

Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal

Hannah M. Bayer^{1,2} and Paul W. Glimcher^{1,*}

¹Center for Neural Science
New York University
New York, New York 10003

²Center for the Decision Sciences
Columbia University
New York, New York 10027

Summary

The midbrain dopamine neurons are hypothesized to provide a physiological correlate of the reward prediction error signal required by current models of reinforcement learning. We examined the activity of single dopamine neurons during a task in which subjects learned by trial and error when to make an eye movement for a juice reward. We found that these neurons encoded the difference between the current reward and a weighted average of previous rewards, a reward prediction error, but only for outcomes that were better than expected. Thus, the firing rate of midbrain dopamine neurons is quantitatively predicted by theoretical descriptions of the reward prediction error signal used in reinforcement learning models for circumstances in which this signal has a positive value. We also found that the dopamine system continued to compute the reward prediction error even when the behavioral policy of the animal was only weakly influenced by this computation.

Introduction

Over the course of the last decade, significant theoretical and empirical evidence has accumulated suggesting that the dopamine neurons of the mammalian midbrain play a role in reinforcement learning. At a theoretical level, reinforcement learning systems have now been carefully studied, and the basic computational properties of simple systems of this type are largely understood. At an empirical level, it has been shown that the activity of dopamine neurons when animals undergo classical or operant conditioning is compatible with the notion that these neurons carry a *reward prediction error* signal of the type required by theoretical models of reinforcement learning. Very little quantitative empirical evidence, however, has been gathered to (1) determine the exact reward-related computation performed by these neurons, (2) test the extent to which dopamine activity alone could actually supply the reward prediction error signal required by reinforcement learning models, and (3) test the extent to which dopamine activity may shape behavior when animals rely on strategies that only weakly involve reinforcement learning.

Nearly all reinforcement learning models share a few critical features. The models typically assume that while

learning in a task, for example, during operant conditioning, the system maintains an estimate of the average value of each possible action. These estimates of value, or *value functions*, encode the average amount of reward that each action has produced in the past. At each moment in time the animal, or more generally the agent, is presumed to use a set of these value functions to predict the rewards or punishments that any action will yield. Any reward actually obtained after an action is complete is then compared to this prediction. The difference between these two values is typically referred to as the *reward prediction error*. The reward prediction error is then multiplied by a constant ranging between 0 and 1, the *learning rate constant*, and the product of this operation is then added to the preexisting value function in order to yield an updated and presumably more accurate reward prediction. Theoretical work indicates that systems having these basic properties, given a set of reasonable constraints, accurately approximate the true values of impending rewards in an environment that does not change or changes only slowly.

The reward prediction error signal thus plays a critical role in reinforcement learning models. After every action or event in the environment, the reward obtained is compared to the reward that was predicted. If these two quantities are equal, then the reward prediction error is zero; if they are not, the prediction error is nonzero and can be used to update the value function. Thus, an animal that receives an unpredictable juice reward should experience a positive reward prediction error after each reward. In contrast, if the animal predicts that a lever press will be followed by 1 ml of juice, the delivery of 1 ml of juice should produce no reward prediction error. Importantly, however, systems of this type do not necessarily account for all of the behavior produced by all animals under all circumstances. If an animal faces an environment in which a strategy of alternating sequentially between two responses yields a reward on every trial, then the output of a reinforcement learning system *may* be only part of the information required to maximize rewards. Under these conditions a simple (first order) reinforcement learning system might use the reward prediction error after each action to correctly deduce that both responses yielded rewards with a high probability but would be unable to determine that it was the sequential structure of the responses that actually yielded the rewards. This observation may suggest the existence of higher-order reinforcement learning mechanisms that integrate the reward prediction error signal with additional information about the structure of the environment, a point developed in Sutton and Barto's actor-critic model (Sutton and Barto, 1981).

Recently acquired physiological evidence strongly suggests a parallel between the activity of dopamine-containing neurons in the midbrain and the reward prediction error term employed in reinforcement learning models. In thirsty monkeys, for example, the dopamine neurons of the substantia nigra pars compacta and

*Correspondence: glimcher@cns.nyu.edu

ventral tegmental area produce bursts of action potentials in response to unexpected juice rewards and suppressions in firing rate when expected rewards do not occur (Mirenowicz and Schultz, 1994; Schultz et al., 1993). In classical conditioning tasks, the first time a stimulus is followed by a reward, the dopamine neurons respond with a burst of action potentials after the reward. However, after repeated presentations of the stimulus with the reward, the dopamine neurons respond only after the presentation of a temporally unpredictable stimulus that signals a later reward. They do not respond after the reward itself, which has been accurately predicted because of the occurrence of the preceding stimulus. Similar results suggest a role for the dopamine neurons in learning during blocking experiments and during operant conditioning as well (Hollerman and Schultz, 1998; Kawagoe et al., 2003; Nakahara et al., 2004; Satoh et al., 2003; Waelti et al., 2001). When animals must learn the values of actions in these tasks, the activity of the midbrain dopamine neurons carries a reward prediction error-like signal. In summary then, these results and many others are consistent with the hypothesis that the dopamine signal is a physiological instantiation of the reward prediction error signal: the difference between the reward an animal expects to receive and the reward it actually receives (Schultz et al., 1997).

We were interested in extending our understanding of the relationship between these dopamine neurons and the theoretically defined reward prediction error by determining what precise function of previous rewards best predicts the activity of these neurons. To this end, we recorded the activity of midbrain dopamine neurons while overtrained awake-behaving primates performed a saccadic timing task that might or might not have required a reinforcement learning system to solve. We hypothesized that, if we could determine the function that related the reward history of the animal to the activity of the dopamine neurons under these conditions, we might be able to further test the hypothesis that these neurons carry a reward prediction error. Given a more complete picture of the function that relates rewards to dopamine firing rates, we then hoped to ask whether this dopamine signal could, in fact, provide all of the reward prediction error information required by an efficient reinforcement learning system. In other words, we hoped to test the notion that the measured pattern of dopamine activity that we observed was sufficient for implementing a complete reward prediction error of the type called for by theory. Finally, we hoped to behaviorally determine whether overtrained animals of the type typically used in studies of this kind rely on reinforcement learning systems for the production of the overtrained behavior. We hypothesized that, under circumstances in which animals employed an identifiable behavioral policy that did not rely directly on a trial-by-trial reinforcement learning-based estimate of the value of an action, the activity of dopamine neurons and the actions of the animal would be only weakly linked even if the reinforcement learning system continued to compute underlying value functions.

Our findings suggest that the responses of midbrain dopamine neurons do indeed encode a signal of the type required by reinforcement learning models. The

activity of midbrain dopamine neurons during a fixed interval immediately after a reward is received or expected reflects an iteratively computed difference between a weighted average of the magnitudes of recently received rewards and the magnitude of the current reward. In our task, we were even able to establish the learning rate constant for this neuronal computation. When we attempted to determine whether the pattern of activity during this fixed postreward interval was sufficient for a simple reinforcement learning model, however, we found that the representation of the theoretically defined reward prediction error was incomplete. The average firing rate of the dopamine neurons in the postreward interval accurately carries information about positive reward prediction errors but not about negative reward prediction errors; neuronal firing rates behaved as if they rectified the reward prediction error. Although physiological mechanisms can be postulated that would correct this asymmetry, we note that this asymmetry may be computationally important. It may be particularly important if reinforcement learning via the reward prediction error signal carried by dopamine neurons is accomplished by dopamine-driven long-term synaptic potentiation. The possibility that the dopamine signal carries only positive reward prediction errors may imply the existence of an independent system carrying negative reward prediction errors that could drive long-term synaptic depression (Daw et al., 2002). Finally, we found that once our highly trained animals had adopted a simple fixed strategy for performing our task, the activity of the dopamine-reinforcement learning system was only weakly coupled to behavior. We found that the dopamine system continued to compute the reward prediction error even when the behavioral policy of the animal was only weakly influenced by this computation.

Results

Behavior

Our monkey subjects were trained to perform a simple task that yielded rewards having a variety of magnitudes, a saccadic timing task (Figure 1). At the start of each 4 s trial, the monkeys fixated a central stimulus. An eccentric target was then illuminated at a single location, and the animals were free to look at that eccentric target at any time during the 4 s trial. There was, however, only one interval during the trial in which a saccade to the eccentric target would yield a reinforcement, and this interval was further subdivided into five subintervals. The earliest of these subintervals was associated with a 0.2 ml water reward, and each subsequent interval was associated with an additional increment of 0.02 ml of fluid. Thus, if a saccade to the eccentric target was executed at the beginning of the reinforced interval, the subject received a smaller reward, and if the saccade was executed at the end of the interval, the subject received a larger reward. Subjects maximized reward magnitude by determining when the end of this unsignaled rewarded interval occurred and choosing to make a saccade at that time. Finally, the time at which the reinforced interval occurred was shifted unpredictably approximately every 100 trials,

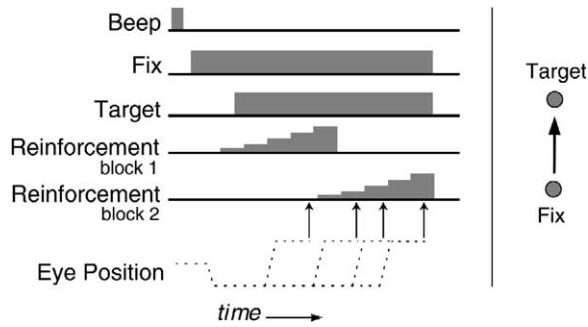


Figure 1. Saccade Timing Task

(Left) The events of an individual trial as a function of time. Animals were reinforced for executing a saccade during an unsignaled temporal window. The height of the reinforcement cartoon indicates that there were five intervals during this window, each of which was associated with an increasing reward size. The delay before the rewarded temporal window was shifted between blocks of trials without cuing the animals. (Right) The spatial configuration of the fixation and target light-emitting diodes.

forcing the subjects to encounter shifting levels of rewards as they attempted to track the rewarded interval.

The animals were able to perform this task in a manner that reliably yielded water rewards. Figure 2A plots delays (in log seconds) selected by a single subject during a typical session as a function of trial number. The shaded rectangles on the graph indicate the time of the reinforced intervals within each block. The darkness gradient from the top to the bottom of each block indicates the underlying reward gradient, with the darker

shading at the top of each block indicating the largest available reward. Note that the subject manages, in general, to select delays that yield rewards, but also note that the size of the reward earned on each trial varies across the range of possible reward values.

To quantify this behavior, Figure 2B plots the average delay adopted by the subjects across all of the block delays that we tested. Each point represents the average delay during the last 50 trials of each block. Each point incorporates at least ten blocks from two monkeys. The error bars indicate the standard error of the mean but, on this scale, do not extend beyond the borders of the points. If the animals had perfectly maximized reward during the second half of each block, then the points would all fall along the main diagonal. Note that in general the responses of the animals approximate this strategy, although for blocks presented at longer delays the animals tended to produce their movement earlier than was optimal. Also note the size of the standard errors, which indicate the low level of variability in the animal's responses at the end of each of these blocks.

To better characterize the trial-by-trial actions of the subjects at the end of each block, Figure 2C plots the correction a subject should have adopted as a function of the correction they did adopt for the block labeled "C" in Figure 2B. Perfect corrections would lie along the main diagonal. The rewarded interval during this block was the earliest possible interval in the task, so errors were always due to waiting too long, and corrections were always to decrease reaction time. However, note that when subjects made responses that lay outside the rewarded interval they tended to produce ac-

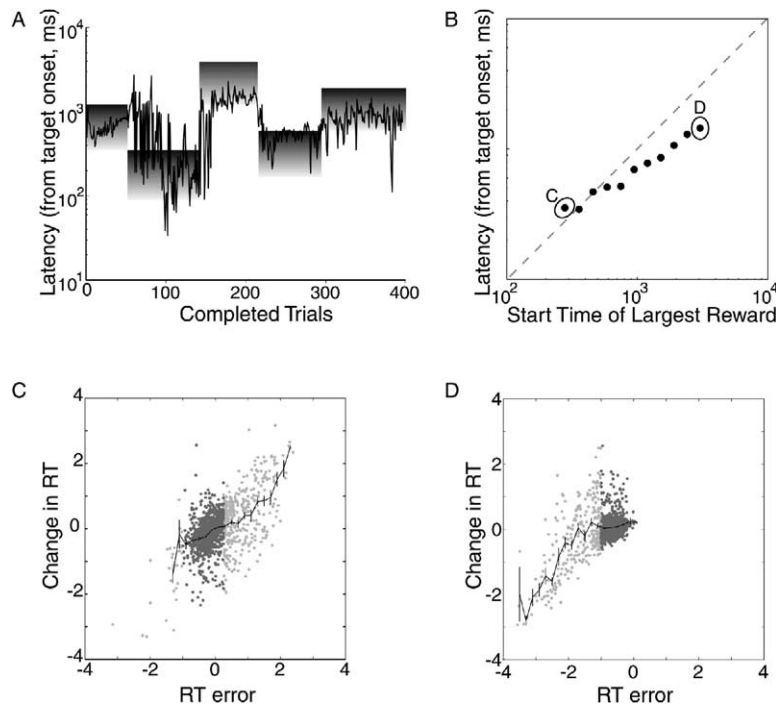


Figure 2. Animals Choose Saccadic Latencies for which They Will Be Reinforced

(A) Saccadic latencies plotted sequentially from a single behavioral session. The gray squares represent the extent of the rewarded temporal window.

(B) Average saccadic latency for the last fifty trials of each block, plotted as a function of the time at which the interval associated with the largest reward size began (error bars showing standard deviation fit inside the points).

(C) Log of the change in reaction time from the current to the next trial (ΔRT) plotted as a function of the log of the difference between the reaction time on the current trial and the reaction time that would have provided the largest volume of juice (RT error). Includes only trials from blocks in which the best reaction time was the earliest one the monkeys ever experienced (including the data used to compute the average point labeled "C"). Rewarded trials are in dark gray, unrewarded trials are in light gray, and mean and standard error are plotted in black.

(D) ΔRT plotted as a function of RT error, including only trials from blocks in which the best reaction time was the latest one the monkeys ever experienced (including the data used to compute the average point labeled "D"). Rewarded trials are in dark gray, unrewarded trials are in light gray, and mean and standard error are plotted in black.

curate corrections that brought them into the rewarded interval. Once subjects were within the overall rewarded interval, however, corrections were only weakly correlated with errors (dark points). This pattern was also true for subjects' reaction time choices during blocks of trials in which the rewarded interval was the latest possible interval in the task, as shown in Figure 2D. This pattern was observed for all blocks of trials, not just the earliest and the latest.

In summary, the monkeys were able to choose a range of saccadic latencies that provided a reward on every trial, usually the largest ones. However, as the interval occurred later in the trial, the animals were more likely to choose a range of latencies that resulted in smaller rewards (Figure 2B). They appeared to prefer a smaller reward in exchange for a shorter waiting time. Even during the earliest rewarded intervals, it appeared that the monkeys were more sensitive to whether or not a reaction time resulted in a reward than the size of that reward (Figures 2C and 2D). These behavioral results suggest that, in this task, the animals learned to find the saccadic latency for which they would receive reinforcement, although the trial-by-trial mechanism by which they accomplished this adjustment behaviorally seemed not to reflect a precise compensatory process driven by the magnitude of the reward received on the preceding trial.

Activity of Individual Dopamine Neurons

In order to study the relationship between the responses of midbrain dopamine neurons and reward history, we recorded from 81 neurons while monkeys performed the saccade timing task. For this report, we analyzed the data from the 50 neurons for which we had gathered at least 100 trials of the task (mean, 292 trials; range, 100 to 1100 trials; SD, 186 trials). Guide tubes and electrodes were placed under real-time ultrasonographic guidance (Glimcher et al., 2001), and electrodes were lowered through ventroposterior medial thalamus prior to entering areas where dopamine neurons were presumed to be located. During recording sessions, neurons were classified as dopaminergic based on three criteria: they had relatively long triphasic action potentials, their baseline firing rates were relatively low (mean, 5.3 ± 1.5 impulses per second), and they showed a phasic response to unpredicted fluid rewards (Figures 3A and 3B). A subset of these neurons, which were typical of the population, were histologically localized to the substantia nigra pars compacta and the ventral tegmental area (Figure 3C).

While monkeys were performing the saccade timing task, all the neurons produced a baseline rate of activity that was modulated at two points during the trial. The first was a burst in activity after the auditory tone that signaled trial onset, and the second was an increase or decrease in rate following the completion of a saccade immediately after the onset of a period in which the delivery of a reward could be anticipated. Figure 4 shows the average response of a single neuron during the task. Averages were produced by aligning individual trials at the time of the tone (left) and the time of reward delivery (right). Plotted above the averages are rasters for a subset of the individual trials that were

used to compute the averages. After the auditory tone, this neuron showed a phasic increase in firing rate of about ten impulses per second, independent of all reward contingences. The response patterns of this neuron were typical of the neurons that we recorded from: most neurons showed an increase in firing rate of about five to ten impulses per second following the auditory tone. These responses were fairly stereotyped under these conditions. We found no evidence that this signal predicted the magnitude of the upcoming reward, a fact which likely reflects the temporal uncertainties of our particular task (Schultz et al., 1997).

In contrast, the responses of this neuron following the delivery of rewards were much more variable. These responses were largest during trials in which the monkey received a large reward after having failed to receive a reward for several trials, a situation that principally arose when the monkeys were searching for the rewarded interval at the beginning of a new block and receiving rewards only infrequently. Following the delivery of a reward under these conditions, the neurons would reach peak firing rates as high as five or six times their baseline firing rates. During long stretches of rewarded trials, the reward that would elicit a large response early in the group of sequentially rewarded trials would elicit only a small response after periods of consistent reinforcement. In general, when the reward a monkey received on a particular trial was greater than the reward he received on the previous trial, the neurons were much more active than when the reward the animal received was smaller than the reward received on the previous trial (Figure 4, right side). This observed pattern of activity was consistent with a simple error signal, suggesting that in this task the firing rates of dopamine neurons following reward delivery carried a reward prediction error of some form.

In order to determine what function of previous rewards best predicted the activity of these neurons, we used a linear regression to determine what combination of previous rewards best predicted neuronal firing rates. This analysis eliminated the need for any assumptions, other than linearity, about the form of the calculation that these neurons might reflect with regard to reward history. It provided a set of weights (or β values) of the following form:

$$\beta_0 \times (R_t) + \beta_1 \times (R_{t-1}) + \beta_2 \times (R_{t-2}) + \dots + \beta_{10} \times (R_{t-10}) \quad (1)$$

where R_t was the amount of fluid reward obtained on the current trial, R_{t-1} was the amount of fluid reward obtained on the previous trial, and so on. The regression gave a single set of β values that described the best rule for predicting the firing rate for any particular trial from the recent history of rewards. The β values thus yield a weighting function that explained how to additively combine the measured values of the last ten rewards in order to best account for dopamine neuron firing rates after the most recent reward had been received. For example, if the neuronal firing rate reflected an error signal that was simply the difference between the current and the previous reward, then it would be expected that the regression would yield a positive

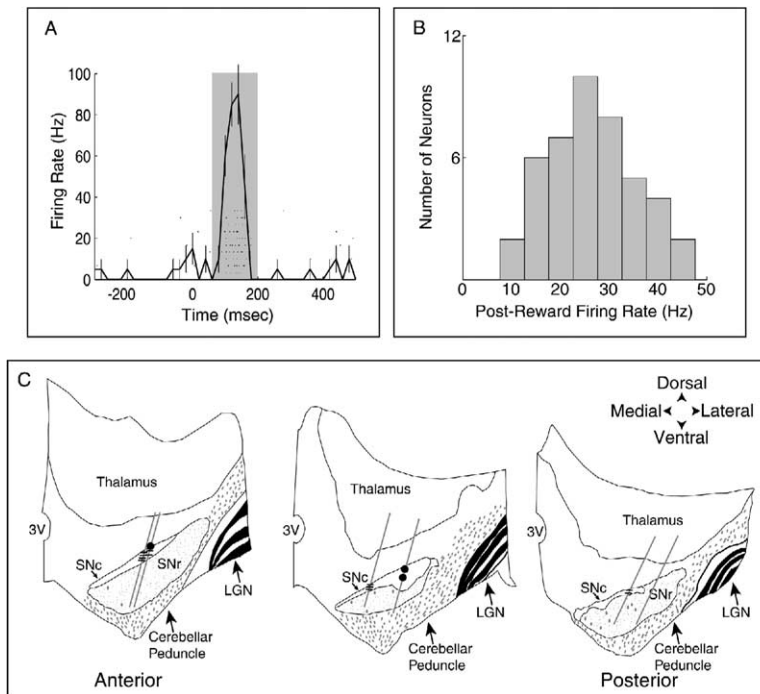


Figure 3. Dopamine Neurons Show a Characteristic Response to Unpredicted Rewards (A) The average response of a single dopamine neuron to the delivery of an unpredicted reward, aligned to the time of reward delivery. Error bars are standard deviation of the mean, and time bins are 20 ms long ($n = 10$ trials).

(B) The distribution of average firing rates during a 150 ms interval starting 75 ms after the delivery of an unpredicted reward, shown in gray in (A), for all neurons in the population. Mean = 26 Hz; standard deviation = 9 Hz ($n = 46$).

(C) Histological localization of a subset of dopamine neurons from this report. Circles are locations of marking lesions placed at the location of recorded neurons, and dashes are estimated locations for additional neurons (where no lesions were made). Distortions in these drawings accurately reflect a significant distortion of the anatomy observed following the perfusion process. The animal had suffered from a blockage of the lateral ventricle during the period in which the marking lesions were made. As a result, the tissue was not sliced in the vertical plane, and sections differ significantly from canonical images.

value, for example, 1.0 and an identical negative value, -1.0 , for β_0 and β_1 , respectively, and 0 for all other β values. However, if the firing rates of the neurons reflected the difference between the current reward and the average of the values of last ten rewards, then the regression would be expected to yield a positive value for β_0 and negative values of one-tenth that magnitude for β_1 through β_{10} .

The results of the regression analysis for a single dopamine neuron are shown in Figure 5A. The weight for the current reward is a large positive number, while the weights for each of the preceding five rewards are negative values between -30 and zero, and the weights for rewards more than five trials into the past are indistinguishable from zero. These weights describe the transformation of reward values in milliliters of water

directly into dopamine firing rates represented as spikes per second. By normalizing the values of all the weights by dividing them by the value of the first weight, we can examine whether the relative weights describe a computation that is proportional to a weighted average of previous rewards. The inset in Figure 5A shows the regression coefficients plotted in this fashion, as they might be used to compute a weighted average.

In order to look at the relationship between firing rate and this computation that the neurons were apparently encoding, and to examine the linearity of the relationship between reward history and firing rate, for each trial we multiplied the values of the current and each of the previous ten rewards by the regression-derived (normalized) β values and then compared it to the firing

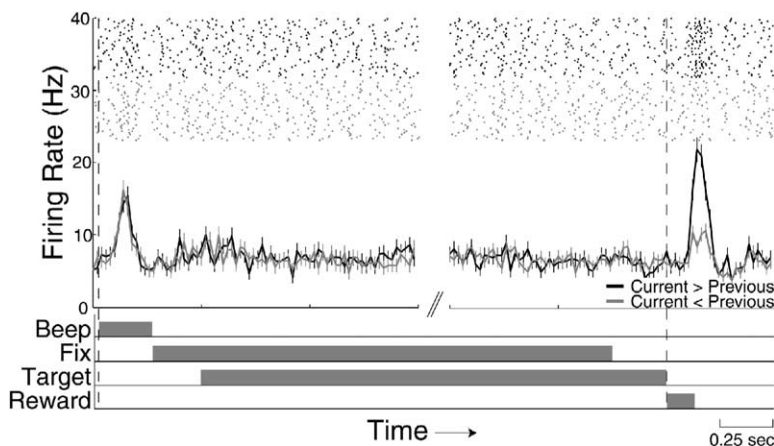


Figure 4. Responses of a Dopamine Neuron during the Saccade Timing Task

(Left) Average response of the neuron aligned to the auditory tone that initiated the trial; error bars represent standard error. (Right) Average response of the neuron aligned to the time of reward delivery; error bars represent standard error. Plotted above the averages are a randomly selected subset of 40 trials from each condition as examples of the raw data that were used to compute the averages. (Both graphs) In black are trials in which there was a large difference between the size of the reward delivered during the trial and the size of the reward during the previous trial ($n = 300$). In gray are trials in which there was a small difference in size between the current and previous rewards ($n = 289$). For reference, the legend underneath represents the events of the trial as a function of time.

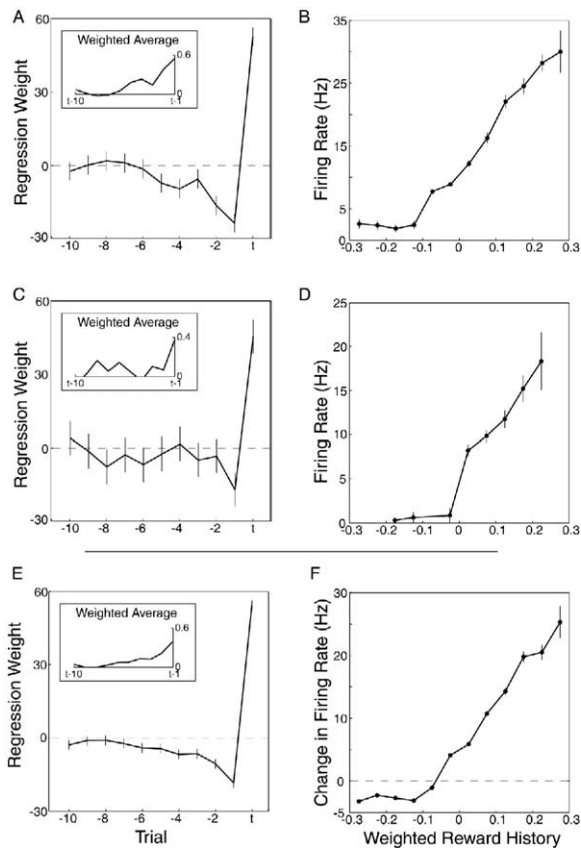


Figure 5. Multiple Linear Regression of Neuronal Firing Rate and Reward History: Single Neuron

(A) Coefficients from multiple linear regression for a single neuron (L041103). (Inset) Last ten coefficients plotted as they would be used to compute a weighted average. Each one is divided by the value of the first coefficient. Error bars represent the 95% confidence intervals. R-squared = 0.50; $p < 0.00001$; $n = 1007$ trials.

(B) Firing rate plotted as a function of weighted reward history. Weighted reward history computed using the coefficients shown in (A) after they have been normalized by dividing all coefficients by the value of the first. Error bars represent standard error.

(C) Coefficients from multiple linear regression for a single neuron (C032504). (Inset) Last ten coefficients plotted as they would be used to compute a weighted average. Error bars represent the 95% confidence intervals. R-squared = 0.42; $p < 0.00001$; $n = 295$ trials.

(D) Firing rate plotted as a function of weighted reward history. Weighted reward history computed using normalized regression coefficients shown in (C). Error bars represent standard error.

(E) Coefficients from multiple linear regression for all neurons combined. (Inset) Last ten coefficients plotted as they would be used to compute a weighted average. Error bars represent the 95% confidence intervals. R-squared = 0.21; $p < 0.0001$; $n = 13919$ trials.

(F) Firing rate plotted as a function of weighted reward history. Weighted reward history computed using normalized regression coefficients shown in (E). Error bars represent standard error.

rate on that trial. Figure 5B shows firing rate plotted as a function of the trial-by-trial weighted average of rewards specified by the linear regression as in Figure 5A. If the underlying function that the regression analyzed had been entirely linear, these data points would all fall along a straight line of slope 1. Interestingly, for all values of this weighted reward history greater than -0.1 , there is a roughly linear relationship with neuronal

firing rate rising to three or four times baseline, suggesting that, in this range, this neuron linearly encodes the weighted averaging calculation described by the β weights. However, for values of this calculation less than -0.1 , there appears to be no modulation in the neuronal response. This lack of modulation seems to occur because the neuron reaches the bottom end of its dynamic range; it produces zero spikes during our measured interval for all large negative values of weighted reward history. These results indicate that the firing rate of this neuron during the rewarded interval is quantitatively correlated with the difference between the current reward and a recency-weighted average of previous rewards for a limited range of reward histories.

Figure 5C shows the set of weights derived by linear regression for another individual neuron. For this neuron, there is also a large positive weight for the current reward and negative weights for the previous rewards that approach zero for rewards farther in the past. This neuron also shows a nearly linear relationship between weighted reward history and firing rate, but again there is a rectification for very negative values of the calculation, suggesting that the firing rate of this neuron during the postreward interval carries a signal encoding this calculation only for positive values (Figure 5D).

Activity of the Neuronal Population

This pattern of neuronal responses, in which firing rate was correlated with the difference between the value of the current reward and a recency-weighted average of the previous rewards, was characteristic of nearly all of our neurons. The responses of 44 out of 50 neurons were significantly ($p < 0.01$) modulated by reward history, with a mean R-squared value of 0.29 (SD = 0.15).

We also performed a single multiple regression analysis on combined data from all of the neurons in our population. In order to combine data from multiple neurons, we normalized the firing rates of each cell by subtracting from them the average baseline firing rate of that cell and compiled the corresponding reward history for each trial. We then performed a single regression of normalized firing rate against reward history for all of these trials. Figure 5E shows the combination of previous rewards that best predicts the change in neuronal firing rate from baseline levels, for all trials from all cells. When this large data set is used, the derived weights describe a signal computed as the difference between the value of the current reward and a recency-weighted average of the values of the last six or seven rewards. Plotting this weighted reward history against firing rate (Figure 5F), there is a linear relationship for positive values of weighted reward history, but not for negative values below about -0.1 . This is similar to the result we show in the single neuron examples, suggesting that the population of dopamine neurons encodes the difference between the current reward and a recency-weighted average of past rewards. However, additional data have not diminished the rectification in the neuronal response for very negative weighted reward history values. The firing rates of these dopamine neurons at the expected time of reward do not encode the difference between the value of the current reward and a recency-weighted average of the values of the last

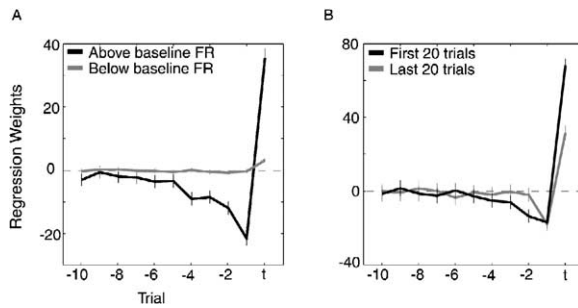


Figure 6. Neuronal Firing Rates Are Better Correlated with Reward History When Firing Rates Are above Baseline, and When There Is Low Correlation between Sequential Rewards

(A) Coefficients from multiple linear regression for all cells. Plotted in black are the results of the regression including only trials with firing rate above baseline (R-squared = 0.16; $p < 0.00001$; $n = 10449$ trials). Plotted in gray are the results of the regression including only trials below baseline (R-squared = 0.03; $p < 0.00001$; $n = 3966$ trials). Error bars represent the 95% confidence intervals. (B) Coefficients from multiple linear regression for all cells. Plotted in black are the results of the regression including only the first 20 trials of each block (R-squared = 0.32; $p < 0.00001$; $n = 3100$ trials). Plotted in gray are the results of the regression including only the last 20 trials of each block (R-squared = 0.09; $p < 0.00001$; $n = 3180$ trials). Error bars represent the 95% confidence intervals.

six or seven rewards when this value is significantly negative.

In order to further examine this apparent rectification, we performed the multiple regression for firing rate against reward history, separating out trials on which the firing rate was below baseline or above baseline. The results of these regressions are shown in Figure 6A. The regression weights for trials with firing rates above baseline (plotted in black) are quite similar to those observed for the population as a whole. However, for trials with reward interval firing rates below baseline, the neuronal firing rate was correlated only with the current reward and a constant offset term (not shown in the figure) reflecting the tonic firing rate of the neuron. There is no significant relationship between firing rate during the reward interval and the difference between the value of the current reward and a recency-weighted average of the values of the last six or seven rewards.

We also examined the possibility that the occurrence of repeated identical rewards might have some effect on the results of our regression, since animals often got the same reward on multiple successive trials. To do this, we performed the multiple regression using only the first 20 trials of each block (when the likelihood of sequential rewards of the same magnitude was low) and the last 20 trials of each block (when the likelihood of sequential rewards of the same magnitude was high). Note that the regression weights for the first 20 trials of each block look similar to the weighting function derived for the complete data set (Figures 6B and 5E). While the regression weights for the last 20 trials of each block also broadly look similar, only two trials contribute significantly to the weighting function. Under these conditions, the recency-weighted average appears to be occurring over a shorter time interval. The R-squared value for the regression using the first 20 tri-

als of each block is greater than that using the last 20 trials of each block.

Neuronal Activity and Choice Behavior

As Figure 2C indicates, immediately after block switches the monkeys tended to search for the rewarded interval by moving toward intermediate delays regardless of the mean delay they were currently employing (longer delays if transitioning from an early reward block or shorter delays if transitioning from a late reward block). Once a reward of any size had been obtained, the animals shifted slowly, and without any obvious trial-by-trial progression, toward delays that yielded larger rewards. In order to determine whether information encoded in the neuronal responses was correlated with these patterns of delay selection, we performed the following additional analyses.

To assess whether firing rate was directly correlated with the behavioral delay selected by the animal on the current trial, we added the temporal subinterval within which the current saccade was produced as an additional parameter to our linear regression. If the firing rates of the dopamine neurons effectively kept track of the difference between the reward just obtained and a weighted average of previous rewards as a function of the delay (or action) that the animals had just produced then this term should have added significant statistical power to our regression. Figure 7A shows the R-squared values for each neuron in the population when the regression included and did not include what delay the animal selected on the subsequent trial. All the cells fall close to the identity line, indicating that neuronal firing rates were not correlated with when the animal made its response.

To ask this same question in another way, we also determined whether neuronal firing rate was better accounted for by a reward history composed only of trials on which the animal had produced a similar action, those trials during which the saccades were all made within the same subinterval. This analysis was intended to determine whether the neurons separately encoded reward information about saccades initiated at different times. Figure 7B shows the R-squared value for each neuron from a regression using this interval-grouped reward history plotted as a function of the R-squared value from the standard reward history regression. For almost every neuron, the interval-grouped regression accounted for much less of the variance in the firing rate. The small but nonzero R-squared values resulting from the interval-grouped reward history represent the amount of variance accounted for by the magnitude of the current reward.

We also performed both of these analyses using information about the timing of the subsequent saccade, in an effort to determine whether these neurons encoded predictive information about the upcoming trial rather than information about the current trial. In neither case did this variable contribute significantly to neuronal firing rate. Taken together, these results suggest that, during our task, the firing rates of the neurons *did not* encode information about the time at which the preceding or subsequent saccade was executed even though the information that they carried clearly related

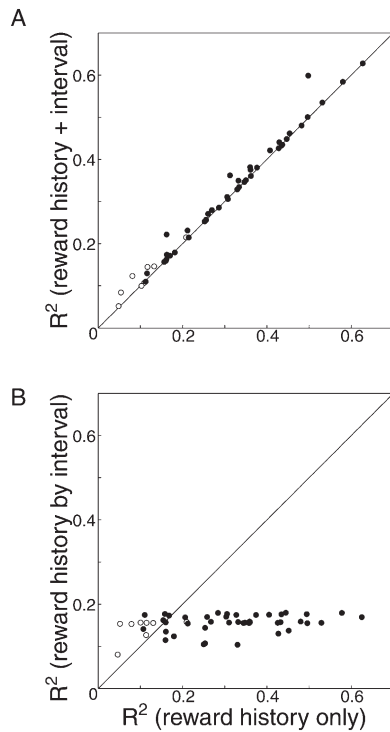


Figure 7. Neuronal Responses Do Not Encode Temporal Properties of the Saccade

(A) R-squared values for individual neurons from regressions with the temporal interval in which the saccade was executed added as an additional variable to the regression plotted as a function of the R-squared values for the regression using reward history only. (B) R-squared values for individual neurons from regressions using only saccades with the same temporal interval in which the saccade was executed to compose the reward history plotted as a function of the R-squared values for the regression using a sequential reward history.

to the value of the saccade irrespective of the time at which it was executed.

In order to examine the relationship between the animals' behavioral choices and the neuronal signals, we plotted the increment in reaction time on the previous trial as a function of the difference between the current reward and the weighted sum of past rewards computed by our regressions (Figures 8A and 8C). We did this separately for blocks of trials with different rewarded intervals in order to ensure that the behaviorally observed differences in time preference did not obscure any effects we might see. (The figures show only the earliest interval and latest interval to represent the range of what was observed.) These graphs allowed us to determine whether there was a relationship between the reward-related signal encoded by the neurons and the choice of reaction time made by the animal on the next trial. Perhaps surprisingly, there was no systematic relationship between reward prediction error and change in reaction time. Thus, the reward-related neuronal signal provided no information about how long the animal had waited before making a movement and appears not to have been used by the animals, in our task, to

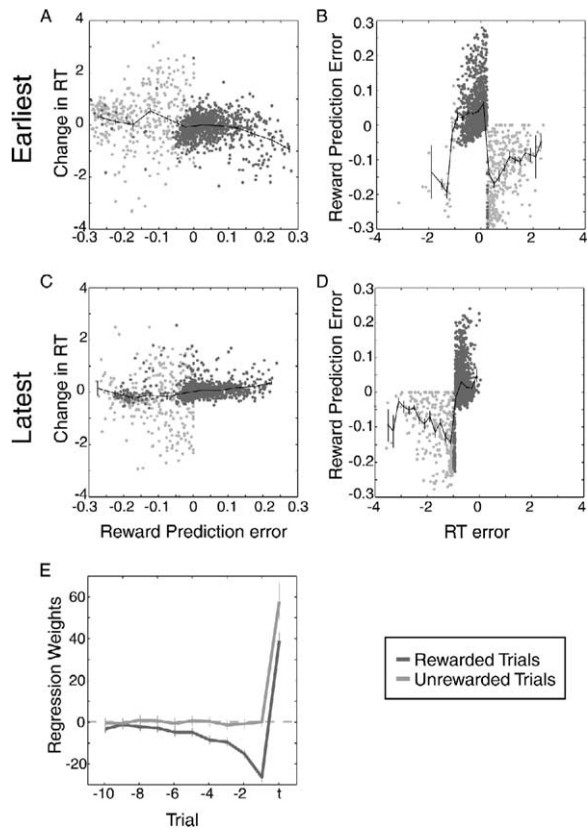


Figure 8. Monkeys Use Information to Perform the Delayed Saccade Task that Is Not Encoded in Neuronal Firing Rates

(All plots) Rewarded trials are in dark gray, and unrewarded trials are in light gray. Mean and standard error are plotted in black. (A) Reward prediction error plotted as a function of RT error for all trials in which the target reaction time was the earliest one the monkeys had ever experienced. (B) Change in reaction time (Δ RT) plotted as a function of reward prediction error for all trials in which the target reaction time was the earliest one the monkeys had ever experienced. (C) Reward prediction error plotted as a function of RT error for all trials in which the target reaction time was the latest one the monkeys had ever experienced. (D) Change in reaction time (Δ RT) plotted as a function of reward prediction error for all trials in which the target reaction time was the latest one the monkeys had ever experienced. (E) Coefficients from multiple linear regression for all cells. Plotted in dark gray are the results of the regression including only rewarded trials (R-squared = 0.43; $p < 0.00001$; $n = 12016$ trials). Plotted in light gray are the results of the regression including only unrewarded trials (R-squared = 0.22; $p < 0.00001$; $n = 2399$ trials). Error bars represent the 95% confidence intervals.

decide how long to wait to make their eye movement during the next trial.

Finally, we examined the relationship between reaction time error and the reward-related signal computed by the neuronally derived regression. Figures 8B and 8D show that there was no systematic relationship between these two quantities either. These plots suggest that, for rewarded trials (in dark gray), there was a weak positive association between reaction time error and reward prediction error but that for unrewarded trials there was no such association. It should be noted, however, that because reward prediction errors do not

map uniquely to optimal changes in reaction time when no reward information is available, they are in principle insufficient feedback for driving changes in reaction time following unrewarded eye movements.

The observation that the correlation between the size of the change in reaction time and the difference between the chosen and best reaction times was strongest when the relationship between reaction time error and reward prediction error was weakest suggests that the animals used information other than the activity of the dopamine neurons to decide when to make their movements in this task. In order to test this hypothesis more explicitly, we performed the multiple regression analysis on all of our neuronal data again, this time separating rewarded trials (shown in dark gray in all panels of Figure 8) and unrewarded trials (shown in light gray in all panels in Figure 8). Note that the regression weights shown in Figure 8E indicate that, for rewarded trials, the firing rate of the neuron encoded information about the reward history, while for unrewarded trials, the firing rate of the neuron encoded information only about the current trial. These data indicate that the dopamine neurons carry information about the reward history primarily for trials during which the animal received a reward in this task. However, the animals appeared to adopt a strategy of changing their behavior most drastically when they did not receive a reward.

Taken together, these results present an interesting dichotomy. The dopamine signal seems to be an efficient way to assess the value of making a saccade to the visual target irrespective of when that saccade is made. From this signal, it would be possible, as will be explained in more detail below, to compute the average value of that movement. However, for deciding precisely when to make the movement, the animals seem to be employing a different mechanism. The behavioral strategy that they employed seems to be one that cannot be accounted for by the pattern of dopamine activity, or perhaps even by the activity of any reinforcement learning system.

Discussion

Relation of Firing Rate and Reward History

We studied the responses of dopamine neurons during a task in which the rewards obtained for taking the same action, making a particular saccadic eye movement, were constantly varying. Immediately after the action was complete, the time at which a reward might or might not be delivered on a given trial, we determined the linear function that related reward history to average neuronal firing rate during this interval. This function, derived by linear regression, computed the difference between the value of the reward received on the current trial and a weighted average of the values of the previous seven rewards. Recent rewards contributed to this average more than earlier rewards, in what approximated an exponentially decaying fashion.

When we used the output of this function, derived from a linear regression, to predict the activity of dopamine neurons, we found an interesting nonlinearity. The output of the function accurately predicted dopamine firing rates when the value of the current reward was

greater than the weighted average of previous rewards. However, when the value of the current reward was significantly less than the weighted average of previous rewards (which occurred principally when the animal received no reward) the dopamine neurons always responded with the same spike rate: 0 Hz.

The function we derived that relates dopamine neuron activity to the reward history associated with the completion of the single action that our animals produced bears a striking resemblance to the exponentially weighted average used by many models of reinforcement learning (Bush and Mosteller, 1955; Rescorla and Wagner, 1972; Sutton and Barto, 1981). In approaches of this type, for example, the widely used model of Sutton and Barto (1998), an animal's current estimate of the value of a state of the environment (V_t) is computed by taking the difference between the value of the most recently acquired reward (R_t) and a weighted sum of the values of previous rewards (V_{t-1}), the *reward prediction error*. The reward prediction is then multiplied by a learning rate constant (α) and used additively to update the animal's estimate of the value of a state:

$$V_t = V_{t-1} + \alpha[R_t - V_{t-1}] \quad (2)$$

In practice, the computation is presumed to be performed iteratively, V being updated incrementally once after every action or stimulus. This iterative process, in which α has a value of less than 1, results in more recent rewards having a larger influence on the value function, V , than rewards obtained farther in the past. More specifically, the iterative calculation yields a weighted average with an exponential rate of decay that is controlled by the size of the α parameter.

A number of researchers have suggested that the activity of dopamine neurons may serve a very specific role in reinforcement learning—it may carry the reward prediction error. While a large body of evidence supports this hypothesis qualitatively, we can test this hypothesis more quantitatively by comparing the function we derived from the dopamine neurons to the functional form of the reward prediction error specified in Equation 1. Figure 9 shows the weighting functions that a theoretical *reward prediction error* ($\alpha[R_t - V_{t-1}]$) would employ. In this particular example, α has been set to 0.5 in gray and 0.7 in black. Note that the general shape of these functions is similar to the empirical functions derived by linear regression of our data set (Figure 5), with a large positive contribution by the reward that has just been received on the current trial, and a negative contribution with an exponential decay for each preceding trial. In summary then, we found that the results of the linear regression described a computation that was quite similar to the computations employed to derive the reward prediction error signal in reinforcement learning models. This further supports the hypothesis that the dopamine neurons of the mid-brain encode, in firing rate, a reward prediction error signal that could be used to drive a reinforcement learning system.

To better relate our data to reinforcement learning theory, we also fit the relationship between dopamine firing rates and reward history using an exponential

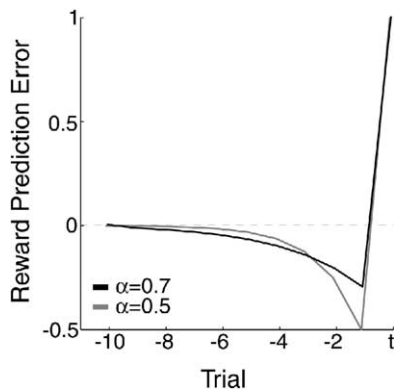


Figure 9. Theoretical Reward Prediction Error Computations
Value of reward prediction error ($\alpha[R_t - V_{t-1}]$) computed using the following equation: $V_t = V_{t-1} + \alpha[R_t - V_{t-1}]$. Plotted in black, $\alpha = 0.5$; plotted in gray, $\alpha = 0.7$. A unit value reward has been simulated during trial t , no rewards for the next ten trials.

function of the type described by Equation 1 to determine how it would describe the weighting of reward history in our experiment. Using an exponent and an offset parameter, we found that the function that captured the most variance in the firing rate data obtained during our task had an exponent equal to 0.7, a physiologically determined estimate of the α parameter under the conditions we studied.

In most models, the reward prediction error function is continuous for a wide range of both negative and positive values. The finding that the firing rates of these dopamine neurons after a reward was expected do not encode a negative reward prediction error suggests that there may be limits to the information encoded by these neurons. Negative reward prediction errors occur when the reward obtained is less than the reward that would be predicted by the exponentially weighted average of previous rewards. This may reflect a constraint imposed on dopamine neurons by their small dynamic range in firing rates, and their low baseline firing rates.

In any case, it is important to note that models of reinforcement learning always encode an equal range of positive and negative reward prediction errors, which dopamine firing rates in the fixed duration postreward interval do not encode. This is a point of significant physiological relevance. Most physiological models of reinforcement learning suggest that when the reward prediction error is positive active neurons in the basal ganglia undergo long-term potentiation, strengthening the synaptic connections that produced the reinforced behavior. When the reward prediction error is negative, in contrast, it is typically hypothesized that active connections are decremented in strength, presumably by a process of long-term depression. Given that dopamine spike rates in the postreward interval seem only to encode positive reward prediction errors and that dopamine is known to produce long-term potentiation under some conditions, there may be reason to suspect that dopamine activity is uniquely associated with the positive reward prediction error term of reinforcement learning models.

In support of this possibility, it has been previously proposed that serotonin neurons of the brainstem may provide signals about negative outcomes (Daw et al., 2002), serving as an opponent process to these dopamine neurons. Although little is known about the signals encoded by individual serotonin neurons, there is some evidence suggesting that this neurotransmitter plays an important role in the control of behavior by aversive events, punishments, and losses (Deakin, 1983; Rogers et al., 2003). For example, lesions of the serotonin system increase the probability that animals will switch between two different reinforcement schedules, suggesting a decrease in the effectiveness of the penalty associated with switching (Al-Ruwaitea et al., 1999). It has also been shown that animals with lesions of the serotonin system have difficulty acquiring stimulus associations that require the inhibition of a response, and difficulty inhibiting the response if the lesion is produced following training (Harrison et al., 1999; Soubrie, 1986). Further examination of the responses of serotonin neurons during learning may yield insight into the potential synergy between these two neurotransmitter systems and may test the hypothesis that serotonin represents information about highly negative outcomes.

This finding may also have implications for the asymmetry in sensitivity to losses and gains that has been observed in both human and animal subjects (Kahneman and Tversky, 1979; Marsh and Kacelnik, 2002). It has long been known by psychologists and economists that people are generally more risk averse for losses than for gains (Kahneman and Tversky, 1979). If information about losses is encoded separately from information about gains, it raises the possibility that these two different systems could be calibrated separately, a potential physiological mechanism for this observed “irrationality” in human and animal decision making.

It is, however, important to note that two important factors may complicate this interpretation of our findings. First, our sample of negative reward prediction errors is mostly made up of trials on which the animal received no reward. While current models of reinforcement learning do not predict the existence of any peculiar discontinuity or nonlinearity in the reward prediction error when reward is precisely equal to zero, this may be important. Our observation that essentially all negative reward prediction errors are associated with silence by the dopamine neurons derives from a data set that oversamples this particular condition. Second, it is important to note that the firing rates of dopamine neurons during the fixed duration postreward interval may not be the only mechanism by which dopamine neurons can signal the reward prediction error. Other properties, like the interspike interval, may still allow target structures to derive the reward prediction error from dopamine neuron activity.

Role of Dopamine in the Saccade Timing Task

The behavioral data that we gathered for this study, unfortunately, suggests that, for fully trained animals performing the saccade timing task, neither dopamine activity in particular nor reinforcement learning algorithms in general play a particularly important role in helping the subjects to decide when to move. Information

about the time at which the preceding or subsequent saccade was executed did not appear to be included in the linear computation represented by the dopamine spike rate, and this information did not seem to significantly influence behavior. Instead, the animals appeared to adopt a fairly stereotyped strategy of searching toward the middle of the 4 s trial interval when no reward was received, a situation that typically occurs immediately after the unsignaled block switches. We believe that this draws attention to an important point. It highlights the preexisting conclusion that reinforcement learning mechanisms are not the only processes that guide behavior. This is a point that [Sutton and Barto \(1981\)](#) made when they developed their actor-critic model. Although this model is now a standard element in many reinforcement learning studies, the full model, which does include a reinforcement learning module, also relies on a second set of mechanisms for producing actions. The other mechanisms, broadly speaking, encode rules and policies for producing behavior given the current state of the reinforcement learning system. In our study, it appears from both the behavioral and physiological data that a system of this type governs behavior when the animals make their saccades in our task.

Pretrial Activity in Dopamine Neurons

Our data also indicate that there is a consistent increase in the firing rates of midbrain dopamine neurons immediately following the auditory stimulus that alerted the subjects to the onset of each trial. We found, however, that this early increase was not quantitatively related to the recent reward history of the monkeys. In contrast, other physiological and theoretical studies have shown a different result. When a tone predicts a reward at a fixed latency, dopamine neurons and some classes of reward prediction errors are active after the tone in a way correlated with the overall likelihood that the stimulus will be followed by a reward ([Fiorillo et al., 2003](#)). The insensitivity to recent reward history that we observed after the tone in our experiments may reflect, amongst other things, the fact that in our task the temporal intervals between the tone initiating one trial and the tone initiating the next were very similar. When one stimulus accurately predicts the occurrence in time of a later stimulus that also predicts a reward, it has been shown that dopamine activity after learning is complete is primarily associated with the earlier reward ([Ljungberg et al., 1992](#); [Montague and Berns, 2002](#)). Although there is no way to be certain from these data, it seems possible that our failure to observe a relationship between the number of action potentials produced at the onset of a trial and the reward history of the animal may reflect the highly stereotyped intertrial timing in our task. It is also interesting to note that previous studies demonstrating a relationship of dopamine responses to conditioned stimuli during operant tasks have shown that greater neuronal responses to a conditioned stimulus are associated with reduced reaction times (e.g., [Satoh et al., 2003](#)). However, in the context of our experiment, a strong correlation of this type would prevent the animals from successfully performing the task.

Insight from Parkinson's Disease

Patients suffering from Parkinson's disease exhibit a broad constellation of symptoms that are largely due to the death of dopamine neurons in the SNc. In addition to the readily apparent motor deficits, it has been known for some time that Parkinson's disease patients also have difficulty performing tasks that require learning based on probabilistic outcomes ([Packard and Knowlton, 2002](#)). The finding that the dopamine system encodes only outcomes that are better than expected suggests that Parkinson's disease patients may be unable to perform such activities because there is an imbalance between feedback about negative and positive outcomes, a hypothesis that has recently begun to be tested ([Frank et al., 2004](#)). Since the dopamine system is damaged, while other brain systems are intact (particularly during early stages of the disease), the representations of negative outcomes are likely to be much more reliable than those of positive outcomes. As a result, it may be that patients have difficulty learning associations based on reinforcements resulting from correct responses but may still be quite sensitive to the results of incorrect responses. In line with this hypothesis, Parkinson's disease patients generally show a decreased ability to learn probabilistic associations between stimuli and outcomes ([Knowlton et al., 1996](#)) but exhibit a relatively normal ability to perform reversals of already learned probabilistic associations ([Cools et al., 2001](#)). Taken together with the physiological results that we have presented, the selective pattern of behavioral impairments observed in Parkinson's disease patients may further suggest the existence of multiple systems for encoding positive and negative errors in reward prediction.

Conclusions

We have shown that the firing rates of dopamine neurons following the delivery of a reward encode a computation reflecting the difference between the current reward and a recency-weighted average of previous rewards. This finding corresponds to the predictions of many current theories of conditioning. However, unlike these models, we have also shown that there may be limits to the range in which this signal veridically represents the reward contingencies of the environment. Taken together, these findings support and extend the hypothesis ([Schultz et al., 1997](#)) that midbrain dopamine neurons encode a reward prediction error signal of the type required by most models of reinforcement learning.

Experimental Procedures

Two male rhesus macaques (*Macaca mulatta*) were used as subjects. All animal procedures were developed in association with the University Veterinarian, approved by the New York University Institutional Animal Care and Use Committee, and designed and conducted in compliance with the Public Health Service's *Guide for the Care and Use of Laboratory Animals*. All surgical and training procedures were performed using standard protocols that have been described in detail previously ([Handel and Glimcher, 1997](#)).

Task

Monkeys were trained to perform a saccade timing task in which they had to learn, by trial and error, when to initiate a saccade to

an eccentric target without an external go cue. Saccade timing trials (Figure 1) began with an audible beep. Three hundred milliseconds later, a central light-emitting diode (LED), which appeared yellow to normal human observers, was illuminated, and the subject was required to align gaze with this stimulus ($\pm 3^\circ$) within 1000 ms. Three hundred milliseconds after gaze was aligned with this central LED, it turned red, and a single red eccentric LED was illuminated at 10° of vertical elevation (the location of the target was identical during all experiments). Simultaneously, a timer was started to ensure that all trials lasted exactly 4 s from the time that the eccentric LED was illuminated. During that 4 s interval, the subject could initiate a saccade to the eccentric target at any time. After gaze was shifted into alignment with the eccentric LED, the subject was required to maintain gaze for another 250 ms. Both LEDs would then be extinguished, and a reward would be delivered if the saccade had been executed at the appropriate time. However, a new trial would not begin until the timer indicated that the 4 s interval (which therefore included the intertrial interval) was over.

During each trial, the subject received a reward if he executed the saccade during an unsignaled temporal window that was embedded in the 4 s interval. From the beginning to the end of the window, the volume of liquid reward that the animal could earn increased in five sequential time steps, from 0.2 ml to 0.28 ml. In order to ensure that the subjects' performance was not limited by intervals shorter than their ability to discriminate temporal differences, based on extensive previous studies of interval timing (Gallistel and Gibbon, 2000) the length of the window was scaled logarithmically as it was moved later in the trial. There were thus 11 possible locations for the rewarded window, which ranged in duration from 125 to 2800 ms. The temporal position of the interval was shifted between blocks of trials in an uncued manner, thus requiring the animal to learn which range of saccadic latencies would be rewarded during each block. In order to discourage the animals from anticipating block changes, there was a fixed 5% chance that the block would end on every trial after the 90th, although any block that reached a length of 125 trials was automatically terminated.

Data Analysis

For each saccade timing trial during which the animal made a saccade that aligned gaze with the eccentric target, we recorded how long the animal had waited to make the saccade, the interval during which the saccade would be rewarded on that trial, the volume of liquid reward that the animal received, and the firing rate of the neuron under study during three different intervals: a 50 ms *baseline interval* ending at the onset of the beep that initiated the trial; a 100 ms *beep interval* starting 50 ms after the beep that initiated the trial; and a 150 ms *reward interval* beginning 75 ms after the offset of the eccentric target (which marked the time at which the reinforcement was delivered on rewarded trials).

In order to determine how previous rewards contributed to the firing rate of the neuron on each trial, we created a reward history for each trial, which consisted of the value of the reward (in ml of the water) that the animal received on that trial, as well as the amount of water he had received for each of the previous 20 trials. This allowed us to perform a multiple linear regression on firing rate and reward history. We used least-squares minimization to determine what linear combination of previous rewards best predicted neuronal firing rate during the reward interval. The regression gave us both a set of weights describing the contribution of each reward to the firing rate of the neuron, and how much of the variance in the firing rate was accounted for by reward history. In further analyses, the interval during which the saccade occurred was also added to the multiple regression analysis to determine whether this time-related variable accounted for a significant amount of firing rate variance. We also performed a regression on firing rate and reward history where the reward history was based only on the trials during which the saccade occurred during the same interval.

For some neurons, perievent time histograms were also generated to examine the temporal relationships between modulations in neuronal activity, variance in neuronal response rate, and significant task events. For each perievent time histogram, we averaged the firing rate of the neuron in 20 ms bins and plotted these aver-

ages as a function of time. For each neuron, two histograms were generated: one centered on the time when the auditory tone initiated the trial, and one centered on the time at which reinforcement was delivered (or would have been delivered, in the case of trials in which the saccade was outside of the rewarded interval).

Acknowledgments

We are grateful to M. Brown for his assistance with the experiments and B. Lau for his comments on the manuscript. We also thank M. Grantner and E. Rykkin for technical support. This work was supported by National Institutes of Health Grant EY-10536, and by an individual NRSA awarded to H.M.B. (MH12790).

Received: August 14, 2004

Revised: April 12, 2005

Accepted: May 17, 2005

Published: July 6, 2005

References

- Al-Ruwaitea, A.S., Chiang, T.J., Ho, M.Y., Bradshaw, C.M., and Szabadi, E. (1999). Effect of central 5-hydroxytryptamine depletion on changeover behaviour in concurrent schedules of reinforcement. *Psychopharmacology (Berl.)* *144*, 264–271.
- Bush, R.R., and Mosteller, F. (1955). *Stochastic Models for Learning* (New York: Wiley).
- Cools, R., Barker, R.A., Sahakian, B.J., and Robbins, T.W. (2001). Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. *Cereb. Cortex* *11*, 1136–1143.
- Daw, N.D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Netw.* *15*, 603–616.
- Deakin, J.F. (1983). Roles of brain serotonergic neurons in escape, avoidance, and other behaviors. *J. Psychopharmacol.* *43*, 563–577.
- Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* *299*, 1898–1902.
- Frank, M.J., Seeberger, L.C., and O'Reilly, R.C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science* *306*, 1940–1943.
- Gallistel, C.R., and Gibbon, J. (2000). Time, rate, and conditioning. *Psychol. Rev.* *107*, 289–344.
- Glimcher, P.W., Ciaramitaro, V.M., Platt, M.L., Bayer, H.M., Brown, M.A., and Handel, A. (2001). Application of neurosonography to experimental physiology. *J. Neurosci. Methods* *108*, 131–144.
- Handel, A., and Glimcher, P.W. (1997). Response properties of saccade-related burst neurons in the central mesencephalic reticular formation. *J. Neurophysiol.* *78*, 2164–2175.
- Harrison, A.A., Everitt, B.J., and Robbins, T.W. (1999). Central serotonin depletion impairs both the acquisition and performance of a symmetrically reinforced go/no-go conditional visual discrimination. *Behav. Brain Res.* *100*, 99–112.
- Hollerman, J.R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* *1*, 304–309.
- Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision-making. *Econometrica* *47*, 263–291.
- Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2003). Reward-predicting activity of dopamine and caudate neurons—a possible mechanism of motivational control of saccadic eye movement. *J. Neurophysiol.* *91*, 1013–1024.
- Knowlton, B.J., Mangels, J.A., and Squire, L.R. (1996). A neostriatal habit learning system in humans. *Science* *273*, 1399–1402.
- Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.* *67*, 145–163.
- Marsh, B., and Kacelnik, A. (2002). Framing effects and risky decisions in starlings. *Proc. Natl. Acad. Sci. USA* *99*, 3352–3355.

- Mirenovic, J., and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *J. Neurophysiol.* 72, 1024–1027.
- Montague, P.R., and Berns, G.S. (2002). Neural economics and the biological substrates of valuation. *Neuron* 36, 265–284.
- Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron* 41, 269–280.
- Packard, M.G., and Knowlton, B.J. (2002). Learning and memory functions of the basal ganglia. *Annu. Rev. Neurosci.* 25, 563–593.
- Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In *Classical Conditioning II: Current Research and Theory*, A.H. Black and W.F. Prokasy, eds. (New York: Appleton-Century-Crofts), pp. 64–99.
- Rogers, R.D., Tunbridge, E.M., Bhagwagar, Z., Drevets, W.C., Sahakian, B.J., and Carter, C.S. (2003). Tryptophan depletion alters the decision-making of healthy volunteers through altered processing of reward cues. *Neuropsychopharmacology* 28, 153–162.
- Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *J. Neurosci.* 23, 9913–9923.
- Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neurosci.* 13, 900–913.
- Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599.
- Soubrie, P. (1986). Reconciling the role of central serotonin neurons in human and animal behavior. *Behav. Brain Sci.* 9, 319–325.
- Sutton, R.S., and Barto, A.G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* 88, 135–170.
- Sutton, R.S., and Barto, A.G. (1998). *Reinforcement Learning* (Cambridge, MA: MIT Press).
- Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412, 43–48.