A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values

Ethan S. Bromberg-Martin, Masayuki Matsumoto, Simon Hong and Okihide Hikosaka

J Neurophysiol 104:1068-1076, 2010. First published 10 June 2010; doi:10.1152/jn.00158.2010

J Neurophysiol 104: 1068–1076, 2010.
First published June 10, 2010; doi:10.1152/jn.00158.2010.

A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values

Ethan S. Bromberg-Martin,[1] Masayuki Matsumoto,[1,2] Simon Hong,[1] and Okihide Hikosaka[1]
[1]*Laboratory of Sensorimotor Research, National Eye Institute, National Institutes of Health, Bethesda, Maryland; and* [2]*Primate Research Institute, Kyoto University, Inuyama, Aichi, Japan*

Submitted 5 February 2010; accepted in final form 9 June 2010

Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J Neurophysiol* 104: 1068–1076, 2010. First published June 10, 2010; doi:10.1152/jn.00158.2010. The reward value of a stimulus can be learned through two distinct mechanisms: reinforcement learning through repeated stimulus-reward pairings and abstract inference based on knowledge of the task at hand. The reinforcement mechanism is often identified with midbrain dopamine neurons. Here we show that a neural pathway controlling the dopamine system does not rely exclusively on either stimulus-reward pairings or abstract inference but instead uses a combination of the two. We trained monkeys to perform a reward-biased saccade task in which the reward values of two saccade targets were related in a systematic manner. Animals used each trial's reward outcome to learn the values of both targets: the target that had been presented and whose reward outcome had been experienced (experienced value) and the target that had not been presented but whose value could be inferred from the reward statistics of the task (inferred value). We then recorded from three populations of reward-coding neurons: substantia nigra dopamine neurons; a major input to dopamine neurons, the lateral habenula; and neurons that project to the lateral habenula, located in the globus pallidus. All three populations encoded both experienced values and inferred values. In some animals, neurons encoded experienced values more strongly than inferred values, and the animals showed behavioral evidence of learning faster from experience than from inference. Our data indicate that the pallidus-habenula-dopamine pathway signals reward values estimated through both experience and inference.

INTRODUCTION

It is thought that the brain contains multiple learning systems that compete to control behavior. An influential distinction is between two learning mechanisms: reinforcement learning by repeated stimulus-reward pairings (Bayley et al. 2005; Knowlton et al. 1996; Wise 2004) and abstract inference using task-specific rules (Daw et al. 2005; Dayan and Niv 2008; Hampton et al. 2006). The reinforcement mechanism is often theorized to be controlled by midbrain dopamine neurons, including those located in the substantia nigra pars compacta and their projection targets in the dorsolateral striatum (Knowlton et al. 1996; Yin and Knowlton 2006). Dopamine neurons are thought to be responsible for reinforcement processes such as forming stimulus-response associations (Wise 2004), habit learning (Knowlton et al. 1996), and model-free reward learning (Daw et al. 2005). In contrast, abstract inference is often theorized to be performed by prefrontal cortical areas responsible for knowledge of task-specific rules, reversal learning, and model-based reward learning (Daw et al. 2005; Hampton et al. 2006; Miller and Cohen 2001; Sakai 2008).

Address for reprint requests and other correspondence: E. S. Bromberg-Martin, Lab. of Sensorimotor Research, National Eye Inst., NIH, 49 Convent Dr., Bldg. 49, Rm. 2A50, Bethesda, MD 20892-4435 (E-mail: bromberge@mail.nih.gov).

However, there is evidence that dopamine neurons are not constrained to treat stimulus-reward pairings in the conventional manner of reinforcement learning. Dopamine neurons have been studied using tasks in which reward delivery on one trial caused a reduction in the reward value of future trials (Nakahara et al. 2004; Satoh et al. 2003). Even though a trial was paired with a rewarding outcome, neurons correctly decreased their estimate of the next trial's value. These results show that dopamine neurons are able to improve their value estimates by learning from stimulus-reward pairings in an unconventional manner. In light of these results, we hypothesized that the dopamine system has even greater flexibility: that it is able to learn stimulus values even in the absence of stimulus-reward pairings through a process of abstract inference. The most direct test of this hypothesis would be whether dopamine neurons can infer that the value of a stimulus has changed even in a situation where the stimulus has not been physically presented and has not been paired with a reward outcome. In addition, it would be ideal to compare neural coding of values that have been inferred with values that have been directly experienced, to see which form of reward learning these neurons preferentially represent.

We tested this hypothesis by training monkeys to perform a task with a "reversal set," in which the reward values of two stimuli are anticorrelated (Hampton et al. 2006; Meyer 1951; Watanabe and Hikosaka 2005). In this task, animals learn the reward statistics of the task environment, so that when the value of one stimulus is changed, the animal infers that the value of the second stimulus has changed in the opposite direction (Watanabe and Hikosaka 2005). This is a form of inference in the sense that animals can tell the second stimulus has changed its value based on their prior knowledge of the task environment, without requiring new exposure to that stimulus or its outcome. We found that, in parallel with the animal's behavior, neurons signaled stimulus values that had been experienced as well as stimulus values that had been inferred. This coding was found in neurons at three locations in a neural pathway controlling the dopamine system: *1*) substantia nigra dopamine neurons themselves; *2*) neurons of the lateral habenula, a major source of input to dopamine neurons that is thought to exert inhibitory control over the dopamine system (Matsumoto and Hikosaka 2007); and *3*) neurons in the globus pallidus internal segment that transmit negative reward signals to the lateral habenula (GPi^{LHb}-negative neurons) (Hong and Hikosaka 2008).

METHODS

Subjects and surgery

Four rhesus monkeys (*Macaca mulatta*), *monkeys E, L, D*, and *N*, were the subjects in this study. All animal care and experimental

procedures were approved by the Institute Animal Care and Use Committee and complied with the Public Health Service Policy on the humane care and use of laboratory animals. A head-holding device, a chamber for unit recording, and a scleral search coil were implanted under general anesthesia. During experimental sessions, monkeys were seated in a primate chair in a sound-attenuated and electrically shielded room.

Behavioral task

Behavioral tasks were under the control of a QNX-based real-time experimentation data acquisition system (REX, Laboratory of Sensorimotor Research, National Eye Institute, National Institutes of Health, Bethesda, MD). The monkeys sat facing a frontoparallel screen with an eye-to-screen distance of ~30 cm. Stimuli were generated by an active matrix liquid crystal display projector (PJ550, ViewSonic) and rear-projected on the screen. The monkeys were trained to perform a one-direction-rewarded version of the visually guided saccade task (Fig. 1*A*). A trial started when a small fixation spot appeared on the screen, typically 0.6° in diameter. After the monkey maintained fixation on the spot for 1,200 (*monkeys E* and *L*) or 1,000 ms (*monkeys D* and *N*), the fixation spot disappeared and a peripheral target appeared on the left or right, typically 15 or 20° from the fixation spot and 1.2° in diameter. The monkey was required to make a saccade to the target within 500 ms. Errors were signaled by a beep sound followed by a repeat of the same trial. Correct saccades were signaled by a 100 ms tone starting 200 ms after the saccade. In rewarded trials, a liquid reward was delivered that started simultaneously with the tone stimulus. The intertrial interval was fixed at 2.2 s or randomized from 2.2 to 3.2 s (*monkeys E* and *L*) or was randomized from 2.5 to 3.5 s (*monkeys D* and *N*). In each block of 24 trials, saccades to one fixed direction were rewarded with 0.3 ml of apple juice, whereas saccades to the other direction were not rewarded. The position-reward contingency was reversed in the next block. There was no external instruction indicating that the block had changed. Both outcomes (rewarded and unrewarded) and target locations (left and right) occurred with equal frequency. For *monkeys E* and *L*, each sub-block of four trials contained two left target and two right target trials in a random order. For *monkeys D* and *N*, each trial's target location was chosen with a computerized coin toss. The rate of correct behavioral performance was high at all times during the task (trial 1, 96 ± 0.5%; trial 2, 93 ± 0.7%; last 10 trials of the block, 95 ± 0.2%).

Single-neuron recording

We used conventional electrophysiological techniques described previously (Hong and Hikosaka 2008; Matsumoto and Hikosaka 2007). Eye movements were monitored using a scleral search coil system with 1-ms resolution. Recording chambers were placed over the midline of the parietal cortex, tilted posteriorly by 38°, and aimed at the habenula; or placed over the frontoparietal cortex, tilted laterally by 35°, and aimed at the globus pallidus internal segment or substantia nigra. The locations of globus pallidus, lateral habenula, and dopamine neurons were mapped based on MRIs (4.7 T, Bruker) and the distinctive activity patterns in nearby brain structures, and the locations were confirmed by histology (Hong and Hikosaka 2008; Matsumoto and Hikosaka 2007). Single-unit recordings were performed using tungsten electrodes (Frederick Haer) that were inserted through a stainless steel guide tube and advanced by an oil-driven micromanipulator (MO-97A, Narishige) or an electrically driven micromanipulator (MicroStepper, LSR/NEI/National Institutes of Health). Single neurons were isolated on-line using custom voltage-time window discrimination software (MEX, LSR/NEI/National Institutes of

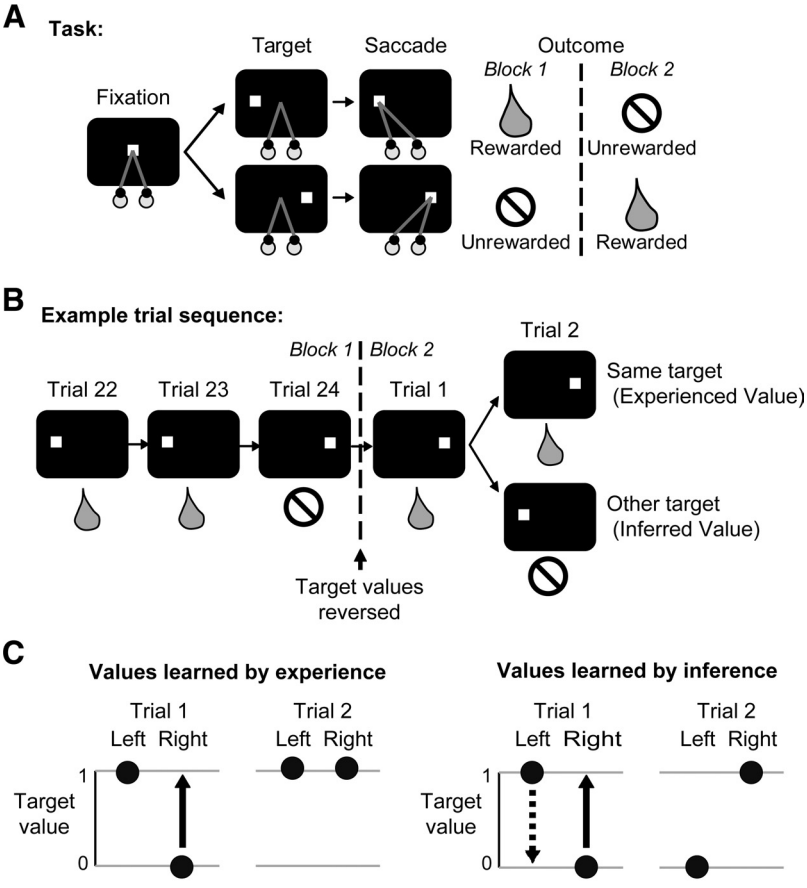

FIG. 1. Reward-biased saccade task. *A*: task diagram. The monkey fixated a central spot for 1.2 s. The spot disappeared and simultaneously a visual target appeared on the left or right side of the screen. The monkey was required to saccade to the target. In 1 block of 24 trials, left saccades were rewarded and right saccades were unrewarded (block 1); in the next block, the reward values were reversed without notice to the animal (block 2). *B*: example sequence of events after a block change. In the 1st trial of the new block, the monkey receives an unexpected reward outcome (trial 1: right target, reward). The 2nd trial of the block could present the same target, whose new reward value had just been experienced (trial 2: same target, experienced value), or it could present the other target, which had been absent on the previous trial and whose new reward value had to be inferred based on the reversal rule of the task (trial 2: other target, inferred value). *C*: 2 ways to learn stimulus values from the pairing right target → reward. *Left*: if the animal learned through experience alone, the right target value would be increased but the left target value would remain unchanged. In trial 2, the animal would show no preference between the targets. *Right*: if the animal learned through inference, the animal would additionally infer that the block had changed to block 2, and hence the left target value had decreased. The animal's preference would switch from the left target to the right target.

Health). We searched for dopamine neurons in and around the substantia nigra pars compacta. Dopamine neurons were identified by their irregular and tonic firing around 5 spikes/s and broad spike potentials. In this experiment, we focused on dopamine neurons that responded to reward-predicting stimuli with phasic excitation. Dopamine-like neurons that were not sensitive to reward-predicting stimuli were not examined further. Globus pallidus internal segment neurons projecting to the lateral habenula were identified with antidromic stimulation techniques (Hong and Hikosaka 2008).

Database

We analyzed behavior and neural activity that had been recorded in two previous studies (Hong and Hikosaka 2008; Matsumoto and Hikosaka 2007). Animals had >30,000 trials of prior experience at the task before these neurons were recorded. In *monkeys E* and *L*, our database consisted of 65 lateral habenula neurons (28 in *monkey E*, 37 in *monkey L*) and 64 reward-positive putative dopamine neurons (20 in *monkey E*, 44 in *monkey L*). In *monkeys D* and *N*, our database consisted of 35 sessions of lateral habenula multiunit activity (18 in *monkey D* and 17 in *monkey N*) and 74 habenula-projecting globus pallidus internal segment neurons (42 in *monkey D* and 32 in *monkey N*). In this study, we analyzed only globus pallidus neurons that had negative reward signals, which were identified using the following procedure (Hong and Hikosaka 2008). We defined each neuron's normalized target response on each trial as its firing rate during a window 150–350 ms after target onset, minus its mean firing rate during a 1,000 ms window before fixation point onset averaged over all trials with the same target direction and reward outcome. We classified a cell a as negative-reward neuron if its normalized target responses had a significant negative main effect of reward in a two-way ANOVA with the factors reward (unrewarded or rewarded) × target position (left or right; $P < 0.01$). This yielded 37 globus pallidus neurons (19 in *monkey D* and 18 in *monkey N*). There was also a population of globus pallidus neurons that carried positive reward signals (Hong and Hikosaka 2008), but the number of such neurons was too small for our analysis.

Data analysis

All statistical tests were two-tailed. The target response was measured using the firing rate in a window 150–350 ms after target onset. The outcome response was measured using the firing rate in a window 200–600 ms after outcome onset. In this analysis, we focused on the first and second trials of each block, because by the third trial of the block the reversals in neural activity and behavior were essentially complete (Hong and Hikosaka 2008;Matsumoto and Hikosaka 2007).

To measure the neural and behavioral change in estimated values of the targets, we defined a reversal index denoted with symbol RI, for each neural population and each set of behavioral sessions. We will first describe the general form of the reversal index and then describe its detailed calculation. Conceptually, the reversal index specifies the degree to which the neural discrimination D between the targets on the current trial (D_{CUR}) has changed from its previous level before the target value reversal (D_{BEF}) to its asymptotic level after target value reversal (D_{AFT}). The reversal index is calculated using the equation

$$RI = (D_{CUR} - D_{BEF})/(D_{AFT} - D_{BEF})$$

Thus if neural discrimination on the current trial is the same as its level before reversal, RI = 0, whereas if neural discrimination is the same as its asymptotic level after reversal, RI = 1. To calculate this for our data, we defined the neural discrimination as the difference in firing rate between the two targets: D = (firing rate for rewarded target) – (firing rate for unrewarded target). In our analysis, D_{CUR} was defined as the mean firing rate difference measured on the second trial of the block (trial 2 in Figs. 2 and 4). D_{AFT} was defined as the mean firing rate difference measured on the last 10 trials of the block. D_{BEF} was defined as the mean firing rate difference measured on the last 10 trials of the block with the firing rates switched for the two targets (such that $D_{BEF} = -D_{AFT}$), thus mimicking the condition that occurred before learning, on the first trial of the block immediately after the two target values had been switched.

The specific calculation procedure was as follows. To calculate the reversal index for experienced value trials, RI_{Exp}, we first calculated the terms D_{CUR}, D_{BEF}, and D_{AFT} for each neuron, using only trials when the target was presented at the same location compared with the previous trial ("same" target in Figs. 2 and 4). Thus for each neuron i, this produced the terms $D_{CUR}(i, \text{same})$, $D_{BEF}(i, \text{same})$, $D_{AFT}(i, \text{same})$. We then calculated RI_{Exp} using the equation

$$RI_{Exp} = \sum_i [D_{CUR}(i, \text{same}) - D_{BEF}(i, \text{same})]/\sum_i [D_{AFT}(i, \text{same}) - D_{BEF}(i, \text{same})]$$

The reversal index for inferred value trials, RI_{Inf}, was calculated in the same way but using only trials when the target was presented at a different location on the screen compared with the previous trial ("other" target in Figs. 2 and 4)

$$RI_{Inf} = \sum_i [D_{CUR}(i, \text{other}) - D_{BEF}(i, \text{other})]/\sum_i [D_{AFT}(i, \text{other}) - D_{BEF}(i, \text{other})]$$

The reversal indexes for behavioral reaction times were calculated in the same way as for neural activity, except for treating each behavioral session as a separate neuron and calculating the difference in reaction times instead of the difference in firing rates. For each animal, the behavioral data were pooled over all recording sessions. Note that each reversal index was calculated using only neurons or behavioral sessions i that had enough data to calculate the relevant measure of neural discrimination $D_{CUR}(i, \text{same})$ or $D_{CUR}(i, \text{other})$, meaning that it had to have at least one rewarded trial and one unrewarded trial. For each reversal index, the mean number of trials available from each neuron was 1.7 for rewarded trials (range, 0–6) and 1.5 for unrewarded trials (range, 0–6). Note also that, because of the relatively small number of trials that were available for each individual neuron, reversal indexes could not be calculated reliably at the single neuron level. For this reason, the reversal index was calculated at the population level, based on the population average activity (see equations above). The SE of the reversal index was estimated using a bootstrap procedure (Efron and Tibshirani 1993). To calculate a bootstrap reversal index, we randomly sampled the neurons with replacement to create a bootstrap dataset and calculated the reversal index for that dataset. The SE was defined as the SD of a set of 20,000 such bootstrap reversal indexes.

This analysis was designed to isolate the effects of task-specific inference while controlling for any other potential difference between experienced value and inferred value trials. In particular, on experienced value trials the target location and reward outcome were the same as the previous trial (repeating), whereas on inferred value trials, the target location and reward outcome were different from the previous trial (switching). We designed the reversal index to control for any idiosyncratic effects of staying versus switching on neural activity or behavior by calculating RI_{Exp} using only neural activity from trials that shared the identical repeating condition (same target) and calculated RI_{Inf} using only neural activity from trials that shared the identical switching condition (other target). This also ensured that the two indexes provided independent measurements, because they were calculated using separate subsets of the data. In addition, note that animals could not be certain in advance whether a trial would be experienced value or inferred value because the target location was chosen pseudorandomly on each trial. Thus the comparison between RI_{Exp} and RI_{Inf} controls for any effect of the animal's preparatory state, such as levels of arousal, motivation, attention, or motor readiness.

Statistical significance and P values were calculated using shuffling procedures. First, we tested the hypothesis that the RI was equal to 0

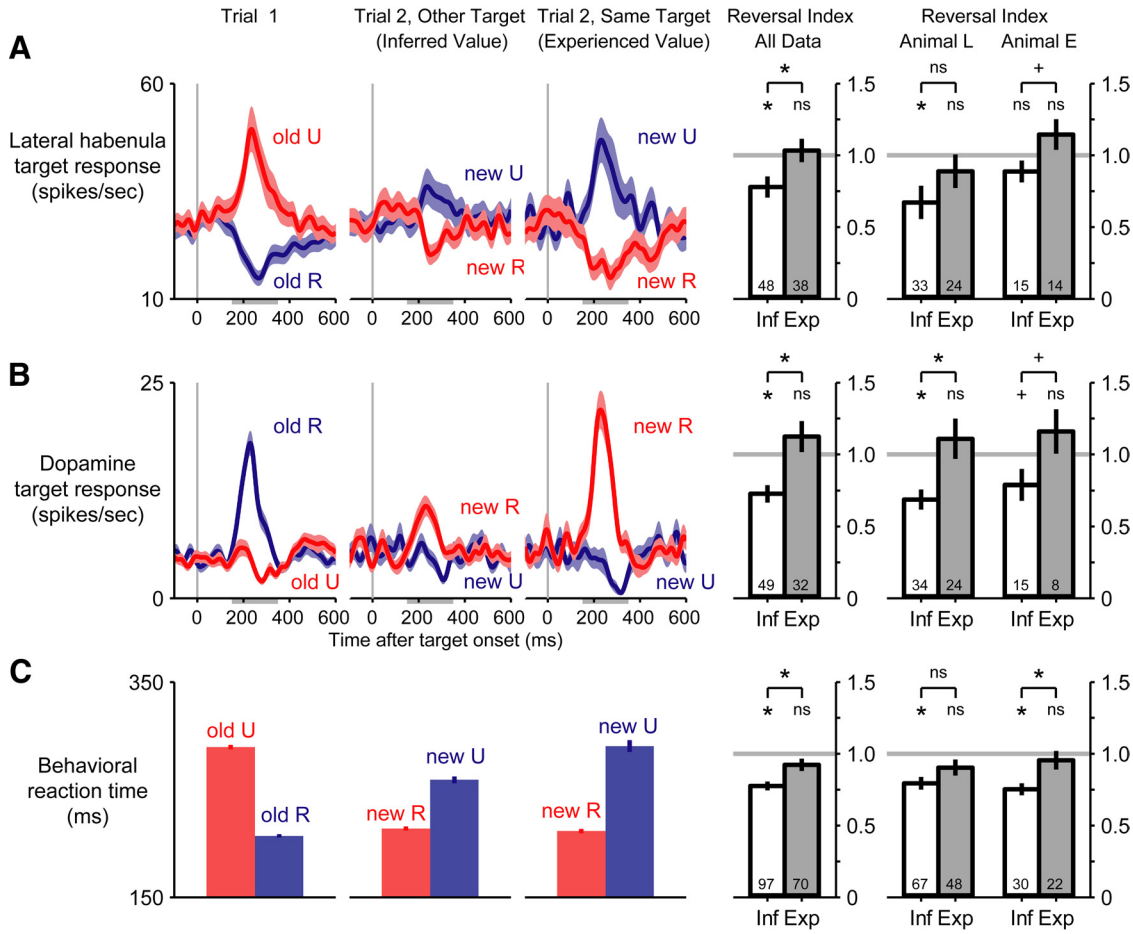

FIG. 2. Combination of experienced and inferred stimulus values in neural activity and behavior in *monkeys E* and *L*. The rows represent (*A*) lateral habenula neurons, (*B*) dopamine neurons, and (*C*) behavioral reaction times. *First 3 columns*: data for the 1st trial of the block (trial 1), for the 2nd trial of the block when the target was different from the 1st trial (trial 2, Other Target), and for the 2nd trial of the block when the target was the same as on the 1st trial (trial 2, Same Target). Data are shown separately for the target that was rewarded in the previous block and unrewarded in the current block (old R, new U, blue) and for the target that was unrewarded in the previous block and rewarded in the current block (old U, new R, red). Neural firing rates were smoothed with a Gaussian kernel (σ = 15 ms) and averaged over neurons. Shaded areas and error bars are ±SE. Gray bars along the time axis indicate the response window for calculation of reversal indexes. Note that each red or blue curve in *A* and *B* only includes data from neurons that had at least 1 trial in which the appropriate current-trial and past-trial targets were presented (n = 42–63 for each curve). *Right 3 column*: reversal index on the 2nd trial of the block, calculated using all data (1st column), using data from *monkey L* (2nd column), and using data from *monkey E* (3rd column). Reversal indexes were calculated separately for other-target trials when the value of the target had to be inferred (white bars, Inf) and for same-target trials when the value of the target had already been experienced on the 1st trial of the block (gray bars, Exp). Numbers at the bottom of each bar indicate the number of neural recording sessions for that bar. Symbols indicate statistical significance measured using a shuffling procedure (*$P < 0.05$; $^{+}P \leq 0.06$; ns $P > 0.06$). Error bars are ±SE. Neural and behavioral measures of stimulus values reversed on both trial types but reversed less fully on inferred value trials.

(no reversal), which occurs when $D_{CUR} = D_{BEF}$. To generate a distribution of RIs representing this hypothesis, we generated 20,000 shuffled datasets in which $D_{CUR}(i)$ and $D_{BEF}(i)$ were randomly shuffled within each neuron. For each shuffled dataset, we calculated the reversal index. We then computed the two-tailed P value by comparing the distribution of reversal indexes from the shuffled datasets to the measured reversal index from the original data. Second, we tested the hypothesis that the RI was equal to 1 (full reversal), which occurs when $D_{CUR} = D_{AFT}$. We tested this hypothesis using the same procedure as before, except by shuffling $D_{CUR}(i)$ and $D_{AFT}(i)$.

Finally, we tested the hypothesis that the reversal indexes RI_{Exp} and RI_{Inf} were equal to each other, i.e., that the difference ($RI_{Exp} - RI_{Inf}$) was equal to 0. To generate a null distribution representing the hypothesis of no difference, we used the following procedure. We first considered the M neurons that contributed data from "same target" trials for calculating RI_{Exp} and collected their neural discrimination values that were used to calculate RI_{Exp}, producing for each of these neurons i a three-element vector [$D_{CUR}(i$, same), $D_{BEF}(i$, same), and $D_{AFT}(i$, same)]. We then considered the N neurons that contributed data from "other target" trials for calculating RI_{Inf}, producing for each of these neurons j the analogous three-element vector [$D_{CUR}(j$, other), $D_{BEF}(j$, other), $D_{AFT}(j$, other)]. Of the total $M + N$ vectors, we randomly reassigned M to the same (experienced value) condition, reassigned the remaining N to the other (inferred value) condition and recalculated the two RIs and their difference, ($RI_{Exp} - RI_{Inf}$). We repeated this shuffling procedure to produce a distribution of 20,000 differences and computed the two-tailed P value by comparing this distribution of differences to the measured difference from the original data.

RESULTS

Behavioral combination of experienced and inferred stimulus values

We trained four monkeys to perform a reward-biased saccade task (Fig. 1*A*). In this task, the monkey began each trial by holding its gaze on a fixation point for 1.2 s. Then the fixation

point disappeared and the monkey made an eye movement to a visual target that appeared on the left or right side of the screen. The location of the saccade target indicated the trial's upcoming reward outcome. In each block of 24 trials, one target location was rewarded, whereas the other target location was unrewarded. Even on unrewarded trials, monkeys still had to make the saccade correctly or else the trial was repeated. As shown in previous studies, monkeys closely tracked the values of the two targets, saccading with short reaction times to the rewarded target and long reaction times to the unrewarded target (Figs. 2*C* and 3*C*). At the end of each block, the reward values of the two target locations were reversed without warning to the animal.

In this study, we analyzed behavior and neural activity on the first two trials after the reversal, when animals had to learn that the reward values of the targets had been changed (Fig. 1*B*). For example, suppose that in the previous block the left target had been rewarded and the right target had been unrewarded. Then on the first trial of the new block, the right target appears and is rewarded, an unexpected outcome. How should the monkey adjust its estimate of the reward values of the two targets? Clearly, the monkey can use the right target's most recently experienced outcome, a reward, to estimate that the right target now has a high value (Fig. 1*B*, same target, experienced value). However, what about the other, left-side target? If the monkey learned only from stimulus-reward pairings, the monkey would not update the value of the left target because the left target had not been physically present and had not been paired with its new reward outcome. The monkey would still believe that the left target had a high value and would have no preference between the targets (Fig. 1*C*, stimulus values learned by experience). On the other hand, if the monkey learned using a strategy tuned to the reward statistics of the task environment, the monkey would correctly infer that an increase in the value of the right target implied a decrease in the value of the left target. The monkey would switch its preference from the left target to the right target (Fig. 1*C*, stimulus values learned by inference).

Consistent with a previous study, we found that monkeys used outcomes gained from one target to infer the value of the other target (Watanabe and Hikosaka 2005). In the following section, we will focus on *monkeys E* and *L,* which were used for recording lateral habenula and dopamine neurons (Fig. 2). On the first trial of each block, the monkeys did not yet know that the target values had reversed. Their reaction times were fast for the old rewarded target and slow for the old unrewarded target (Fig. 2*C*, trial 1). By the second trial of the block, however, the monkeys changed their reaction times to match the new reward values of both targets. This happened when the second-trial target was the same as the first-trial target and its new reward value had been experienced (Fig. 2*C*, same target); it also happened when the second-trial target was different from the first-trial target and its new reward value had to be inferred based on the reward statistics of the task (Fig. 2*C*, other target). The reaction time bias favoring the rewarded target was somewhat weaker on inferred-value trials than experienced-value trials, suggesting that animals learned more fully through direct experience than through inference alone (Fig. 2*C*).

Inferred value signals in lateral habenula and dopamine neurons

We next asked whether inferred stimulus values could be accessed by neurons that control the dopamine system. We analyzed the activity of 65 neurons in the lateral habenula and 64 reward-responsive putative dopamine neurons in the substantia nigra pars compacta, recorded in *monkeys E* and *L* (Matsumoto and Hikosaka 2007 and METHODS). As shown in previous studies, these neurons responded to the targets with strong reward-predictive signals (Matsumoto and Hikosaka 2007). Lateral habenula neurons were inhibited by the rewarded target and strongly excited by the unrewarded target

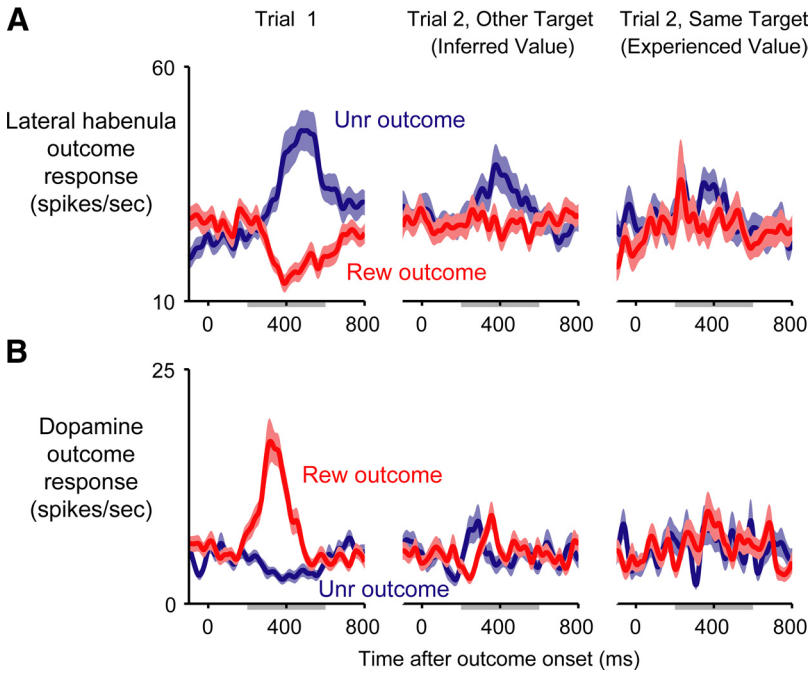

FIG. 3. Neural responses to outcome delivery in *monkeys E* and *L*. The rows represent (*A*) lateral habenula neurons and (*B*) dopamine neurons. Same format as the *left 3 columns* of Fig. 2. Data are plotted from the same neurons and trials as in Fig. 2, *A* and *B*, but aligned on outcome delivery. Gray bars along the time axis indicate the time window for measuring the outcome response. On the 1st trial of each block when an unexpected outcome was delivered, lateral habenula and dopamine neurons had a strong outcome response (*left column*). On inferred-value trials, lateral habenula neurons had a tendency for a small residual outcome response (*middle column*).

(Fig. 2*A*, trial 1). Dopamine neurons had the opposite response pattern, excited by the rewarded target and inhibited by the unrewarded target (Fig. 2*B*). Crucially, on the second trial of each block, neurons changed their reward-predictive activity to match the new reward values of both targets. This could be seen when the value of the second-trial target had already been experienced (Fig. 2, *A* and *B*, same target); it could also be seen when the second-trial target had been absent on the previous trial and its value had to be inferred (Fig. 2, *A* and *B*, other target). Thus lateral habenula and dopamine neurons signaled both experienced and inferred stimulus values. As in the monkey's behavior, the difference in neural response strength between the two targets was somewhat weaker on inferred-value trials, suggesting that the neurons accessed the same estimate of target value that was controlling the monkey's behavior. These data indicate that monkeys and neurons did not estimate stimulus values exclusively based on either stimulus-reward pairings or task-specific inference rules; instead, they used a combination of the two.

A more detailed statistical analysis supported these conclusions. Neurons distinguished between the rewarded and unrewarded targets on both inferred-value trials and experienced-value trials (Wilcoxon signed-rank test, inferred-value: habenula, $P = 0.004$; dopamine, $P < 10^{-3}$; experienced value: habenula $P < 10^{-4}$; dopamine, $P < 10^{-4}$). We quantified the degree of neural reversal expressed on the second trial of each block using a reversal index denoted RI (see METHODS; Fig. 2, *A* and *B, right*). The reversal index was defined based on the population average neural activity (METHODS). This index was 0 if the population continued to respond as it had in the old block (no reversal) and was 1 if the population completely reversed and reached its full response strength in the new block (full reversal). An analogous reversal index was defined for behavior, based on saccadic reaction times (METHODS; Fig. 2*C, right*). Statistical significance was measured using shuffling procedures (METHODS). All reversal indexes were greater than zero ($P < 10^{-4}$), and the reversal indexes were larger for experienced values than for inferred values ($RI_{Exp} > RI_{Inf}$: habenula, $P = 0.02$; dopamine, $P = 0.0008$; behavior, $P = 0.006$). The reversal index on experienced value trials was not significantly different from 1, consistent with full reversal (habenula, $P = 0.68$; dopamine, $P = 0.25$; behavior, $P = 0.08$). The reversal index on inferred value trials was close to 0.75 and was significantly <1, indicating partial reversal (habenula, $P = 0.001$; dopamine, $P = 10^{-4}$; behavior, $P < 10^{-4}$). Inspection of data from individual animals indicated that the same qualitative pattern of effects was present in both monkeys, although not all effects reached statistical significance in each animal possibly due to the smaller number of neurons (Fig. 2).

In a previous study, we found that dopamine neurons can be classified into multiple types that carry different motivational signals (Matsumoto and Hikosaka 2009). One type of dopamine neurons is inhibited by punishments, as though encoding their negative motivational value, whereas other types of dopamine neurons react to punishments with no response or with excitation. These types of dopamine neurons could potentially carry different inferred value signals. Although we cannot identify these neuron types directly in this study because we did not measure neural responses to punishments, we can take advantage of the fact that dopamine neurons that are inhibited by punishments are also strongly inhibited by reward omission cues, whereas other types of dopamine neurons have weaker or no inhibition (Matsumoto and Hikosaka 2009). We therefore sorted dopamine neurons into two subpopulations based on whether they were strongly inhibited below their baseline firing rate in response to the unrewarded target ($P < 0.001$, signed-rank test using data from trials 3–24 of the block). Both subpopulations of dopamine neurons had experienced and inferred value signals consistent with those seen in the population as a whole, although the results were somewhat noisier because of the smaller number of neurons. This could be seen in the subpopulation that was strongly inhibited by the unrewarded target ($n = 31$, $RI_{Exp} = 0.91$, $RI_{Inf} = 0.73$; both indexes >0, $P < 0.001$), as well as in the subpopulation that was not ($n = 33$, $RI_{Exp} = 1.44$, $P = 0.20$; $RI_{Inf} = 0.79$, $P < 10^{-4}$). Thus it is likely that inferred value signals are prevalent in the general population of dopamine neurons as a whole.

We also examined neural responses to reward outcome delivery. In this task, lateral habenula and dopamine neurons responded to the unpredicted reward outcomes delivered on the first trial of each block (Matsumoto and Hikosaka 2007) (Fig. 3, *left column*). These outcome responses might remain on inferred-value trials because the target values were only partially learned and therefore the outcome might be only partially predicted. There was evidence for such a tendency in lateral habenula neurons (Fig. 3*A*). We measured the outcome response as the difference between firing rates for the rewarded and unrewarded outcomes. Lateral habenula neurons had a significant negative outcome response on the first trial of each block ($P < 10^{-4}$, Wilcoxon signed-rank test) and on the second trial of each block on inferred value trials ($P = 0.001$) but not on experienced value trials ($P = 0.14$). However, the difference between inferred and experienced value responses fell short of statistical significance ($P = 0.06$, Wilcoxon signed-rank test). There was no clear evidence for such a tendency in dopamine neurons (Fig. 3*B*). Dopamine neurons had a significant positive outcome response on the first trial of each block ($P < 10^{-4}$) but not on the second trial of each block (inferred value trials, $P = 0.23$; experienced value trials, $P = 0.72$).

Inferred value signals in globus pallidus neurons that project to the lateral habenula

Given that lateral habenula and dopamine neurons encoded inferred stimulus values, we wondered which upstream site in the reward pathway could be the source of these signals. One candidate is a subpopulation of neurons in the globus pallidus internal segment that project to the lateral habenula (Parent et al. 1981; Hong and Hikosaka 2008). We therefore analyzed data from *monkeys D* and *N* in which we recorded the activity of globus pallidus neurons as well as lateral habenula multiunit activity (Hong and Hikosaka 2008). We focused our analysis of pallidus neurons on a subpopulation, which we will refer to as GPi^{LHb}-negative neurons—those that projected to the lateral habenula (confirmed using antidromic stimulation techniques) and that carried negative reward signals similar to those seen in lateral habenula neurons, responding with a higher firing rate to the unrewarded target than to the rewarded target (METHODS).

Again, we found that behavior and neural activity reflected inferred values (Fig. 4). The reversal index for inferred-value trials was significantly greater than zero in each monkey for

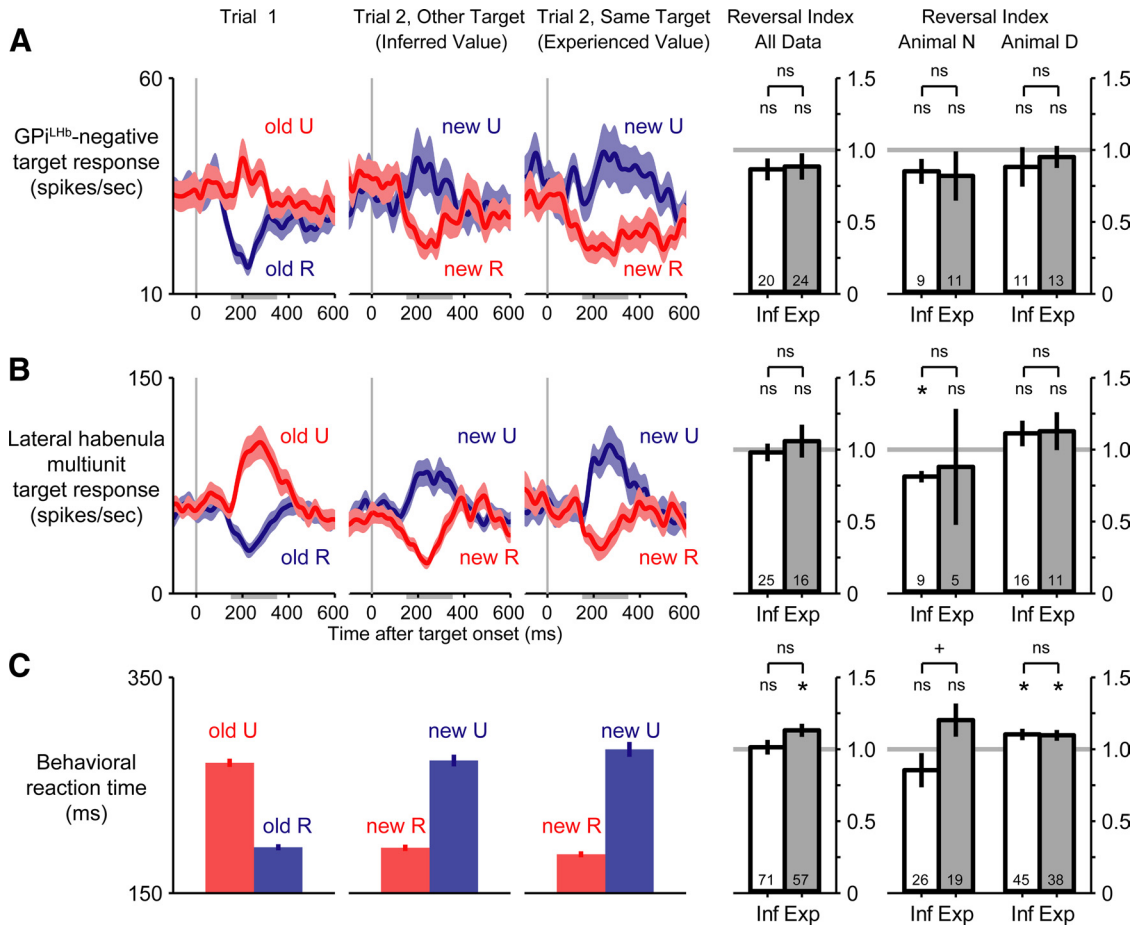

FIG. 4. Experienced and inferred stimulus values in neural activity and behavior in *monkeys N* and *D*. Same format as Fig. 2. The rows represent (*A*) GPi^{LHb}-negative neurons, (*B*) lateral habenula multiunit activity, and (*C*) behavioral reaction times. Note that each red or blue curve in *A* and *B* only includes data from neurons that had at least 1 trial in which the appropriate current-trial and past-trial targets were presented ($n = 24–37$ for each curve). In *monkey D*, neural and behavioral measures of stimulus values reversed similarly on both experienced value and inferred value trials (*right column*).

each variable measured: GPi^{LHb}-negative population activity, lateral habenula population activity, and behavioral reaction times (all $P < 0.01$).

Whereas *monkeys E* and *L* learned better from experience (Fig. 2), the averaged behavioral data from *monkeys D* and *N* reflected similar learning from both experience and inference (Fig. 4*C*; $RI_{Inf} = 1.04$, $RI_{Exp} = 1.13$, $P = 0.10$; although RI_{Exp} was >1, $P < 0.01$). In parallel, the population average neural activity in these animals also reflected similar learning from experience and inference (Fig. 4*C*; habenula, $P = 0.49$; GPi^{LHb}-negative, $P = 0.87$). Inspection of data from single animals indicated that *monkey D* had behavioral reversal indexes that were slightly >1 and were very similar to each other ($RI_{Inf} = 1.10$, $RI_{Exp} = 1.10$, $P = 0.91$), and likewise, the animal's neural reversal indexes were also very similar to each other (habenula, $RI_{Inf} = 1.11$, $RI_{Exp} = 1.13$, $P = 0.92$; GPi^{LHb}-negative, $RI_{Inf} = 0.88$, $RI_{Exp} = 0.95$, $P = 0.66$). *Monkey N* also had no consistently detectable difference in reversal indexes, although this animal did have a modest tendency for lower reversal indexes on inferred value trials (the behavioral RI_{Inf} and RI_{Exp} had a trend to be different, $P = 0.06$, although RI_{Inf} was not significantly <1, $P = 0.24$; for habenula multiunit, RI_{Inf} was significantly <1, $P < 0.01$, although it was not significantly less than RI_{Exp}, $P = 0.62$). Taken together, these data were consistent with the possibility that neural signals reflect an animal's knowledge of the task. In some animals, both behavioral and neural reversal were less than complete (*monkeys E* and *L*; Fig. 2), whereas in at least one other animal, both behavioral and neural reversal were fully complete (*monkey D*; Fig. 4).

DISCUSSION

We found that GPi^{LHb}-negative neurons, lateral habenula neurons, and substantia nigra dopamine neurons were able to infer the new reward value of a stimulus even when the stimulus had not been presented and had not been paired with its new reward outcome. This form of inferential learning had a large influence on neural activity and behavior; when stimulus values were changed, animals were able to accomplish 75–100% of neural and behavioral reversal through inference alone. This shows that a neural pathway controlling the dopamine system does not learn exclusively from stimulus-reward pairings; it can also infer stimulus values based on knowledge of the task at hand.

Our study makes three contributions toward understanding the neural mechanism of the inference process. First, to our knowledge, our study is the first demonstration that inferred value signals are present in the basal ganglia including the dopaminergic reward system. One previous study using a

similar task showed that a subset of neurons in the caudate nucleus can reverse their activity as quickly as the neurons in this study, within a single trial after a block change (Watanabe and Hikosaka 2005). However, these caudate neurons did not encode the value of the presented stimulus; rather, these neurons discriminated between the two blocks of trials, with some neurons activated during block 1 and other neurons activated during block 2. This block-coding activity would be useful in the inference process because knowing the current block of trials would allow animals to deduce the proper stimulus values, although it is currently unknown whether this activity is used for value inference. Other previous studies discovered inferred value signals in prefrontal cortical areas, notably in single neuron activity in the lateral prefrontal cortex (Pan et al. 2008b) and in blood oxygen level–dependent signals in the right ventromedial prefrontal cortex (Hampton et al. 2006). Thus prefrontal cortical neurons may be the source of the inferred value signals in the pallidus-habenula-dopamine pathway. The reverse direction of causality is also possible, because the prefrontal cortex receives a substantial input projection from dopamine neurons.

Second, we found that inferred stimulus values were represented in multiple brain regions in a manner suggesting a pallidus → habenula → dopamine route of transmission. GPi^{LHb}-negative neurons are likely to send their inferred value signals to the lateral habenula because they project to the lateral habenula and their reward signals occur at shorter latencies than in lateral habenula neurons (Hong and Hikosaka 2008). Likewise, lateral habenula neurons are likely to send their inferred value signals to dopamine neurons because electrical stimulation of the lateral habenula is known to inhibit dopamine neurons at short latencies (Christoph et al. 1986; Ji and Shepard 2007; Matsumoto and Hikosaka 2007). Thus our data suggest that stimulus values are likely to be inferred several steps upstream of dopamine neurons and transmitted to them through the pallidus-habenula-dopamine pathway.

Third, we found that some animals and neurons learned through a combination of experience and inference, not relying exclusively on either factor (Fig. 2). The relative influence of experience and inference is likely to depend on the animal's learning strategy and on the amount of skill the animal has gained at the task. During early training sessions, dopamine neuron activity reflects gradual learning and extinction of stimulus-reward associations through repeated pairings (Hollerman and Schultz 1998; Pan et al. 2008a; Roesch et al. 2007; Takikawa et al. 2004). In this study, however, animals had extensive prior knowledge of the task at hand, and neural activity was strongly influenced by task-specific inference. The transition from experiential to inferential learning could take place through at least two possible mechanisms. In the first mechanism, this transition could represent a change between different experience-based and inference-based task representations within a single brain structure. Regions such as the prefrontal cortex and dorsal striatum are thought to contain representations of multiple rules for task performance (or "task sets") (Miller and Cohen 2001; Sakai 2008; Watanabe and Hikosaka 2005), and could learn to select the most appropriate set of rules through repeated trial and error (Doya et al. 2002; Haruno et al. 2001). In the second mechanism, this transition could represent a change between experience-based and inference-based learning mechanisms located in separate brain structures. Extended training could cause a shift in the balance of control between these structures, similar to the shifts of control between neural systems during the learning of visuomotor sequences (Hikosaka et al. 1999) and spatial mazes (Packard and McGaugh 1996). These structures might correspond to the lateral prefrontal cortex and dorsomedial striatum versus the dorsolateral striatum, which have been proposed to have different roles in the early and late stages of reinforcement learning (Daw et al. 2005; Haruno and Kawato 2006; Yin and Knowlton 2006). In both of the mechanisms discussed above, information about experienced and inferred values could converge on the dorsal striatum, after which it could be sent directly to the pallidus-habenula-dopamine pathway through striato-pallidal projections. Both of these mechanisms also have the common feature that some prefrontal and striatal neurons would selectively learn experienced values, whereas other neurons would selectively learn inferred values. In this view, our data indicates that the pallidus-habenula-dopamine pathway can integrate value signals from both experience-based and inference-based learning systems, without being exclusively associated with either one.

Implications for computational models of reinforcement learning

The combination of experienced and inferred values also has an important implication for computational models of dopamine function. Most existing computational models learn stimulus values purely through repeated experience with stimulus-reinforcement pairings, similar to the experience-based learning system hypothesized above (Montague et al. 1996). However, these models are formally designed to assign value to abstract "states" of the environment (Sutton and Barto 1998), which do not have to correspond to individual sensory stimuli but can include abstract features of the environment that permit the use of specialized inference rules. For example, one computational model of dopamine function is able to perform inference by dividing the environment into distinct contexts where stimulus values are stable (e.g., block 1 and block 2) and learning stimulus values separately for each of these contexts (Nakahara et al. 2004). This procedure allows stimulus values to be updated immediately after a change between blocks, reproducing the pattern of perfect inference seen in *monkey D* (Fig. 4, *right column*); however, this model cannot reproduce the bias toward experienced value learning seen in *monkeys E* and *L* (Fig. 2). Thus to reproduce our data, a model is needed that can achieve both experiential and inferential learning, perhaps by using a mixture of the state representations that are appropriate for each learning strategy. This would resemble a class of algorithms that improve the reliability and flexibility of learning by averaging across multiple state representations (Daw et al. 2005; Doya et al. 2002; Haruno et al. 2001) or multiple learning rules (Camerer and Ho 1999).

The inference process seen in our data bears a resemblance to the inference performed by model-based reinforcement learning algorithms (Daw et al. 2005; Doya et al. 2002; Gläscher et al. 2010; Sutton and Barto 1998). These algorithms learn by observing transitions between states. If state A is known to transition to state B, and state B is paired with reward, the algorithm infers that the value of state A has

increased. Note, however, that standard model-based learning is different from the inference seen in our task, in two ways. First, model-based algorithms can infer stimulus values in novel situations, even before a stimulus has been directly paired with reward (e.g., after separately observing A → B and B → reward, they infer that A → reward). In contrast, our task required an inference rule to be learned through extensive training in a familiar task environment. Second, most model-based algorithms assume that the correct representation of task state is already known at the start of training (Daw et al. 2005; Doya et al. 2002; Sutton and Barto 1998). In contrast, the stimulus values in our task depended on hidden task states, block 1 and block 2, which were not known at the start of training and had to be discovered during the learning process. In summary, model-based inference requires subjects to learn new transitions between states, whereas inference in our task required subjects to learn a new state representation itself (Daw et al. 2006; Gluck and Myers 1993; Nakahara et al. 2004). In this view, our data indicate that the neural source of input to the pallidus-habenula-dopamine pathway can be trained to construct a faithful internal representation of the hidden structure of the environment. An important goal of future research will be to discover how the brain constructs these internal representations of its environment and how it uses them to drive the pallidus-habenula-dopamine pathway and reward-seeking behavior.

ACKNOWLEDGMENTS

We thank K. Doya, H. Nakahara, S. Kaveri, P. Dayan, K. Krueger, D. Lee, M. Yasuda, Y. Tachibana, S. Yamamoto, and H. Kim for valuable discussions.

GRANTS

This research was supported by the Intramural Research Program at the National Eye Institute.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

REFERENCES

Bayley PJ, Frascino JC, Squire LR. Robust habit learning in the absence of awareness and independent of the medial temporal lobe. *Nature* 436: 550–553, 2005.

Camerer C, Ho T-H. Experience-weighted attraction learning in normal form games *Econometrica* 67: 827–874, 1999.

Christoph GR, Leonzio RJ, Wilcox KS. Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *J Neurosci* 6: 613–619, 1986.

Daw ND, Courville AC, Touretzky DS. Representation and timing in theories of the dopamine system. *Neural Comput* 18: 1637–1677, 2006.

Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8: 1704–1711, 2005.

Dayan P, Niv Y. Reinforcement learning: the good, the bad and the ugly. *Curr Opin Neurobiol* 18: 185–196, 2008.

Doya K, Samejima K, Katagiri K-i, Kawato M. Multiple model-based reinforcement learning. *Neural Comput* 14: 1347–1369, 2002.

Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC, 1993.

Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66: 585–595, 2010.

Gluck MA, Myers CE. Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus* 3: 491–516, 1993.

Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision-making in humans. *J Neurosci* 26: 8360–8367, 2006.

Haruno M, Kawato M. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *J Neurophysiol* 95: 948–959, 2006.

Haruno M, Wolpert DM, Kawato M. MOSAIC model for sensorimotor learning and control. *Neural Comput* 13: 2201–2220, 2001.

Hikosaka O, Nakahara H, Rand MK, Sakai K, Lu X, Nakamura K, Miyachi S, Doya K. Parallel neural networks for learning sequential procedures. *Trends Neurosci* 22: 464–471, 1999.

Hollerman JR, Schultz W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci* 1: 304–309, 1998.

Hong S, Hikosaka O. The globus pallidus sends reward-related signals to the lateral habenula. *Neuron* 60: 720–729, 2008.

Ji H, Shepard PD. Lateral habenula stimulation inhibits rat midbrain dopamine neurons through a GABA(A) receptor-mediated mechanism. *J Neurosci* 27: 6923–6930, 2007.

Knowlton BJ, Mangels JA, Squire LR. A neostriatal habit learning system in humans. *Science* 273: 1399–1402, 1996.

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

Matsumoto M, Hikosaka O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* 459: 837–841, 2009.

Meyer DR. Food deprivation and discrimination reversal learning by monkeys. *J Exp Psychol* 41: 10–16, 1951.

Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24: 167–202, 2001.

Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16: 1936–1947, 1996.

Nakahara H, Itoh H, Kawagoe R, Takikawa Y, Hikosaka O. Dopamine neurons can represent context-dependent prediction error. *Neuron* 41: 269–280, 2004.

Packard MG, McGaugh JL. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol Learn Mem* 65: 65–72, 1996.

Pan WX, Schmidt R, Wickens JR, Hyland BI. Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. *J Neurosci* 28: 9619–9631, 2008a.

Pan X, Sawa K, Tsuda I, Tsukada M, Sakagami M. Reward prediction based on stimulus categorization in primate lateral prefrontal cortex. *Nat Neurosci* 11: 703–712, 2008b.

Parent A, Gravel S, Boucher R. The origin of forebrain afferents to the habenula in rat, cat and monkey. *Brain Res Bull* 6: 23–38, 1981.

Roesch MR, Calu DJ, Schoenbaum G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci* 10: 1615–1624, 2007.

Sakai K. Task set and prefrontal cortex. *Annu Rev Neurosci* 31: 219–245, 2008.

Satoh T, Nakai S, Sato T, Kimura M. Correlated coding of motivation and outcome of decision by dopamine neurons. *J Neurosci* 23: 9913–9923, 2003.

Sutton RS, Barto AG. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press, 1998.

Takikawa Y, Kawagoe R, Hikosaka O. A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *J Neurophysiol* 92: 2520–2529, 2004.

Watanabe K, Hikosaka O. Immediate changes in anticipatory activity of caudate neurons associated with reversal of position-reward contingency. *J Neurophysiol* 94: 1879–1887, 2005.

Wise RA. Dopamine, learning and motivation. *Nat Rev Neurosci* 5: 483–494, 2004.

Yin HH, Knowlton BJ. The role of the basal ganglia in habit formation. *Nat Rev Neurosci* 7: 464–476, 2006.