nature neuroscience

A reservoir of time constants for memory traces in cortical neurons

Alberto Bernacchia, Hyojung Seo, Daeyeol Lee & Xiao-Jing Wang

According to reinforcement learning theory of decision making, reward expectation is computed by integrating past rewards with a fixed timescale. In contrast, we found that a wide range of time constants is available across cortical neurons recorded from monkeys performing a competitive game task. By recognizing that reward modulates neural activity multiplicatively, we found that one or two time constants of reward memory can be extracted for each neuron in prefrontal, cingulate and parietal cortex. These timescales ranged from hundreds of milliseconds to tens of seconds, according to a power law distribution, which is consistent across areas and reproduced by a 'reservoir' neural network model. These neuronal memory timescales were weakly, but significantly, correlated with those of monkey's decisions. Our findings suggest a flexible memory system in which neural subpopulations with distinct sets of long or short memory timescales may be selectively deployed according to the task demands.

© 2011 Nature America, Inc. All rights reserved.

In economic behavior, choices that have a higher reward expectation are favored and adaptive decision making depends on our ability to learn reward expectation through past rewards associated with our actions. The neural mechanisms underlying this process have been the subject of growing interest, as they could provide important insights on how learning occurs in the brain and how humans and other animals make economic decisions. Neural correlates of reward valuation have been observed in different studies[1–3] and interpreted in the framework of reinforcement learning theory[4,5]. In the reinforcement learning model, reward expectation is computed by weighting the previous rewards through a temporal filter, which quantifies the memory trace of rewards. The optimal duration of the filter (memory) depends on the predictability of the environment. If the payoffs for the same option change often and unpredictably, then rewards should be filtered on short timescales to track the fast changes in a volatile environment; in contrast, if past rewards reliably predict future ones, then they should be filtered on long timescales to exploit a stable environment[6,7]. The neural mechanism underlying switching between long and short time constants for computing reward expectation remains poorly understood.

On which timescale does the brain filter rewards? To date, a few studies have estimated the time constant of this filter from behavior and assessed how past rewards affect choice selection[8–12], but the neural mechanisms responsible for such timescales are still unknown. To address this issue, we analyzed the activity of cortical neurons in monkeys performing a competitive game task. Using a method based on the idea that reward memory modulates neural activity multiplicatively, we found that memory time constants can be extracted from the activity of single neurons. We found that a different timescale for reward memory can be associated with each recorded neuron and that there is a wide range of timescales across neurons, obeying a power law distribution. The same distribution is found across three different cortical areas: anterior cingulate cortex (ACCd), dorsolateral prefrontal cortex (DLPFC) and lateral intraparietal cortex (LIP). Hence, each area is endowed with a reservoir of time constants for reward memory, which are distributed heterogeneously across neurons.

We found that the time constants estimated from pairs of simultaneously recorded neurons are uncorrelated, implying that our results cannot be explained by a single time constant for all neurons that changes slowly over time. On the other hand, our analysis of an animal's behavior suggests that the timescale over which reward events affect decisions changes across experimental sessions, possibly reflecting the animal's attempt to increase its payoff by exploring different strategies. The time constants for reward memory at the behavioral and neuronal levels were weakly correlated across experimental sessions. Finally, we found that a randomly connected circuit model, akin to a reservoir network[13–15], can reproduce the observed distribution of timescales, provided that the network operates at the critical point (or edge of chaos)[16–18]. Taken together, these findings suggest a distributed, flexible neural system for reward valuation and memory.

RESULTS

Multiplicative memory traces in cortical neurons

We analyzed single-neuron activity recorded from three cortical areas, ACCd[19] (154 neurons), DLPFC[20] (322 neurons) and LIP[21] (205 neurons) of six monkeys performing a matching pennies task[11,22] (**Fig. 1a**). In each trial, the monkey chose one of two targets by shifting its gaze and the computer made its choice by simulating a rational opponent; the animal received reward if its choice matched that of the computer. We computed firing rates of each neuron by counting the spikes in 12 time intervals of 250 ms (**Fig. 1a**), which are referred to as epochs. This includes six epochs (1.5 s) before saccade initiation (pre-fixation, fore-period and delay) and six epochs (1.5 s) after saccade completion (choice fixation, feedback and post-feedback). Consistent with previous studies[23–25], we found that the activity of neurons varied substantially in different trial epochs (99% of neurons,

Department of Neurobiology and Kavli Institute of Neuroscience, Yale University School of Medicine, New Haven, Connecticut, USA. Correspondence should be addressed to X.-J.W. (xjwang@yale.edu).

Received 18 October 2010; accepted 11 January 2011; published online 13 February 2011; doi:10.1038/nn.2752

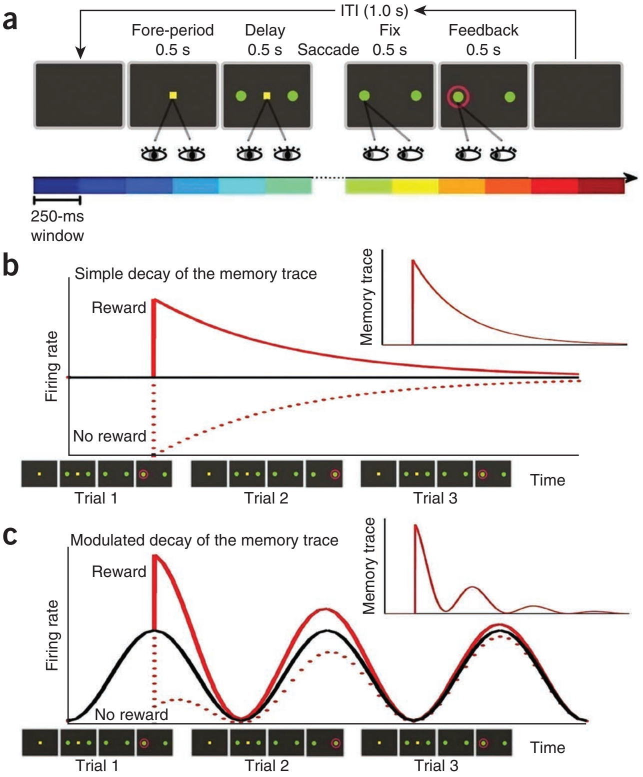

Figure 1 Behavioral task and schematic illustration of memory traces. (**a**) In the matching pennies task, the monkey was required to fixate a central spot during the fore-period (500 ms) and delay period (500 ms) while the two choice targets (green disks) were displayed. The central spot then disappeared and the monkey made a saccadic eye movement to one of the two choice targets and maintained its gaze on the chosen target for 500 ms (choice fixation). A red ring appearing around the correct target revealed the computer's choice, and if it matched the animal's choice (as illustrated), reward was delivered 500 ms later. ITI, inter-trial interval. Colored bars at the bottom show the 12 250-ms intervals (epochs) used to compute the firing rates in the analysis. (**b**,**c**) Two hypothetical neurons. The neuron in **b** has a constant average firing rate (black line), whereas the firing rate of neuron in **c** depends on the trial epoch, repeating in each of the three consecutive trials. Red lines show the change in activity as a result of the outcome in the first trial (continuous line indicates reward, dashed line indicates no reward). The inset shows the memory trace of the reward, given by the difference between the red and black lines. The memory trace of the neuron in **b** shows a simple decay, whereas that of the neuron in **c** is multiplicatively modulated by the epoch-dependent activity.

© 2011 Nature America, Inc. All rights reserved.

675 of 681, ANOVA, $P < 0.05$). The time course of the activity in successive epochs differs substantially in different neurons.

We then examined the effect of reward on the activity of neurons. Neural activity in all three cortical areas carries the information of past reward events[19–22]. We characterized the memory trace of each neuron using a new approach (**Fig. 1b**,**c**). Consider the time course of neural activity in different epochs, averaged across trials and hence across reward/no reward conditions. The difference in activity from the average time course, triggered by a reward/no reward event, was defined as the memory trace of the reward, which is positive for one of the outcomes and negative for the other. Thus, if the average activity of a neuron is zero in a given epoch, then the change by either outcome must be zero, and the memory trace in that epoch is therefore also zero. Starting with this intuition, we hypothesized that the memory trace in a given epoch is proportional to the average firing rate in that epoch. In that case the memory trace is modulated (multiplied) by the average firing rate.

We define the epoch code as the firing rate averaged across all trials, as a function of the different epochs, denoted by $g(k)$ ($k = 1,\ldots,12$ epochs, in temporal order). In one neuron recorded in ACCd, firing rate decreased after the saccade to a chosen target and increased after the feedback period (**Fig. 2a**). To separate the contributions of epoch and reward memory to neural activity, we modeled the firing rate measured in trial n and epoch k, denoted by $\mathrm{FR}(n,k)$, as the sum of the epoch code $g(k)$ and a filter $f(n',k)$ convolved with the animal's reward history in previous trials (last five trials; in each trial, Rew = +1 indicates reward; Rew = −1 indicates no reward).

$$\mathrm{FR}(n,k) = g(k) + \sum\nolimits_{n'=0:5} f(n',k)\cdot \mathrm{Rew}(n-n') \qquad (1)$$

The filter f describes how the reward in a given trial affects neural activity in the subsequent trials, assuming that the effects of rewards in successive trials are additive. For example, $f(3,4)$ describes the effect of a reward after 3 trials during epoch 4. The filter f corresponds to our definition of memory trace (**Fig. 1b**,**c**); it reflects the deviation from the epoch-dependent time course $g(k)$ resulting from a reward event. Because Rew(n) is nearly a random sequence[11] of +1 and −1, averaging the firing rates over all trials recovers the epoch code $g(k)$ (sum over n of $\mathrm{FR}(n,k)$). We estimated the memory trace $f(n',k)$ by applying multiple linear regression to the data according to equation (1). In an example negative neuron (that is, reward decreases the activity of this neuron in subsequent trials), the memory trace does not decay monotonically, but its strength is modulated throughout the trial consistent with the epoch code (**Fig. 2b**). According to the multiplicative

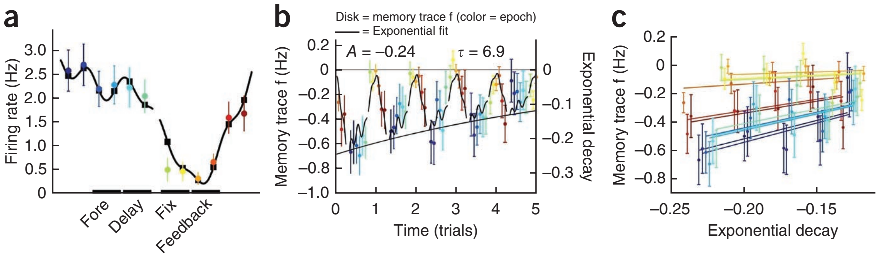

Figure 2 An example neuron in ACCd showing multiplicative modulation of memory traces by the epoch code. The colors in all panels denote trial epochs, following the format of **Figure 1a**. (**a**) The epoch code for an example neuron; that is, the firing rate computed in 12 250-ms epochs in a trial and averaged over all trials (black squares, interpolated by the black line, broken during the saccade). Colored disks correspond to the slopes fitted in **c** (error bars represent ±s.e.); their correlation with the epoch codes quantifies the multiplicative modulation and is referred to as the factorization index (0.97 in this example). (**b**) The memory trace f of past rewards in the same neuron, up to five trials in the past. Colored dots and error bars (±s.e.) show the results of the multiple linear regression model (equation (1)) and the black line is the exponential fit (equation (2), continuous line, exponential ex(t); broken line, modulated envelope g·ex(t)). The parameters for the fit are shown (A, amplitude; τ, timescale). (**c**) The memory trace f (from **b**), plotted as a function of the exponential function ex. The lines are least-squares fit, each line encompassing a particular epoch and all five trial lags. According to the factorization, the slopes should correspond to the epoch code, $f = g$·ex. The values of the slopes are plotted in **a** (colored squares) and compared with the epoch code $g(k)$.

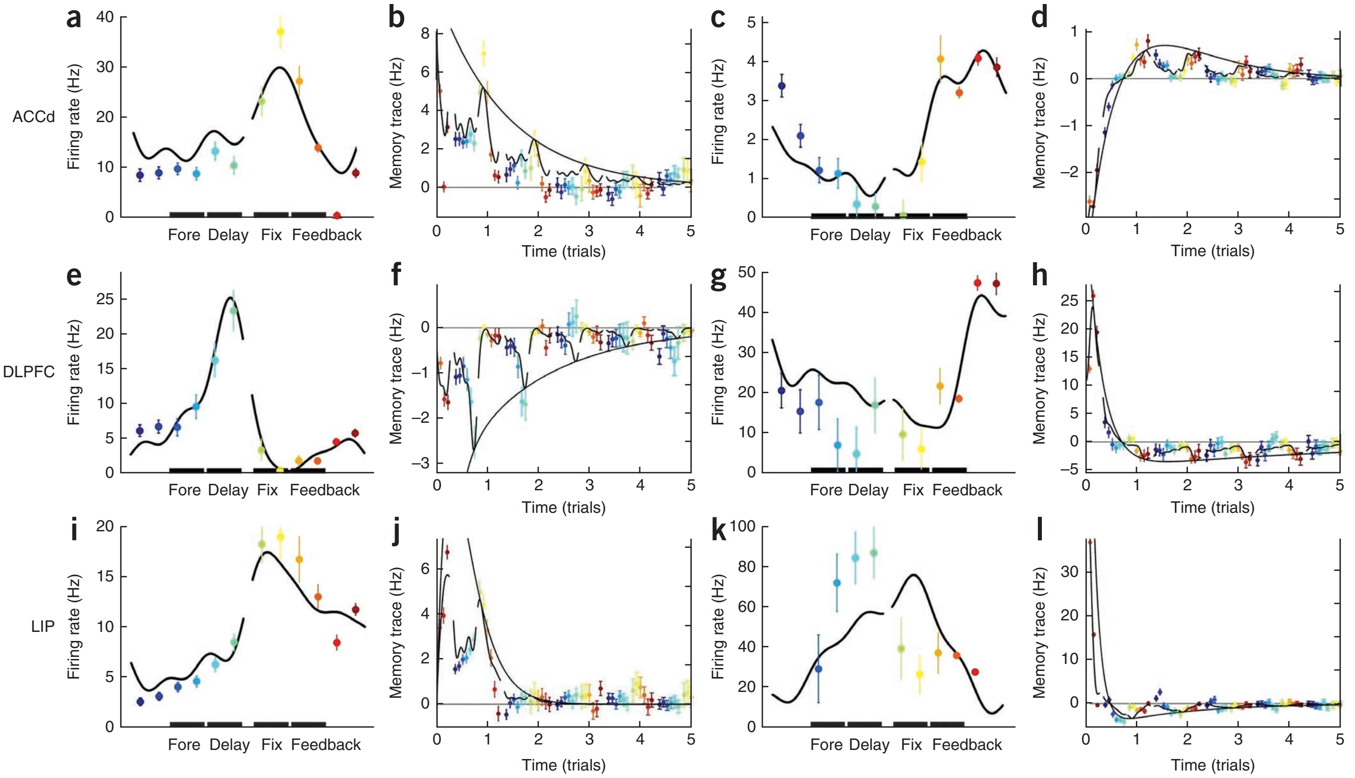

Figure 3 Firing rates and memory traces for six neurons, two for each of the three recorded areas. For each of the six neurons, epoch codes (first and third column) and memory traces (second and fourth column) are shown, presented as in **Figure 2a**,**b**. The second column shows monotonic decay of the memory trace and the fourth column shows biphasic memory traces (double exponential). Different neurons had different firing rates, both in magnitude and time course, and different types of memory decay, but they were all consistent with an exponential (single or double) decay of the memory modulated by the epoch code. The factorization indexes for those neurons are 0.98 (**a**,**b**), 0.91 (**c**,**d**), 0.98 (**e**,**f**), 0.84 (**g**,**h**), 0.97 (**i**,**j**) and 0.61 (**k**,**l**).

© 2011 Nature America, Inc. All rights reserved.

model (**Fig. 1b**,**c**), we assumed that the memory trace f is factorized into the epoch code $g(k)$ and an exponential function $\mathrm{ex}(t)$.

$$\mathrm{FR(n,k)=g}(k)+\mathrm{g(k)}\cdot\textstyle\sum_{n'=0:5}\mathrm{ex}(t)\times\mathrm{Rew}(n-n') \qquad (2)$$

The filter f considered in equation (1) is now replaced by the product of two factors $g(k)\cdot\mathrm{ex}(t)$, where $\mathrm{ex}(t)=Ae^{\frac{-t}{\tau}}$ is an exponential decay function and t is the time elapsed since the outcome (Online Methods). By applying this model to the example neuron (**Fig. 2a**,**b**), we obtained a timescale of memory decay τ = 6.9 trials and an amplitude A = −0.24. According to the factorization ($f = g\cdot\mathrm{ex}$), the constant of proportionality between the memory trace f and the exponential function ex, estimated in different epochs (**Fig. 2c**), should reproduce the epoch code $g(k)$. The epoch codes for the neuron closely followed these predictions, indicating that the factorization is nearly exact (**Fig. 2a**). The factorization index of a neuron, defined as the correlation coefficient between the epoch code and the proportionality constants (slopes), was 0.97 for this neuron.

The modulated decay of the memory trace was observed in the majority of the recorded neurons in all three cortical areas. In some cases, the sum of two exponential functions, $\mathrm{ex}(t)=A_1e^{\frac{-t}{\tau_1}}+A_2e^{\frac{-t}{\tau_2}}$, fitted the data better than a single exponential, in which case the memory trace often exhibited a biphasic characteristic (with A_1 and A_2 of the opposite sign; **Fig. 3**). Using the Bayesian Information Criterion, we found that the best fit was a single exponential for 269 neurons and double exponentials for 268 neurons, whereas the remaining 144 neurons were fitted best by a model with ex(t) = 0. The latter is interpreted as no memory and the corresponding neurons were excluded from further analysis. We tested the validity of the fitting procedure by randomly reshuffling the order of trials in each session and we consistently found that 96% of neurons (656 of 681) showed no memory after reshuffling.

We examined the average firing rates and memory traces of all recorded neurons (examples shown in **Fig. 3**). Although the activity of most neurons is consistent with an exponential decay of the memory trace (79%, 537 of 681, single and double exponentials), a fraction of them did not show a modulation of the memory by the epoch code. This is quantified by the factorization index, which is significantly positive for approximately half of the neurons showing a memory effect (46%, 249 of 537, P < 0.05, t test). We found a small, but significant, difference in the fraction of neurons with memory across different areas (87% in ACCd, 75% in DLPFC and 78% in LIP, χ^2 test, P = 0.01).

We next investigated how the timescales of memory traces were distributed across neurons in different cortical areas. We determined that the distribution of timescales in all cortical areas could be fit with a power law with an exponent of −2 (**Fig. 4**). The power law implies that timescales are distributed in a wide range of values. In fact, for a power law distribution ~τ^{-2}, the variance increases with the sample size and, in principle, arbitrarily large timescales would be observed with a proportionally large increment in the number of recorded neurons. Note that the power law tail applies for timescales equal to or larger than one trial, which are those timescales that might be involved in memory (see below). About 20% of all recorded neurons (133 of 681)

© 2011 Nature America, Inc. All rights reserved.

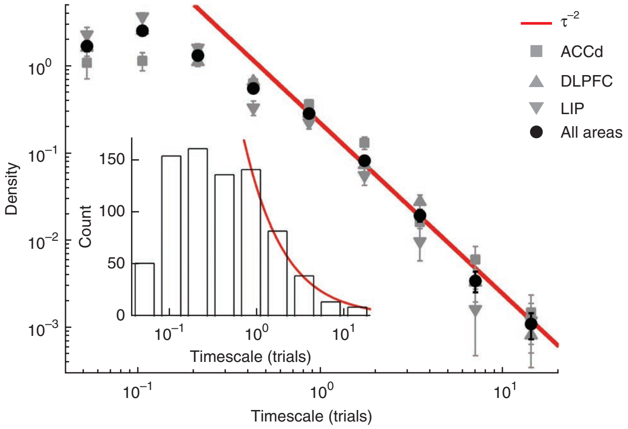

Figure 4 Distribution of the timescales characterizing the reward memory traces across neurons. Black disks show the density for the neurons in all three cortical areas in the corresponding bin; that is, the count of timescales divided by the bin length (error bars represent ±s.e.). The inset shows the count of the timescales in the same bins, in a linear scale (a total of 805 timescales). Grey markers show the density separately for each of the three different cortical areas (square, ACCd, 197 timescales; upward triangle, DLPFC, 362; downward triangle, LIP, 246). The red line (red curve in the inset) shows a power law fit (exponent = −2).

had timescale larger than one trial (29% in ACCd, 19% in DLPFC and 13% in LIP; χ^2 test, $P = 0.0005$; see **Supplementary Fig. 1c**–**e**). Because the timescales from one- (τ) and two-exponential functions (τ_1, τ_2) were distributed similarly (**Supplementary Fig. 1a**,**b**), we pooled all timescales (a total of 805 timescales from 269 single exponential and 268 double exponential; that is, 269 τ, 268 τ_1 and 268 τ_2). ACCd contributed 197 timescales from 71 single exponential and 63 double exponential functions (71 τ, 63 τ_1 and 63 τ_2), whereas 20 neurons had no memory. A total of 362 timescales were obtained from DLPFC with 124 single and 119 double exponential functions (124 τ, 119 τ_1 and 119 τ_2) and 79 DLPFC neurons had no memory. LIP neurons contributed 246 timescales from 74 single and 86 double exponential functions (74 τ, 86 τ_1 and 86 τ_2) and 45 LIP neurons showed no memory.

Comparison with behavior

Are the neural memory timescales relevant for learning and decision making? The matching pennies task that we used does not necessarily require the memory of past rewards and the optimal strategy for the monkey is to choose randomly and unpredictably. Although the overall performance of monkeys was nearly optimal, their trial-by-trial decisions, locally in time, were influenced by previous rewards and actions[11,19–22]. We analyzed the behavior of monkeys in different experimental sessions by fitting their decisions with a standard reinforcement learning model[5] (reinforcement learning, Online Methods). The learning rate parameter (α) of the reinforcement learning model quantifies the behavioral timescale of the memory trace ($\alpha \sim 1/\tau$). The resulting likelihood was significantly larger than the likelihood for reshuffled trials and the model fit with behavioral data was significant in 78% of the sessions (196 of 250, $P < 0.05$). We found that the timescales of behavioral memory varied across sessions, possibly suggesting that monkeys adopted different strategies in successive sessions. For the 196 sessions fitted by the reinforcement learning model, the distribution of behavioral timescales followed a power law distribution (**Fig. 5a**) and the exponent was consistent with that measured in the neural distribution. Hence, the distributions of behavioral and neuronal timescales qualitatively matched with each other. This result suggests that there might be a relationship between the memory trace observed at the neural level and that observed at the behavioral level. We tested this hypothesis by comparing the neural timescale for reward memory observed during a given recording session with the behavioral timescale fit in that session (when both are available) and we found a small, but significant, correlation across sessions ($R = 0.12$, $P = 0.003$; **Fig. 5b**), suggesting that the activity of single neurons is related, albeit weakly, to the behavioral strategy of the animals.

Do the reward memory timescales also change in a single session? We determined whether the timescales are stable in a single recording session by dividing each session into two separate blocks (halves) of trials and we re-estimated both the neural and behavioral timescales separately in the two blocks. Both the behavioral and neural memory timescales were fairly stable in a single session (**Fig. 6**).

The neural and behavioral timescales might fluctuate together across sessions, but their small correlation indicates that there is only a weak coupling. Indeed, we found that at any moment, the timescales of reward memory varied across cortical neurons. In each recording session, only few neurons were simultaneously recorded (about two on average). When we estimated memory timescales for pairs of simultaneously recorded neurons, the correlation between their time constants was not significantly different from zero (312 pairs of timescales, $R = 0.07$, $P = 0.2$). This result suggests that the broad distribution of memory time constants observed in the data reflects a variability of timescales across different neurons, rather than resulting from a memory timescale fixed for all neurons that collectively changes across sessions.

Taken together, our results support the conclusion that a diverse collection of neural memory timescales, a reservoir, is available across cortical neurons at any given time. The animal's behavior may be determined by a readout system that is able to sample, at different times,

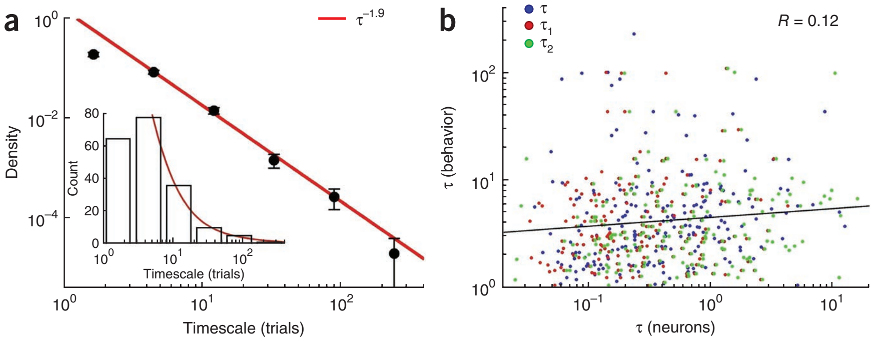

Figure 5 Distribution of behavioral timescales and their relationship with the neural memory timescales. (**a**) Time constant τ estimated from the learning rate α ($\tau \sim 1/\alpha$) of a reinforcement learning model fit to the monkey's behavioral data. Black disks show the density in the corresponding bin; that is, the count of timescales divided by the bin length (error bars represent ±s.e.). The inset shows the count of the timescales in the same bins, in linear scale (a total of 196 timescales). The red line (red curve in the inset) shows a power law fit (exponent = −1.9). (**b**) The scatterplot of behavioral versus neural memory timescales obtained from all sessions where both were available. Neural timescales from different types of fit (τ from single exponential and τ_1, τ_2 from double exponential) are shown in different colors. Behavioral and neural timescales show a small, but significant, correlation ($R = 0.12$, $P = 0.003$).

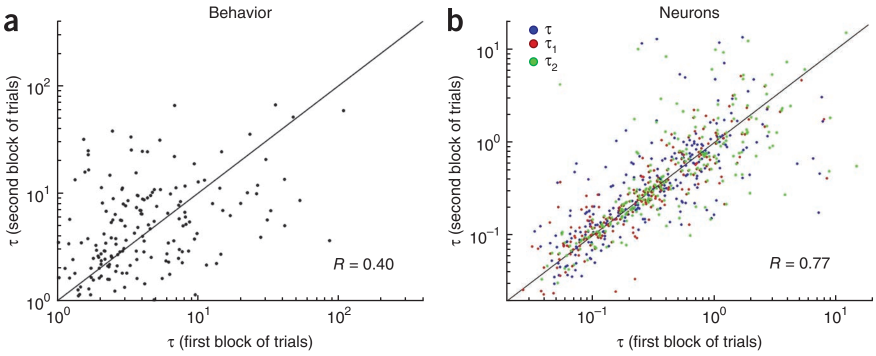

Figure 6 Stability of behavioral and neural memory timescales in an experimental session. (**a**,**b**) In both panels, the scatterplot of the timescales fitted in the second half of the trials is plotted against the timescales fitted in the first half of the trials in the same session. The correlation was significantly different from zero in both cases ($R = 0.4$ for behavioral timescales, $R = 0.77$ for neural timescales), suggesting that both types of timescales are fairly stable in a single session. Neural memory timescales from different types of fit (τ from single exponential and τ_1, τ_2 from double exponential) are shown in different colors.

from a variety of timescales present in the reservoir. The reservoir might not be static and it may change its distribution of timescales from day to day. During competitive games, the subjects might also take into account their recent choices to determine their future behavior. We therefore tested whether any memory trace of choice exists in the recorded neurons by applying the same analysis of equations (1) and (2) and substituting reward with choice. We found that multiplicative modulation and a power law distribution of memory timescales also hold for memory trace of past choices (**Supplementary Fig. 2**).

Neural network model for memory traces

What neural mechanism(s) accounts for the statistical properties of reward memory described above? To address this question, we constructed a simple neural network model that reproduces the observed neural memory traces (**Fig. 7** and **Supplementary Fig. 3**). Model neurons integrate the reward signals by receiving a current impulse whenever a reward is obtained. Because neurons are recurrently connected and form loops, their activities reverberate and are able to maintain the memory of reward events. However, those memories decay and are slowly forgotten according to a time course that depends on the pattern of synaptic connections among neuron pairs. Specifically, the activity of neurons evolve according to $\frac{\mathrm{d}v}{\mathrm{d}t} = J\cdot v(t) + h\cdot \mathrm{Rew}(t)$, where **v** is a vector of M components, each component is the activity of a different neuron in the reservoir (M = 1,000 neurons in simulations), J is the synaptic connectivity matrix of their interactions and **h** is a vector representing the relative strength of the reward input Rew(t) to each neuron. For our purposes, the specific form of the input signals is not important; the results depend only on the synaptic matrix J. We assumed that the connection weights (the entries of the matrix J) were randomly distributed and we looked for candidate probability distributions such that the network model reproduces the distributions of timescales and amplitudes observed in the neural data from behaving monkeys (see **Supplementary Text**). Amplitudes determine the extent of the immediate response of neurons to reward, with respect to the average activity. Time constants had a power-law distribution (**Fig. 4**) and the distribution of amplitudes was exponential (**Fig. 8a**, where we used A for one exponential and $A_1 + A_2$ for two exponentials).

First, we found that the connection weights must be broadly distributed among neuron pairs and that this endows the network with a wide variety of timescales. Intuitively, the stronger the connection, the longer the reverberation of the input and hence the timescale of the memory trace. However, if connections are also heterogeneous, then weaker connections and smaller timescales will also contribute to the memory traces. If the width of the distribution of connection weights reaches a certain threshold, a power-law distribution of timescales is observed (**Fig. 7b**), which is characterized by a high probability for both small and large timescales. This is a distinct type of network state 'at a critical point' (or edge of chaos in nonlinear systems), which have been proposed to be desirable for many kinds of computations[16–18]. In our model, the criticality corresponds to the situation where the system is on the verge of losing stability. When the width of the connection distribution exceeds the critical level, the linear system is unstable and the model would need to be extended to include nonlinearities such as saturation of neural activity. For the sake of simplicity, we limited ourselves to the linear model, which is sufficient for the purpose of reproducing the observed power-law distribution of timescales under specific conditions.

A second desirable property of the network is that its dynamics are robust with respect to small changes of the connection strengths. If the coding of the memory changes markedly as a result of small changes in the connection strengths (for example, synaptic noise), it would be difficult for a downstream system to interpret that code. A known property of the connection matrix J that ensures that kind of robustness is normality, which guarantees that there is an orthogonal set of eigenvectors[26] (but see refs. 27–29 for non-normal neural network models). If J is normal, we found that the amplitudes of the memory traces followed an exponential distribution (**Fig. 8b**), consistent with the experimental observations (**Fig. 8a**). To the best of our knowledge, our results provide the first complete statistical description of

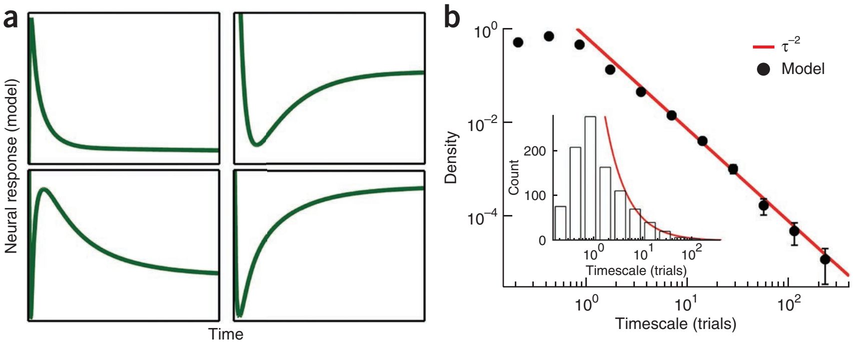

Figure 7 Neural responses (memory traces) in the model and distribution of timescales of the memory traces in model neurons. (**a**) The memory traces of four model neurons. (**b**) The black disks show the density of timescales in the corresponding bin; that is, the count of timescales divided by the bin length (error bars represent ±s.e.). The inset shows the count of the timescales in the same bins, in linear scales (a total of 1,000 timescales). The red line (red curve in the inset) shows a power law fit (exponent = –2).

© 2011 Nature America, Inc. All rights reserved.

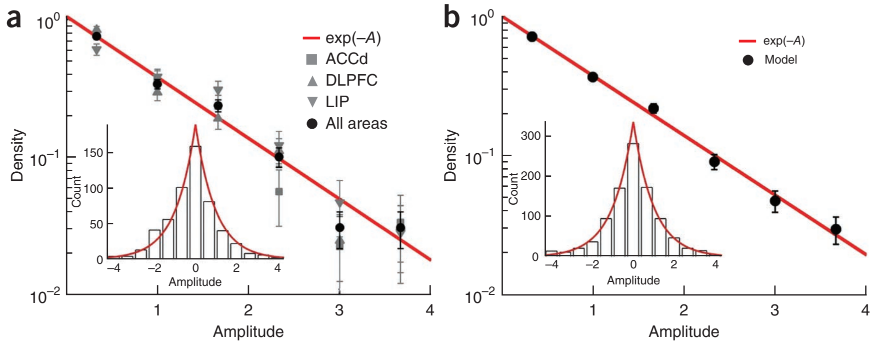

Figure 8 Distribution of amplitudes of the memory traces in the neural data and model. (**a**,**b**) In both panels, black disks show the density in the corresponding bin; that is, the count of timescales divided by the bin length (error bars represent ±s.e.). The inset shows the count of the amplitudes in the same bins, in a linear scale (537 amplitudes in the data, 1,000 in the model). Amplitudes are plotted as absolute values, as the distribution was approximately symmetric (symmetry is shown in the inset). Grey markers show the density separately for the three different recorded areas (squares, ACCd, 134 amplitudes; upward triangles, DLPFC, 243; downward triangles, LIP, 160). The red line (red curve in the inset) shows an exponential fit ($e^{-|A|}$).

© 2011 Nature America, Inc. All rights reserved.

a network connection matrix based on *in vivo* neuronal recordings of behaving animals (see also refs. 30–32).

DISCUSSION

The power law of timescales suggests that the duration of reward memory trace is highly diverse across cortical neurons. The same diversity is observed across three cortical areas, suggesting that the computation of reward memory is a distributed process. This finding is consistent with an increasing appreciation that neural encoding of cognitive variables is highly heterogeneous and distributed[33,34]. Prefrontal cortex is important for dynamic decision processes encoding and updating values[1–4]. Although anterior cingulate cortex has been implicated in monitoring conflict between incompatible response processes[35] or detecting performance errors[36], recent studies have placed more emphasis on its role in representing both positive and negative values[19,37]. Parietal cortex has also been implicated in decision making on the basis of the value representation and the accumulation of sensory evidence[38,39].

Our work provides a comprehensive description of memory traces in terms of a specific distribution of timescales across a population of neurons and introduces a framework that could potentially be applicable to different brain areas and different types of memory. The concept of multiplicative modulation of memory traces can be used to deduce the neural memory timescales in various tasks and to test the idea that a different set of time constants is selected to adapt to a specific environment[6,7]. Although the global optimal strategy for the matching pennies task is to choose randomly and therefore does not require memory, the animals made their decisions largely on the basis of their reward history[11,19–22]. Perhaps in the persistent search for an appropriate strategy, they sampled different timescales across experimental sessions. We found that those behavioral timescales followed a similar distribution and were weakly, but significantly, correlated with the timescales observed at the neural level. This suggests the possibility that the behavior might be driven by a mechanism that appropriately samples from a range of timescales in a neural network, which has yet to be elucidated. Alternatively, this weak correlation might be caused by factors that are currently not understood. Note that the observed range is different for the neuronal versus behavioral time constants. Also, we have not attempted to fit the behavioral data by a reinforcement learning model endowed with multiple time constants. Future work is needed to further assess the correlation between neural memory traces and behavior. Regardless, our results suggest that reward memory with multiple time constants might be used to compute the value functions in reinforcement learning theory in more than one timescale. Similarly, the double exponential decay of memory may correspond to a reward prediction error signal; if the short timescale (τ_1) is small enough (about one trial or smaller), then the corresponding exponential filter will respond primarily to the reward in the present trial, whereas the long timescale (τ_2) may provide a value signal by weighting the rewards in the past few trials. When the two exponentials have opposite signs, they roughly subtract the value from the actual reward signal, therefore providing a reward prediction error. It has been noted that a biphasic filtering in dopamine neurons might provide a reward prediction error[40].

Besides the memory for reward, the activity of primate cortical neurons reflects other types of short-term memory. The time course of memory-related activity varies across different neurons and different task protocols, including persistent, ramping and multi-phasic activity[41–43]. Memory traces in the neural signals are mixed with other task-dependent factors[44,45] and it has been debated as to whether other processes involved in goal-directed behavior could be inter-mixed with a memory trace, such as spatial attention[46], motor planning[47], anticipation of future events[48] or timing[49]. The epoch code in the present task might include many of those processes and we found that memory signals could be dissociated from those factors by assuming a multiplicative computation. The hypothesis of a multiplicative effect of memory on neural activity could be tested by looking more closely at the multi-phasic time course of memory-related activity observed in other experiments. The computational advantage of the multiplicative effect of memory needs to be further investigated. For example, it may serve the appropriate recall of memories at different epochs (see **Supplementary Text**), as observed in a recent study[50].

Reservoir-type networks have been the subject of active research in computational neuroscience and machine learning[13–15], but experimental support that such networks are adopted by the brain has been lacking. Those models predict that the memory of input signals is stored in a large, recurrent and heterogeneous network (reservoir) in a distributed manner and that a desired output is obtained by a trainable combination of the response signals in the reservoir. The heterogeneous encoding of the input allows the flexible learning of different output functions. In our context, that may correspond to a flexible change in strategy resulting from the variety of timescales for reward memory present in the reservoir. We present direct experimental evidence, at the level of single neurons, for a high-dimensional reservoir network of reward memory traces in prefrontal, cingulate and parietal areas of the primate cortex. This empirical finding is reproduced by a simple computational model, which suggests that reward filtering in the cortex involves a dynamic reservoir network operating at the critical point, leading to a power-law distribution of time constants. The output of the network, supposedly driving the animal's behavior, is not explicitly modeled in our equations. Further studies are necessary to elucidate how the motor areas read out the memory of reward and choices and how the two are combined to subserve adaptive choice behavior.

Power-law distributions are unusual, as they imply a high probability for both large and small time constants. A diversity of time constants also means a broad range of learning rates, as the two are inversely

© 2011 Nature America, Inc. All rights reserved.

related to each other. This is noteworthy, as a shift from an exploitive to an exploratory strategy as the environment becomes uncertain is often assessed by an increase in the learning rate[10]. Our work suggests that a broad range of learning rates are available in the system, a subset of which (fast or slow) might be selectively utilized according to which strategy is behaviorally desirable. Ultimately, this framework could lead to a new model for predicting how reward expectation is computed and how reward memory affects decision making.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank J. Mazer and M.W. Jung for comments on an earlier version of the manuscript, and R. Chaudhuri, M. Harre and J. Murray for discussions. This work was supported by the US National Institutes of Health grant R01 MH062349 and the Swartz Foundation (A.B. and X.-J.W.), and by US National Institutes of Health grants R01 MH073246 (X.-J.W. and D.L.) and DA029330 (D.L.).

AUTHOR CONTRIBUTIONS

All of the authors participated in the research design and the preparation of the manuscript. H.S. collected the data, A.B. and H.S. analyzed data, and A.B. and X.-J.W. performed modeling.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Kable, J.W. & Glimcher, P.W. The neurobiology of decision: consensus and controversy. *Neuron* **63**, 733–745 (2009).
2. Rushworth, M.F. & Behrens, T.E. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* **11**, 389–397 (2008).
3. Wang, X.-J. Decision making in recurrent neural circuits. *Neuron* **60**, 215–234 (2008).
4. Soltani, A., Lee, D. & Wang, X.-J. Neural mechanism for stochastic behavior during a competitive game. *Neural Netw.* **19**, 1075–1090 (2006).
5. Sutton, R.S. & Barto,, A.G. *Reinforcement Learning, an Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
6. Behrens, T.E., Woolrich, M.W., Walton, M.E. & Rushworth, M.F. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
7. Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B. & Dolan, R.J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
8. Lau, B. & Glimcher, P.W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* **84**, 555–579 (2005).
9. Corrado, G.S., Sugrue, L.P., Seung, H.S. & Newsome, W.T. Linear-nonlinear-Poisson models of primate choice dynamics. *J. Exp. Anal. Behav.* **84**, 581–617 (2005).
10. Kennerley, S.W., Walton, M.E., Behrens, T.E., Buckley, M.J. & Rushworth, M.F. Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.* **9**, 940–947 (2006).
11. Lee, D., Conroy, M.L., McGreevy, B.P. & Barraclough, D.J. Reinforcement learning and decision making in monkeys during a competitive game. *Brain Res. Cogn. Brain Res.* **22**, 45–58 (2004).
12. Kim, S., Hwang, J., Seo, H. & Lee, D. Valuation of uncertain and delayed rewards in primate prefrontal cortex. *Neural Netw.* **22**, 294–304 (2009).
13. Maass, W., Natschläger, T. & Markram, H. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* **14**, 2531–2560 (2002).
14. Jaeger, H., Lukosevicius, M., Popovici, D. & Siewert, U. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Netw.* **20**, 335–352 (2007).
15. Verstraeten, D., Schrauwen, B., D'Haene, M. & Stroobandt, D. An experimental unification of reservoir computing methods. *Neural Netw.* **20**, 391–403 (2007).
16. Sussillo, D. & Abbott, L.F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).
17. Bertschinger, N. & Natschlager, T. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Comput.* **16**, 1413–1436 (2004).
18. Langton, C.G. Computation at the edge of chaos: phase transitions and emergent computations. *Physica D* **42**, 12–37 (1990).
19. Seo, H. & Lee, D. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *J. Neurosci.* **27**, 8366–8377 (2007).
20. Seo, H., Barraclough, D.J. & Lee, D. Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cereb. Cortex* **17**, i110–i117 (2007).
21. Seo, H., Barraclough, D.J. & Lee, D. Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. *J. Neurosci.* **29**, 7278–7289 (2009).
22. Barraclough, D.J., Conroy, M.L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* **7**, 404–410 (2004).
23. Lapish, C.C., Durstewitz, D., Chandler, L.J. & Seamans, J.K. Successful choice behavior is associated with distinct and coherent network states in anterior cingulate cortex. *Proc. Natl. Acad. Sci. USA* **105**, 11963–11968 (2008).
24. Sigala, N., Kusonoki, M., Nimmo-Smith, I., Gaffan, D. & Duncan, J. Hierarchical coding for sequential task events in the monkey prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **105**, 11969–11974 (2008).
25. Jin, D.Z., Fujii, N. & Graybiel, A.N. Neural representation of time in cortico-basal ganglia circuits. *Proc. Natl. Acad. Sci. USA* **106**, 19156–19161 (2009).
26. Trefethen, L.N. & Embree, M. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators* (Princeton University Press, Princeton, New Jersey, 2005).
27. Murphy, B.K. & Miller, K.D. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron* **61**, 635–648 (2009).
28. Ganguli, S., Huh, D. & Sompolinsky, H. Memory traces in dynamical systems. *Proc. Natl. Acad. Sci. USA* **105**, 18970–18975 (2008).
29. Goldman, M.S. Memory without feedback in a neural network. *Neuron* **61**, 621–634 (2009).
30. Schneidman, E., Berry, M.J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–1012 (2006).
31. Brunel, N., Hakim, V., Isope, P., Nadal, J.-P. & Barbour, B. Optimal information storage and the distribution of synaptic weights: perceptron versus purkinje cell. *Neuron* **43**, 745–757 (2004).
32. Ganguli, S. *et al.* One-dimensional dynamics of attention and decision making in LIP. *Neuron* **58**, 15–25 (2008).
33. Duncan, J. An adaptive coding model of neural function in prefrontal cortex. *Nat. Rev. Neurosci.* **2**, 820–829 (2001).
34. Rigotti, M., Rubin, D.B.D., Wang, X.-J. & Fusi, S. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front. Comput. Neurosci.* **4**, 24 (2010).
35. Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S. & Cohen, J.D. Conflict monitoring and cognitive control. *Psychol. Rev.* **108**, 624–652 (2001).
36. Holroyd, C.B. & Coles, M.G.H. The neural basis of human error processing: reinforcement learning, dopamine and error-related negativity. *Psychol. Rev.* **109**, 679–709 (2002).
37. Wallis, J.D. & Kennerley, S.W. Heterogeneous reward signals in prefrontal cortex. *Curr. Opin. Neurobiol.* **20**, 191–198 (2010).
38. Platt, M.L. & Glimcher, P.W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238 (1999).
39. Roitman, J.D. & Shadlen, M.N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).
40. Bayer, H.M. & Glimcher, P.W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
41. Rainer, G. & Miller, E.K. Time course of object-related neural activity in the primate prefrontal cortex during a short-term memory task. *Eur. J. Neurosci.* **15**, 1244–1254 (2002).
42. Machens, C.K., Romo, R. & Brody, C.D. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J. Neurosci.* **30**, 350–360 (2010).
43. Shafi, M. *et al.* Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience* **146**, 1082–1108 (2007).
44. Curtis, C.E. & Lee, D. Beyond working memory: the role of persistent activity in decision making. *Trends Cogn. Sci.* **14**, 216–222 (2010).
45. Passingham, D. & Sakai, K. The prefrontal cortex and working memory: physiology and brain imaging. *Curr. Opin. Neurobiol.* **14**, 163–168 (2004).
46. Lebedev, M.A., Messinger, A., Kralik, J.D. & Wise, S.P. Representation of attended versus remembered locations in prefrontal cortex. *PLoS Biol.* **2**, e365 (2004).
47. Funahashi, S., Chafee, M.V. & Goldman-Rakic, P.S. Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature* **365**, 753–756 (1993).
48. Rainer, G., Rao, S.G. & Miller, E.K. Prospective coding for objects in primate prefrontal cortex. *J. Neurosci.* **19**, 5493–5505 (1999).
49. Brody, C.D., Hernandez, A., Zainos, A. & Romo, R. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cereb. Cortex* **13**, 1196–1207 (2003).
50. Bromberg-Martin, E.S., Matsumoto, M., Nakahara, H. & Hikosaka, O. Multiple timescales of memory in lateral habenula and dopamine neurons. *Neuron* **67**, 499–510 (2010).

© 2011 Nature America, Inc. All rights reserved.

ONLINE METHODS

Animal preparation and electrophysiological recording. All of the data were collected using the same behavioral task and electrophysiological techniques. These techniques have been described previously[19–21]. We used six rhesus monkeys (five male and one female). The animal's head was fixed during the experiment and eye movements were monitored at a sampling rate of 225 Hz with a high-speed eye tracker (Thomas Recording). Animals performed an oculomotor free-choice task[22] (matching pennies; **Fig. 1a**). Trials began with the animal fixating a small yellow square (0.9° × 0.9°) displayed at the center of the computer screen for a 0.5-s fore-period. Two identical green disks were presented at 5° eccentricity in diametrically opposed locations along the horizontal meridian for a 0.5-s delay period. The extinction of the central target signaled the animal to shift its gaze toward one of the targets within 1 s. After the monkey maintained its fixation on the chosen peripheral target for 0.5 s, a red ring appeared around the target selected by the computer. The animal was rewarded only if it chose the same target as the computer, which simulated a rational decision maker in the matching pennies game trying to minimize the animal's expected payoff. Before each trial, the computer made a prediction for the animal's choice by computing the conditional probabilities for the animal to choose each target given its choices and rewards in the preceding four trials. The computer made a random choice if the probabilities were consistent with unbiased behaviors, otherwise it would bias its selection against the prediction. Single-unit activity was recorded using a five-channel multi-electrode recording system (Thomas Recording) from three cortical regions: the ACCd[19] (area 24c, two male monkeys, 8–12 kg), DLPFC[20,22] (anterior to the frontal eye field; four male and one female monkeys, 5–12 kg) and LIP[21] (two male and one female monkeys, 5–11 kg). All the neurons were recorded without pre-screening. The placement of the recording chamber was guided by magnetic resonance images and confirmed by metal pins inserted in known anatomical locations at the end of the experiment in some animals. In three animals, two recording chambers were used for simultaneous recording of DLPFC and LIP. All the experimental procedures were approved by the Institutional Animal Care and Use Committee at Yale University and conformed to the Public Health Services Policy on Humane Care and Use of Laboratory Animals and the Guide for the Care and Use of Laboratory Animals.

Multiple regression analysis of memory traces. To estimate the memory traces $f(n,k)$ from the observed neuronal firing rates and sequence of rewards, we computed the firing rates in each trial in 12 time intervals of 250 ms each (**Fig. 1a**). The following model was used to fit the firing rates. The firing rate of a neuron depends on the trial epoch k, following the epoch code $g(k)$; after the outcome is revealed (feedback period) in each trial, the firing rate is changed by an amount of $+f(n',k)$ for reward and $-f(n',k)$ for no reward, where n' is the number of trials elapsed since that outcome. The effects of outcomes in successive trials are additive. The firing rate $\mathrm{FR}(n,k)$ is thus described by

$$\mathrm{FR}(n,k) = g(k) + \sum\nolimits_{n'=0:5} f(n',k)\mathrm{Rew}(n-n') + \mathrm{noise} \qquad (3)$$

where the index k labels the epoch ($k = 1,\ldots,12$) and the indices n and n' label trials. The effect of reward extends up to five trials ($n' = 0,\ldots,5$), while the index n runs over all N trials available in each neuron recording (starting after the first five trials, $n = 6,\ldots, N$). To determine $f(n,k)$ and $g(k)$, we applied a multiple regression model by using the known $\mathrm{FR}(n,k)$ and $\mathrm{Rew}(n)$ (+1/−1 for reward/no reward). Note that the epoch code $g(k)$ depends on the twelve different epochs within a trial, whereas the reward $\mathrm{Rew}(n)$ depends only on trial number. As a consequence, the regression can be applied separately for each epoch. For a fixed epoch k, the seven unknown variables $g(k)$, $f(0,k)$, $f(1,k)$, $f(2,k)$, $f(3,k)$, $f(4,k)$ and $f(5,k)$ can be determined by using the known values of $\mathrm{FR}(n,k)$ and $\mathrm{Rew}(n)$ in $N - 5$ trials ($n = 6,\ldots, N$). Using a parsimonious matrix notation and omitting the epoch label k, equation (3) can be rewritten as

$$\mathrm{FR} = \mathrm{Rew}\bullet f + \mathrm{noise} \qquad (4)$$

where the vector of the known firing rates FR is equal to

$$\mathrm{FR} = [\mathrm{FR}(6,k), \mathrm{FR}(7,k),\ldots, \mathrm{FR}(N,k)]^T \qquad (5)$$

The seven unknown variables have been rewritten by a single vector **f**

$$\mathbf{f} = [g(k), f(0,k), f(1,k), f(2,k), f(3,k), f(4,k), f(5,k)]^T \qquad (6)$$

The matrix Rew is known, given by

$$\mathrm{Rew} = \begin{pmatrix} 1 & \mathrm{Rew}_6 & \mathrm{Rew}_5 & \mathrm{Rew}_4 & \mathrm{Rew}_3 & \mathrm{Rew}_2 & \mathrm{Rew}_1 \\ 1 & \mathrm{Rew}_7 & \mathrm{Rew}_6 & \mathrm{Rew}_5 & \mathrm{Rew}_4 & \mathrm{Rew}_3 & \mathrm{Rew}_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & \mathrm{Rew}_N & \mathrm{Rew}_{N-1} & \mathrm{Rew}_{N-2} & \mathrm{Rew}_{N-3} & \mathrm{Rew}_{N-4} & \mathrm{Rew}_{N-5} \end{pmatrix} \qquad (7)$$

Because the sequence of rewards is nearly random and N is large, different columns of the matrix Rew are nearly orthogonal. This implies that the matrix product ($\mathrm{Rew}^T{\cdot}\mathrm{Rew}$) is well conditioned and that the solution f_{sol} minimizing the variance of the noise (or squared error) is robust and given by

$$f_{\mathrm{sol}} = (\mathrm{Rew}^T \bullet \mathrm{Rew})^{-1} \bullet \mathrm{Rew}^T \bullet \mathrm{FR} \qquad (8)$$

This expression is used to obtain the results. The confidence intervals for f_{sol} are derived from the residual errors according to the Matlab (Mathworks) function regress.

The matrix product ($\mathrm{Rew}^T{\cdot}\mathrm{Rew}$) is approximately proportional to the identity matrix. When $\mathrm{Rew}^T{\cdot}\mathrm{Rew} = I$, the filter is equal to the firing rate averaged over all trials, where the average is conditioned on the past rewards. This is equivalent to the cross-correlation between the input (rewards) and output (firing rates) and its application would correspond to a reverse correlation method, commonly used in the analysis of sensory neural coding. Here, however, we only showed results from the multiple regression analysis. For simplicity, we used an average over all trials as the definition of epoch code $g(k)$ in the main text, making use of the above approximation.

Exponential memory traces and model selection. The model considered here is similar to that of equation (3), but we assumed that memory traces are exponential function $\mathrm{ex}(t)$ rescaled by the epoch code $g(k)$.

$$\mathrm{FR}(n,k) = g(k) + g(k)\sum\nolimits_{n'=0.5} \mathrm{ex}(t)\mathrm{Rew}(n-n') + \mathrm{noise} \qquad (9)$$

The filter f considered in equation (3) is replaced by $g(k){\cdot}\mathrm{ex}(t)$. We considered two different exponential functions, a single exponential and the sum of two exponentials.

$$\mathrm{ex}_1(t) = Ae^{\frac{-t}{\tau}} \qquad (10)$$

$$\mathrm{ex}_2(t) = A_1 e^{\frac{-t}{\tau_1}} + A_2 e^{\frac{-t}{\tau_2}} \qquad (11)$$

where $\tau_1 < \tau_2$. The physical time t depends on all indices k, n and n' because the time elapsed between different epochs and between successive trials is variable, due to the variability in the time taken by the animal to start a trial and to make a saccade to one of the two targets. On the basis of the time stamps generated during the experiment, we computed the physical time $t = t(n,k,n')$ as the difference between the time corresponding to a given trial and epoch (n,k) and the time corresponding to the feedback epoch of n' trials in the past (up to five trials). Note that the memory trace f obtained by the multilinear regression is not computed in physical time. In that case, we assumed that the saccade reaction time of the animal in all trials is equal to 120 ms (average) and that the time elapsed between the initiation of two successive trials is 3.4 s (median).

The epoch code $g(k)$ was fixed by the firing rates averaged across trials, whereas the parameters of the exponential function (two parameters (A,τ) when using equation (10) and four parameters (A_1,τ_1, A_2,τ_2) when using equation (11)) were estimated using a nonlinear curve-fitting procedure, implemented by the Matlab function fminsearch, minimizing the variance of the noise (sum of squared errors) in equation (9). Fitting was repeated ten times for each neuron and each model in the search for a global minimum of the error. Any parameters resulting in unrealistic values were discarded, such as negative values of τ, τ_1 or τ_2, values of

τ larger than 20 trials, and the absolute value of A or ($A_1 + A_2$) larger than 4. We determined the parameters for all neurons in both exponential models, single and double exponential and denoted the corresponding square errors by $\sigma_1{}^2$ and $\sigma_2{}^2$, respectively. We also computed the variance of firing rate, $\sigma_0{}^2$, as the square error for a zero filter model, that is, ex = 0 or FR = g + noise. Among the three models, the selection of the appropriate one for each neuron was determined according to the Bayesian information criterion (BIC)

$$\mathrm{BIC}_i = m\log(\sigma_i^2) + p_i\log(m) \tag{12}$$

where p_i denotes the number of parameters in the model, and $p_0 = 1$, $p_1 = 3$, $p_2 = 5$, for 0, 1 and 2 exponential fit, respectively (note that the variance σ_i^2 is also a parameter), and m is the number of data points ($m = 12(N - 5)$; 12 epochs and $N - 5$ trials for each neuron). The model with the minimum BIC was chosen for each neuron. As a control of the fitting procedure, we reshuffled the label n in the firing rates FR(n,k), assigning to each firing rate the value of a random trial, and we repeated the entire procedure.

Reinforcement learning fit of behavior. We applied a standard reinforcement learning model[5], separately for each recording session, to analyze how the animal's choice was influenced by the outcomes of its previous choices. For example, when right target R was chosen in trial t, the value function for R, denoted by $Q_R(t)$, was updated according to

$$Q_R(t+1) = Q_R(t) + \alpha[\mathrm{Rew}(t) - Q_R(t)] \tag{13}$$

where Rew(t) denotes the reward received by the animal in trial t, and the term inside square is commonly defined as the reward prediction error; that is, the discrepancy between the actual reward and the expected reward. A similar equation holds for the left value function $Q_L(t)$. The probability that the animal would choose the rightward target in trial t, $P_R(t)$, was determined by the SoftMax transformation

$$P_R(t) = \frac{\exp(\beta Q_R(t))}{\exp(\beta Q_L(t)) + \exp(\beta Q_R(t))} \tag{14}$$

where β, referred to as the inverse temperature, determines the randomness of the animal's choices. Model parameters (α,β) were estimated separately for each recording session by using a maximum likelihood procedure, where the likelihood is the product of probabilities in all trials (equation (14)), in each trial using R or L according to the actual monkey's choice. The parameter values maximizing the likelihood were found by using the Matlab function fminsearch. The significance of the estimation was assessed, for each session, by constructing 100 surrogate sessions, each one obtained by reshuffling of the order of trials. The distribution of 100 maximum likelihoods obtained by the estimation procedure was then compared with the maximum likelihood of the non-reshuffled case, which was considered to be significant if not smaller than the five largest reshuffled likelihoods.

Value functions and reward prediction error signals can be related to the exponential filters estimated for individual neurons. If a single value function (for a given stimulus/action) and a single reward (delivered at time zero) are considered, the solution of equation (13) can be approximated by an exponential response, $Q(t) = (1 - \frac{1}{\tau})^t \sim \exp(\frac{-t}{\tau})$, provided that τ is larger than one trial. When a sequence of rewards is delivered instead of a single one, the value is a superposition of the exponential responses for each reward.

© 2011 Nature America, Inc. All rights reserved.