikosaka O: **Lateral habenula as a source of negative reward signals in dopamine neurons**. *Nature* 2007, **447**:1111-1115.

11. • Nakahara H, Itoh H, Kawagoe R, Takikawa Y, Hikosaka O: **Dopamine neurons can represent context-dependent prediction error**. *Neuron* 2004, **41**:269-280.
This is the first study to show that dopamine neurons can signal reward prediction error which depends on the context. The context in this task is the number of trials since the last reward, which indicates the growing probability that a reward will appear in the upcoming trial.

12. Tobler PN, Fiorillo CD, Schultz W: **Adaptive coding of reward value by dopamine neurons**. *Science* 2005, **307**:1642-1645.

13. • Matsumoto M, Matsumoto K, Abe H, Tanaka K: **Medial prefrontal cell activity signaling prediction errors of action values**. *Nat Neurosci* 2007, **10**:647-656.
This work reveals that both positive and negative reward prediction error signals are coded by an increase in firing activity, in two distinct subpopulations of neurons in the medial prefrontal cortex. The response of these neurons changes as monkeys learn the action–outcome association, in a way that validates the hypothesis that these neurons signal positive and negative prediction errors.

14. Rushworth MFS, Behrens TEJ: **Choice, uncertainty and value in prefrontal and cingulate cortex**. *Nat Neurosci* 2008, **11**:389-397.

15. Barraclough DJ, Conroy ML, Lee D: **Prefrontal cortex and decision making in a mixed-strategy game**. *Nat Neurosci* 2004, **7**:404-410.

16. • Seo H, Lee D: **Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game**. *J Neurosci* 2007, **27**:8366-8377.
The authors record from ACC neurons in behaving monkeys during a competitive game and show that spiking activity in some neurons is correlated with rewards in previous trials while activity in others encodes a signal related to reward prediction errors.

17. • Samejima K, Ueda K, Doya K, Kimura M: **Representation of action-specific reward values in the striatum**. *Science* 2005, **310**:1337-1340.
This paper provides the clearest evidence so far that in a free-choice task, a significant fraction of recorded striatal neurons encode possible action values but are not selective for chosen responses.

18. Lau B, Glimcher PW: **Action and outcome encoding in the primate caudate nucleus**. *J Neurosci* 2007, **27**:14502-14514.

19. Lau B, Glimcher PW: **Value representations in the primate striatum during matching behavior**. *Neuron* 2008, **58**:451-463.

20. Montague PR, Berns GS: **Neural economics and the biological substrates of valuation**. *Neuron* 2002, **36**:265-284.

21. Padoa-Schioppa C, Assad JA: **Neurons in the orbitofrontal cortex encode economic value**. *Nature* 2006, **441**:223-226.

22. Sugrue LP, Corrado GC, Newsome WT: **Matching behavior and representation of value in parietal cortex**. *Science* 2004, **304**:1782-1787.

23. Lau B, Glimcher PW: **Dynamic response-by-response models of matching behavior in rhesus monkeys**. *J Exp Anal Behav* 2005, **84**:555-579.

24. Kennerley SW, Walton ME, Behrens TEJ, Buckley MJ, Rushworth MFS: **Optimal decision making and the anterior cingulate cortex**. *Nat Neurosci* 2006, **9**:940-947.

25. Reynolds JN, Wickens JR: **Dopamine-dependent plasticity of corticostriatal synapses**. *Neural Netw* 2002, **15**:507-521.

26. Calabresi P, Picconi B, Tozzi A, di Filippo M: **Dopamine-mediated regulation of corticostriatal synaptic plasticity**. *Trends Neurosci* 2007, **30**:211-219.

27. Reynolds JN, Hyland BI, Wickens JR: **A cellular mechanism of reward-related learning**. *Nature* 2001, **413**:67-70.

28. Fino E, Glowinski J, Venance L: **Bidirectional activity-dependent plasticity at corticostriatal synapses**. *J Neurosci* 2005, **25**:11279-11287.

29. • Pawlak V, Kerr JND: **Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity**. *J Neurosci* 2008, **28**:2435-2446.
The authors show that cortico-striatal synapses in spiny projection neurons in the striatum follow a form of plasticity which depends on the precise timing of presynaptic and postsynaptic action potentials. Interestingly, the activation of D_1/D_5 dopamine receptors is required for this plasticity.

30. Wang Z, Kai L, Day M, Ronesi J, Yin HH, Ding J, Tkatch T, Lovinger DM, Surmeier DJ: **Dopaminergic control of corticostriatal long-term synaptic depression in medium spiny neurons is mediated by cholinergic interneurons**. *Neuron* 2006, **50**:443-452.

31. Matsuda Y, Marzo A, Otani S: **The presence of background dopamine signal converts long-term synaptic depression to potentiation in rat prefrontal cortex**. *J Neurosci* 2006, **26**:4803-4810.

32. Huang YY, Simpson E, Kellendonk C, Kandel ER: **Genetic evidence for the bidirectional modulation of synaptic plasticity in the prefrontal cortex by D1 receptors**. *Proc Natl Acad Sci U S A* 2004, **101**:3236-3241.

33. Mockett BG, Guevremont D, Williams JM, Abraham WC: **Dopamine d1/d5 receptor activation reverses nmda receptor-dependent long-term depression in rat hippocampus**. *J Neurosci* 2007, **27**:2918-2926.

34. Grace AA: **Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia**. *Neuroscience* 1991, **41**:1-24.

35. Schultz W: **Multiple dopamine functions at different time courses**. *Annu Rev Neurosci* 2007, **30**:259-288.
This paper focuses on different time courses of dopamine signaling and proposes a role for these signals in guiding the behavior at different time scales.

36. Seung HS: **Learning in spiking neural networks by reinforcement of stochastic synaptic transmission**. *Neuron* 2003, **40**:1063-1073.

37. Soltani A, Wang XJ: **A biophysically-based neural model of matching law behavior: melioration by stochastic synapses**. *J Neurosci* 2006, **26**:3731-3744.
The authors show that in a neural circuit model for a foraging task, plastic synapses with a biophysically plausible learning rule can estimate reward values in terms of return. The information stored in synapses is reflected in the spiking activity of decision neurons, and reward-dependent plasticity generates adaptive choice behavior in accordance with the matching law.

38. Soltani A, Lee D, Wang XJ: **Neural mechanism for stochastic behavior during a competitive game**. *Neural Netw* 2006, **19**:1075-1090.

39. Izhikevich EM: **Solving the distal reward problem through linkage of STDP and dopamine signaling**. *Cereb Cortex* 2007, **17**:2443-2452.
The author shows that a spiking network model endowed with a STDP learning rule, which is modulated by dopamine signals, can reproduce observations in classical and instrumental conditioning.

40. Montague PR, McClure SM, Baldwin PR, Phillips PE, Budygin EA, Stuber GD, Kilpatrick MR, Wightman RM: **Dynamic gain control of dopamine delivery in freely moving animals**. *J Neurosci* 2004, **24**:1754-1759.

41. 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**:111-121.

42. Fusi S, Asaad WF, Miller EK, Wang XJ: **A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales**. *Neuron* 2007, **54**:319-333.

43. Wang XJ: **Probabilistic decision making by slow reverberation in cortical circuits**. *Neuron* 2002, **36**:955-968.

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

45. 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 U S A* 2006, **103**:15224-15229.

46. Seo H, Barraclough DJ, Lee D: **Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex**. *Cereb Cortex* 2007, **17(Suppl 1)**:i110-i117.

47. Basso MA, Wurtz RH: **Modulation of neuronal activity by target uncertainty**. *Nature* 1997, **389**:66-69.
This is the first study to show the effect of uncertainty in the motor output (saccade location) on the preparatory activity of neurons that control this behavior. The authors find that the premovement activity of saccade-related neurons in SC decreases as the probability that a particular saccade will be selected decreases.

48. Basso MA, Wurtz RH: **Modulation of neuronal activity in superior colliculus by changes in target probability**. *J Neurosci* 1998, **18**:7519-7534.

49. Basso MA, Wurtz RH: **Neuronal activity in substantia nigra pars reticulata during target selection**. *J Neurosci* 2002, **22**:1883-1894.

50. Dorris MC, Munoz DP: **Saccadic probability influences motor preparation signals and time to saccadic initiation**. *J Neurosci* 1998, **18**:7015-7026.

51. Ikeda T, Hikosaka O: **Reward-dependent gain and bias of visual responses in primate superior colliculus**. *Neuron* 2003, **39**:693-700.

52. Lee KM, Keller EL: **Neural activity in the frontal eye fields modulated by the number of alternatives in target choice**. *J Neurosci* 2008, **28**:2242-2251.
The authors show that at different times during a color-to-location saccade task, FEF neurons show different forms of modulations with respect to the number of alternatives (NAs); visual neurons decrease their activity during stimulus presentation as NA is increased, whereas the visuo-motor neurons increase their activity during presaccadic time.

53. Fiorillo CD, Tobler PN, Schultz W: **Discrete coding of reward probability and uncertainty by dopamine neurons**. *Science* 2003, **299**:1898-1902.

54. Takikawa Y, Kawagoe R, Hikosaka O: **Reward dependent spatial selectivity of anticipatory activity in monkey caudate neurons**. *J Neurophys* 2002, **87**:508-515.

55. Platt ML, Glimcher PW: **Neural correlates of decision variables in parietal cortex**. *Nature* 1999, **400**:233-238.

56. Ernst MO, Bulthoff HH: **Merging the senses into a robust percept**. *Trends Cogn Sci* 2004, **8**:162-169.

57. Ma WJ, Beck JM, Latham PE, Pouget A: **Bayesian inference with probabilistic population codes**. *Nat Neurosci* 2006, **9**:1432-1438.

58. Yang T, Shadlen MN: **Probabilistic reasoning by neurons**. *Nature* 2007, **447**:1075-1080.