Behavioral/Systems/Cognitive

Basal Ganglia Neurons Dynamically Facilitate Exploration during Associative Learning

Sameer A. Sheth,[1] Tarek Abuelem,[2] John T. Gale,[1] and Emad N. Eskandar[1]

[1]Department of Neurosurgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, and [2]Department of Neurosurgery, Baylor College of Medicine, Houston, Texas 77030

The basal ganglia (BG) appear to play a prominent role in associative learning, the process of pairing external stimuli with rewarding responses. Accumulating evidence suggests that the contributions of various BG components may be described within a reinforcement learning model, in which a broad repertoire of possible responses to environmental stimuli are evaluated before the most profitable one is chosen. The striatum receives diverse cortical inputs, providing a rich source of contextual information about environmental cues. It also receives projections from midbrain dopaminergic neurons, whose phasic activity reflects a reward prediction error signal. These coincident information streams are well suited for evaluating responses and biasing future actions toward the most profitable response. Still lacking in this model is a mechanistic description of how initial response variability is generated. To investigate this question, we recorded the activity of single neurons in the globus pallidus internus (GPi), the primary BG output nucleus, in nonhuman primates (*Macaca mulatta*) performing a motor associative learning task. A subset (29%) of GPi neurons showed learning-related effects, decreasing firing during the early stages of learning, then returning to higher baseline rates as associations were mastered. On a trial-by-trial basis, lower firing rates predicted exploratory behavior, whereas higher rates predicted an exploitive response. These results suggest that, during associative learning, BG output is initially permissive, allowing exploration of a variety of responses. Once a profitable response is identified, increased GPi activity suppresses alternative responses, sharpening the response profile and encouraging exploitation of the profitable learned behavior.

Introduction

The basal ganglia (BG) are a central component of multiple parallel loops that, in the aggregate, engage virtually all parts of the cortex (Alexander and Crutcher, 1990; Parent and Hazrati, 1995; McFarland and Haber, 2000). Information follows known patterns of connections—cortex to striatum to pallidum to thalamus and back to cortex—establishing a loop within which the BG exert their influence. Evidence is mounting that the BG play an important role in associative motor learning (Graybiel, 2005). The connectivity described above reflects that of a reinforcement learning (RL) model, in which the results of a variety of actions are evaluated in a trial-and-error fashion. Those actions resulting in a favorable outcome are reinforced. Over time, this process biases behavior toward profitable actions and suppresses the occurrence of undesirable actions. Requirements for an RL model include mechanisms to generate variability in action, evaluate the results of those actions, and modify future behavior accordingly (Sutton and Barto, 1998).

The second of these mechanisms has been the subject of considerable previous work. Schultz and colleagues (2003) have convincingly demonstrated the role of dopaminergic neurons in the midbrain in encoding the requisite evaluative signal. These neurons fire in proportion to the difference between expected and actual reward, generating a reward prediction error signal. This error signal from midbrain dopaminergic neurons, along with environmental cues from cortex (Flaherty and Graybiel, 1993; Matsumoto et al., 2001; Haber et al., 2006), is coincident upon the striatum. A growing body of work demonstrates that striatal neurons modify their activity based on this convergent information (Kawagoe et al., 1998, 2004; Lauwereyns et al., 2002; Samejima et al., 2005). Dynamic modulation of striatal activity appears causally related to the behavioral changes observed in associative learning, providing evidence for the third mechanism (Brasted and Wise, 2004; Pasupathy and Miller, 2005; Williams and Eskandar, 2006).

Whereas a great deal of prior work has concentrated on elucidating the mechanisms of reward sensitivity and attendant behavioral modification, little is known regarding the first mechanism requisite for an RL model: generation of variability in action. The main output structures of the BG are the globus pallidus internus (GPi) and substantia nigra pars reticulata, which send GABAergic projections to the motor thalamus and superior colliculus (DeLong, 1971; Parent and Hazrati, 1995). A role for the GPi in motor learning has been implicated in a range of species. In songbirds, for example, the homologous structure is essential for vocal exploration and learning (Olveczky et al., 2005; Andalman and Fee, 2009; Gale and Perkel, 2010). In macaques, GPi neurons show enhancement of premovement modulation

Received July 12, 2010; revised Jan. 31, 2011; accepted Feb. 3, 2011.

This work was supported by National Institutes of Health Grants R25 NS065743 (to S.A.S.) and R01 EY017658 and R01 NS063249 (to E.N.E.), National Science Foundation Grant IOB 0645886 (to E.N.E.), and Howard Hughes Medical Institute Early Career Award (to E.N.E.).

Correspondence should be addressed to Emad N. Eskandar, Massachusetts General Hospital, 55 Fruit Street, White 502, Boston, MA 02114. E-mail: eeskandar@partners.org.

DOI:10.1523/JNEUROSCI.3658-10.2011

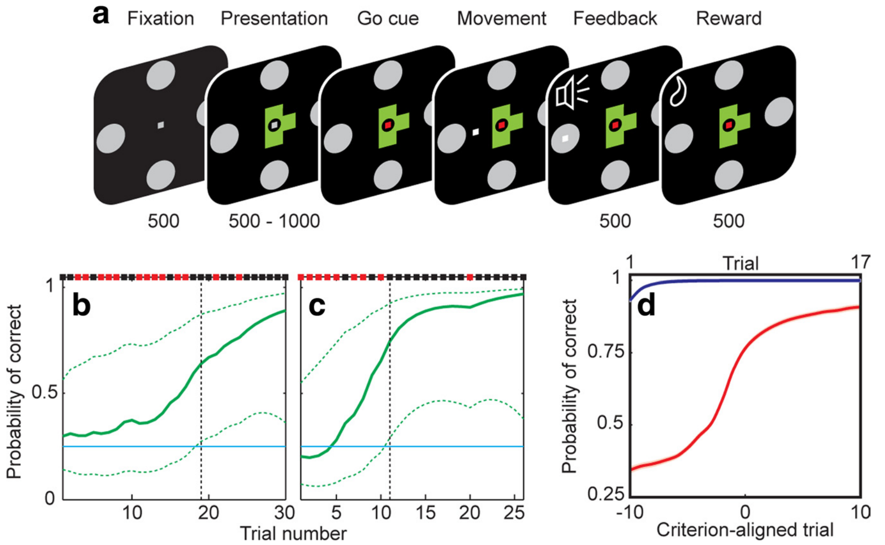

Figure 1. Behavioral task and performance. ***a***, The sequence of task epochs and their duration in milliseconds is shown for each representative screen. A fixation spot and the four targets appeared for 500 ms, followed by presentation of an object. After a variable delay, the go cue was indicated by a color change in the fixation spot, allowing joystick movement. Once the target was reached and held for 50 ms, a high or low tone indicated correct or incorrect response, respectively, and correct responses were rewarded with a drop of water. Fixation within a 1° window was required until target acquisition. ***b***, ***c***, Behavior during example learning blocks. Binary results (black, correct; red, incorrect) are shown along the top. The estimated learning curve is shown in green, with 99% confidence intervals indicated by the dashed lines. The criterion trial (vertical black dashed line) was defined as the point at which the lower 99% confidence interval surpassed chance (25%, horizontal blue line). These two learning blocks are the same as those depicted in Figure 5, *a* and *c*. ***d***, Population performance during familiar object (blue) and novel object (red) trials over learning, averaged across all learning blocks. Novel object trials are aligned to the criterion trial (lower *x*-axis labels) to allow for comparison across blocks regardless of learning rate. Familiar object trials are ordinally numbered (upper *x*-axis labels), as a criterion learning trial does not apply. Performance indicated mastery of familiar objects, and a gradually improving learning curve for novel objects. SEs are indicated by shading but are too small to be visible.

specific to novel but not familiar visual cues (Inase et al., 2001). Given these roles in learning and motor control, we hypothesized that the GPi may be a candidate structure responsible for generating the response variability characteristic of reinforcement learning. We therefore recorded the activity of GPi neurons in two monkeys while they performed a visuomotor associative-learning task.

Materials and Methods

Preparation and electrophysiology. Two adult male rhesus monkeys (*Macaca mulatta*) were studied in accordance with Massachusetts General Hospital and National Institutes of Health guidelines on animal research. The animals were fitted with a titanium head post, plastic recording chamber, and scleral search coils. The recording chamber was placed in a vertical orientation over the left hemisphere, centered 1 mm posterior to the posterior border of the anterior commissure and 9 mm lateral to midline. These surgical procedures were performed using standard sterile technique under isoflurane anesthesia, with postoperative administration of analgesics and antibiotics. During experiments, animals were comfortably seated in a primate chair with head fixation. A sipper tube was positioned at the mouth and fitted with side vents to prevent water availability with sucking. A joystick was oriented vertically and situated immediately in front of the animal's chair. The chair was designed so that the animal was forced to use the right hand (contralateral to the recording site). Eye position and joystick deflection were sampled and recorded at 1 kHz.

Acute recordings were performed daily with 0.5–1.0 MOhm tungsten microelectrodes (FHC). A single electrode per recording session was loaded in a millimeter-spaced grid, with at least 3 d intervening between loading the same grid location. Recording sites were confined to the GPi using confirmation between stereotactic coordinates, postimplant magnetic resonance imaging, and physiological signatures of deep nuclei and white matter boundaries. Animals were not sacrificed for histology following the recordings. Analog signals were bandpass filtered between 200 Hz and 5 kHz and sampled at 20 kHz (Spike2; Cambridge Electronic Design). Spikes were sorted offline using a principal components analysis-based template-matching algorithm (Spike2; Cambridge Electronic Design). We ensured that the waveforms and interspike intervals were consistent with single-unit activity.

Behavioral task. The animals were required to learn to associate a geometric object with a joystick movement in one of four possible directions (Fig. 1*a*). In each block of trials, the monkeys were presented with four objects, two of which were highly familiar (the association between the object and the correct joystick movement having been well established in previous training), and two of which were randomly generated novel shapes. The monkeys had to learn by trial-and-error, with water reward reinforcement, the correct direction associated with the novel objects. Within a block, each object was uniquely mapped to its correct direction without overlap, so that each direction was represented. The four objects were presented in semirandom sequence, such that all four objects were represented within a consecutive group of four trials, thereby ensuring an even temporal distribution of the objects over the block. Incorrect trials were repeated immediately until a correct response was achieved. Block switches would occur, without overt indication, after a minimum of 17 correct responses per object.

Each trial started with a fixation period, in which the animal had to acquire visual fixation of a central point around which four peripheral targets were arranged in the cardinal directions. After 500 ms, the object was presented centrally for a random interval between 500 and 1000 ms, followed by a go cue indicated by color change of the fixation point. Joystick movement was permitted after the go cue, and was required to follow a direct trajectory within a narrow corridor toward one of the peripheral targets. When the cursor reached the target and was held for 50 ms, a feedback tone sounded, indicating either a correct (high pitch) or incorrect (low pitch) choice. After 500 ms, water was delivered via the sipper tube on correct trials. Eye position was monitored and required to stay within 1° of the central fixation point through presentation, go cue, and movement epochs, until the feedback signal. The trial would abort if eye fixation broke, joystick movement began before the go cue, joystick trajectory deviated from a straight corridor to the target, or if the target was not held for a sufficient duration.

Animals also performed a control task in which the correct direction for each visual stimulus was indicated by a green (instead of gray) target. Control trials were similar to the regular task in all other aspects of appearance and timing. In a control block, all four objects were thus visually guided. Blocks of the control task were interleaved with regular task blocks, such that a control block guided the monkey through the same movements and reward schedule as the preceding regular block, but with cued targets.

Data analysis. We used a state-space smoothing algorithm for point processes (Smith et al., 2004) to estimate the learning curve and criterion learning point as the animals learned the correct associations (Fig. 1*b*,*c*). This algorithm uses a Bernoulli probability model to estimate the animal's continuous learning from his binary performance on each trial. The analysis is conducted from the perspective of an ideal observer, with complete knowledge of the performance during the block, rather than that of the subject, who only has information of previously completed trials. In the first step, the learning state process is estimated from the entire block's performance by fitting the binary (correct vs incorrect) responses to a model in which the unobservable value of the learning state in the current trial is defined to be a random step from the previous trial's value. Each step in this random walk is assumed to behave as an independent Gaussian variable, with a variance that determines how rapidly changes take place in the subject's trial-to-trial performance. This variance is estimated from the behavioral data using a maximum likelihood algorithm (Dempster et al., 1977). The fact that the entire block's performance (ideal observer's perspective) is used to estimate this parameter imparts a smooth progression to the calculated learning curve, which would otherwise appear jagged if only the performance up to the current trial (subject's perspective) were used.

Once the learning curve is estimated, the second step is calculation of the criterion learning point. To do so, 99% confidence intervals are determined around each point in the learning curve to account for the fact that the curve is an uncertain estimation. The criterion point is defined as the first trial where the lower 99% confidence bound surpasses chance (25% for four possible targets). At this point, the ideal observer would be 99% certain that the actual probability of a correct response was greater than chance. The criterion trial therefore represents the estimated point at which the animal first learned the correct association.

Only learning blocks reaching criterion were included in subsequent analyses. Because novel objects were learned at different rates, behavioral and neuronal data were aligned to the criterion trial across blocks (defined as trial zero) to evaluate changes in activity during comparable phases of learning (Brasted and Wise, 2004; Williams and Eskandar, 2006). Because learning did not occur for familiar objects, alignment to criterion was not applicable.

Firing rates were calculated within a 500 ms window centered on the task epochs (fixation, image presentation, go cue, movement, feedback sound, and reward). The peri-go cue directional preference for each cell was calculated by choosing the movement direction that was associated with the highest firing rate in the 500 ms window centered on the go cue.

To identify the subset of neurons whose activity was related to the task, we calculated the Pearson correlation coefficient (*r*) between the learning curve and firing rates. Only correct trials were included in this analysis. Population responses were calculated on normalized data to account for variation in firing rates across neurons. Normalization was performed for each neuron by subtracting the epoch-averaged minimum firing rate across all trials and dividing by the range, such that normalized rates ranged from 0 to 1. Changes in firing during learning were determined by comparing individual precriterion points to all postcriterion points (two-tailed *t* test) and by requiring at least two consecutive significant points.

To investigate the relationship between firing rates and the animals' behavior, we used a receiver operating characteristic (ROC) analysis. We tested the hypothesis that the firing rate near the go cue predicted the animal's choice on the subsequent movement. Two hypotheses were operationally defined to predict behavior after correct trials. For the exploration hypothesis, trials were sorted into two groups based on whether the chosen action in the current trial was the same or different from the chosen action for the previous trial with the same object. Choosing a different option would be consistent with exploration. For the

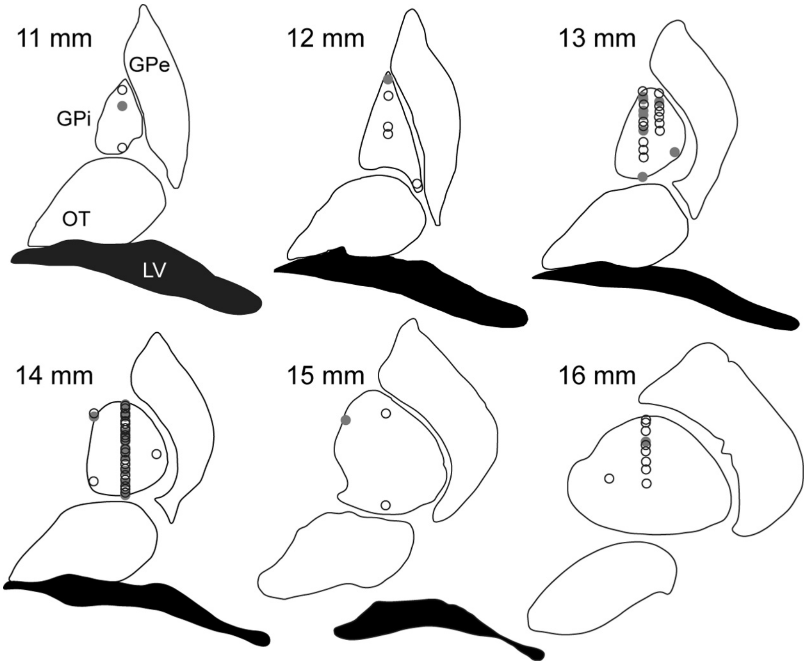

Figure 2. Location of GPi recording sites. Recording site location was determined by confirmation between stereotactic coordinates and physiological characteristics of deep nuclei and white matter boundaries. Coronal sections anterior to the interaural plane (noted in millimeters) from a standard atlas are diagramed with recording site locations. Sites with learning-related neurons are indicated with a filled circle, and learning-unrelated neurons with an open circle. GPe, globus pallidus externus; OT, optic tract; LV, lateral ventricle.

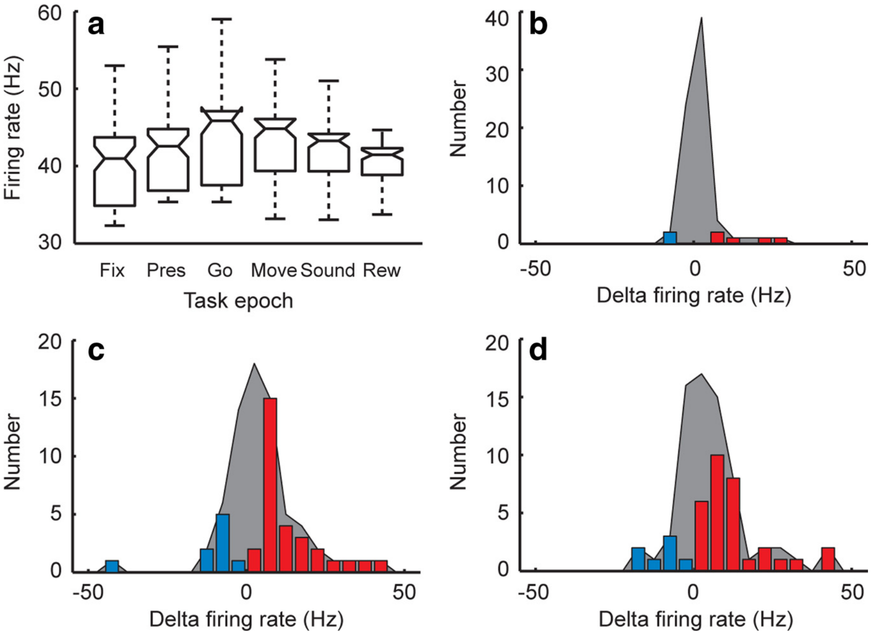

Figure 3. Population task responsiveness. ***a***, Box plot of population firing rates across the six task epochs: fixation (Fix), presentation (Pres), go cue (Go), movement (Move), feedback sound (Sound), and reward (Rew). The central line in each box represents the median, the box edges the 25% and 75% quartiles, and the whiskers 2.7 SDs (representing ~99.3% of the data). ***b–d***, Distribution of changes in firing rate (with respect to fixation) at the presentation, go cue, and movement epochs, respectively. The shaded region represents any change in firing rate relative to fixation, and the colored bars represent statistically significant increases (red) or decreases (blue).

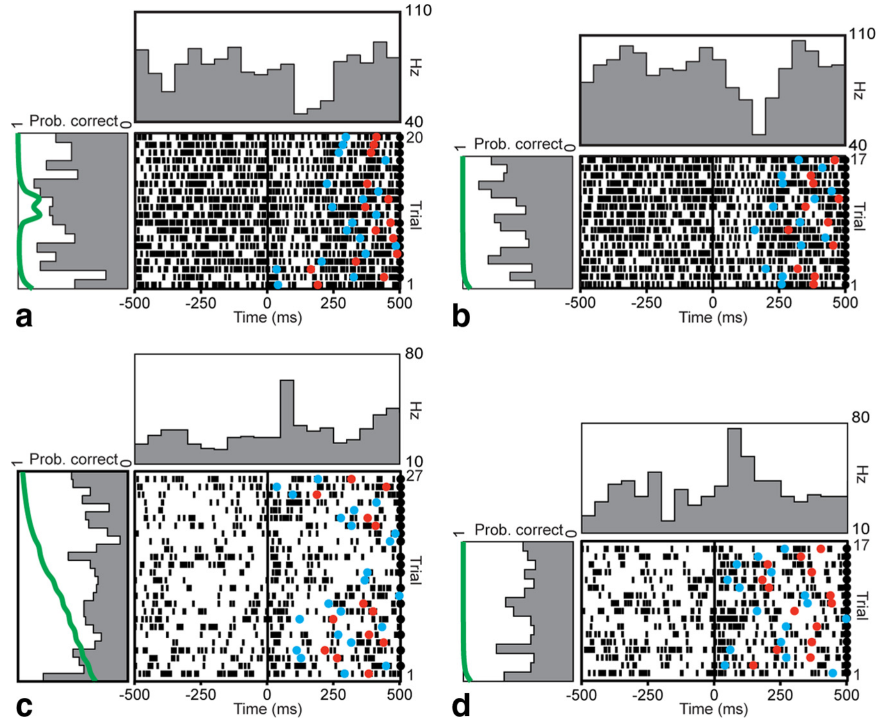

Figure 4. Example firing pattern of two learning-unrelated neurons. Rasters and peristimulus time histograms for sequential trials (bottom to top) aligned to the go cue ($t = 0$) are shown for two example neurons, over the course of a single learning block. ***a***, ***b***, This neuron decreased firing near the go cue when presented with both novel (***a***) and concurrently presented familiar (***b***) objects. Mean fixation firing rate during novel object trials was 60 Hz. ***c***, ***d***, This neuron increased firing near the go cue during both novel (***c***) and familiar (***d***) object trials. Mean fixation firing rate during novel object trials was 29 Hz. These changes are evident in the histograms in the panel above the raster. The histograms in the panels to the left of the rasters show changes in peri-go cue activity over the course of the learning block, averaged within a 500 ms window centered on the go cue. The green line in the left panel depicts the learning curve for that block (Prob. correct). Neither neuron demonstrated any learning-related modulation over the course of the block. Correct trials are indicated by black circles on the right edge of the raster. Reaction times and movement times in each trial are indicated by blue and red circles, respectively.

exploitation hypothesis, trials were sorted into two groups based on whether a correct or incorrect action was chosen. In this model, choosing the same option would be consistent with exploitation.

The behavioral response following incorrect trials cannot be easily categorized. If the choice following an incorrect response is identical, it can neither be confidently called exploitive (since it is not optimizing or exploiting reward attainment) nor exploratory (since it is not exploring other choice options). If the following choice is different, it may be either exploitive (since it may be trying to optimize choice) or exploratory (since it is different from previous). This analysis was therefore restricted to correct trials. An ROC curve was then constructed, and the area under the curve used as the discrimination value. Discrimination values were calculated within a window 400 ms wide, stepped in 100 ms increments, centered on the go cue. Significance was estimated with a bootstrap analysis by shuffling the neuronal–behavioral data pairs 1000 times and considering discrimination values ranking <5% or >95% as significant.

Results

Behavioral data

The animals completed an average of 6.3 ± 0.2 (mean ± SEM) learning blocks within each recording session. Behavioral responses for two separate learning blocks are depicted in Figure 1, *b* and *c*, and the overall population response in Figure 1*d*. The monkeys' behavior indicated mastery of the object–direction pairing for familiar objects. Performance on novel objects improved over the course of a learning block as the monkeys learned the correct associations. Because each object was mapped to a unique direction, presentation of a familiar object before a novel object at the start of a new block decreased the potential responses to the novel object. The animals' behavior indicated awareness of this feature of the task, as evidenced by the fact that the population learning curves began slightly >25% (Fig. 1*d*).

Between both animals, a total of 460 learning blocks were completed, encompassing 920 novel cue-movement associations. The animals successfully achieved criterion in 72% of blocks, on trial number 9.3 ± 0.8 (mean ± SEM, counting preceding incorrect and correct trials). Across all trials, reaction time (time between go cue and initiation of movement) was 276 ± 2 ms (mean ± SEM) for familiar objects and 352 ± 2 ms for novel objects. Movement time (time between initiation of movement and acquisition of target) was 111 ± 0.01 ms for familiar objects and 107 ± 0.02 ms for novel objects.

Neuronal data

We recorded 73 individual GPi neurons as the animals performed the associative-learning task. Recording sites are shown in Figure 2. We bisected each axis to define anterior versus posterior, medial versus lateral, and dorsal versus ventral subdivisions. Forty-three neurons (59%) were recorded in the skeletomotor region of the GPi, as defined by location in the posterior-lateral-ventral region of the GPi.

To determine task responsiveness of the neuronal population, we calculated average firing rates of each neuron within the six relevant task epochs (Fig. 3*a*). Firing rates differed significantly across task epochs ($p < 10^{-3}$, Kruskall–Wallis test). Individual comparisons of the epochs showed that firing rates first significantly differed from fixation at the go cue ($p < 0.01$, Tukey–Kramer *post hoc* test). As it is known that the composition of GPi neurons is heterogeneous, we divided the population based on whether an individual cell tended to increase or decrease its firing during the trial. The distribution of cells that significantly changed their firing rate during the presentation, go cue, and movement are indicated by the colored bars in Figure 3, *b*–*d*. Whereas a few individual cells increased ($N = 5$) or decreased ($N = 2$) firing at the presentation (Fig. 3*b*), a much larger number did so ($N = 30$ and 9, respectively) at the go cue (Fig. 3*c*), and persisted into the movement (Fig. 3*d*).

Because the first significant change in neuronal firing at the population level occurred at the go cue, we examined firing patterns of individual neurons during this epoch over the course of learning. For example, Figure 4 depicts peri-go cue rasters of a neuron's firing during trials in which one of the novel objects was being learned (Fig. 4*a*), and during trials in the same block in which one of the familiar objects was being presented (Fig. 4*b*). This neuron consistently decreased firing at the go cue on every trial, but this pattern did not modulate with learning over successive trials as the animal learned the correct association (Fig. 4*a*, left panel). Its pattern over the course of the block was relatively stationary and similar for both novel and familiar objects. Figure 4, *c* and *d*, depicts a neuron that increased its firing at the go cue during novel and familiar object trials, also in a pattern unrelated to learning.

In contrast, Figure 5 depicts two example neurons whose firing changed over the course of learning a novel object. In the first example (Fig. 5*a*), the peri-go cue firing rate was relatively higher at the beginning and end of the learning block, but was lower in the middle for several trials. In the second example (Fig. 5*c*), the decrease in firing occurred earlier in the learning block. Examination of the animal's behavioral performance in these blocks revealed a similar difference. In the first case, criterion performance was achieved after 19 trials (counting previous correct and incorrect trials). In the second, criterion was achieved earlier, after 11 trials. Firing in these cells correlated significantly with behavioral performance, as measured by the probability of a correct response ($p < 0.05$, Pearson's linear correlation). This modulation in firing over the course of a learning block did not take place during concurrent familiar object trials (Fig. 5*b*,*d*), indicating an effect specific to novel object learning.

In other neurons whose firing transiently decreased for several trials over the course of learning, we observed a similar relationship between the timing of the change and the rate at which learning criterion was reached. To better study these dynamic learning-related changes, we aligned trials to the trial at which criterion learning performance was attained. Alignment to the criterion trial allowed us to compare activity across similar stages of learning, compensating for different learning rates across blocks. Only correct trials were included in this analysis. We sought to identify the subset of neurons most involved in learning by choosing those whose activity correlated with the learning curve (Williams and Eskandar, 2006). Twenty-one of the 73 neurons (29%) showed a significant positive correlation ($p < 0.05$, Pearson's test) with the learning curve, including the described transient decrease in firing lasting several trials, and were included in subsequent analyses. The two neurons depicted in Figure 5 are representative examples of these learning-related neurons.

To determine whether location was a significant factor in identifying learning-related neurons, we compared the fraction of learning neurons identified in the posterior versus anterior region, medial versus lateral region, and dorsal versus ventral region. None of these comparisons were significant ($p > 0.05$, Fisher's exact test). Four of the 73 neurons showed a significant negative correlation ($p < 0.05$, Pearson's test) with the learning curve.

The population peri-go cue activity in this subset of learning-related neurons is shown in Figure 6*a*. There was no modulation in firing rates across trials when the animals were presented with familiar objects. During novel association learning, however, we observed a consistent transient decrease in peri-go cue firing rate starting eight trials before criterion was reached. This suppression was significant compared with postcriterion points ($p < 0.05$, two-tailed *t* test) and lasted for four trials. By the time learning criterion was reached, activity had returned to the higher baseline firing rate, and was indistinguishable from the activity observed in familiar trials. This difference between novel and familiar object trials suggests that the effect seen during novel object trials was specifically related to the learning required during those trials, and absent during contemporaneous familiar object trials, in which there was no active learning. This effect was not present when rates were aligned to the fixation or presentation.

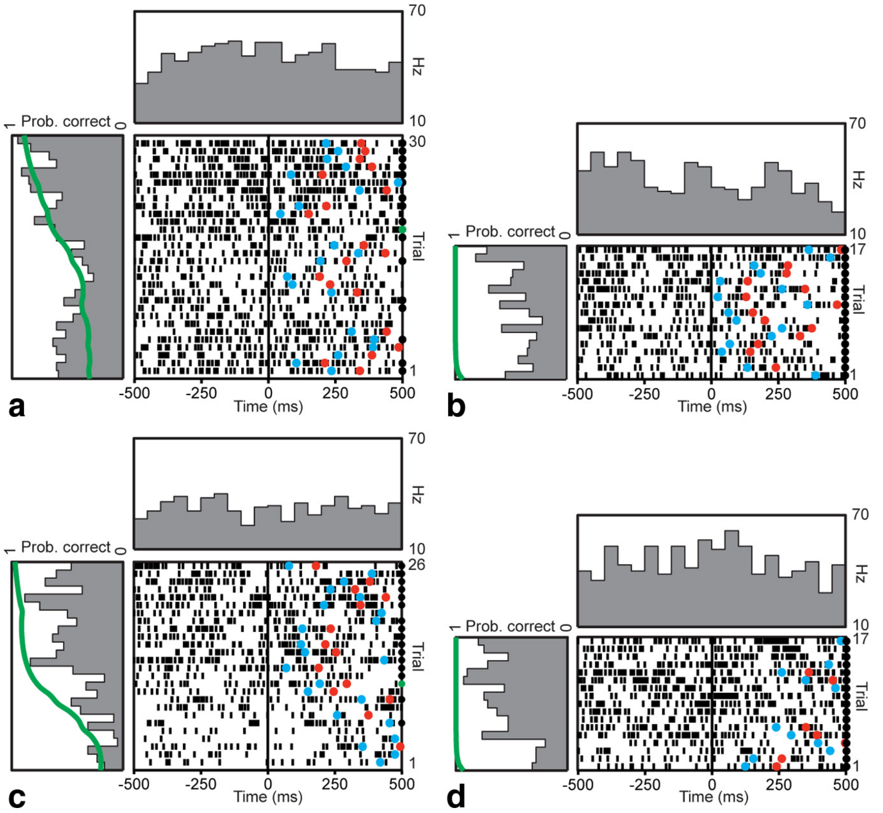

Figure 5. Example firing patterns of two learning-related neurons. Rasters and peristimulus time histograms aligned to the go cue are shown for two neurons. ***a***, ***b***, Example firing pattern of a learning-related neuron during presentation of a novel object (***a***) and a concurrently presented familiar object (***b***). Correct trials are indicated by a black circle on the right edge of the raster. The trial at which learning criterion was achieved is indicated with a green circle. This neuron decreased its firing during novel object trials near the middle of the learning block, but exhibited no such pattern during familiar object trials. Learning criterion occurred at trial number 19. Mean fixation firing rate during novel object trials was 20 Hz. ***c***, ***d***, Example firing pattern of a second learning-related neuron during novel (***c***) and familiar (***d***) object trials. This neuron decreased firing early in the block during novel object trials. Learning criterion occurred at trial number 11. Mean fixation firing rate was 34 Hz. Histograms were calculated as in Figure 4. Behavioral curves for ***a*** and ***c*** are the same as those depicted with explanation in Figure 1, *b* and *c*.

The rate of learning did not differ between sessions during which learning-related and learning-unrelated neurons were recorded. The number of trials to criterion was 9.2 ± 0.3 in the former and 9.4 ± 0.3 in the latter ($p = 0.71$, two-tailed *t* test), and the shape of the learning curves was also identical (Fig. 6*d*). The population activity of all 52 learning-unrelated neurons did not exhibit a similar decrease in GPi firing in the peri-go cue period (Fig. 6*e*).

To ensure that this change in neuronal firing was not simply a reflection of variations in movement parameters over the course of learning, we plotted reaction times and movement times as a function of criterion-aligned trial number (Fig. 6*b*,*c*). There was no significant difference between precriterion and postcriterion reaction times ($p = 0.15$, *t* test), nor between precriterion and postcriterion movement times ($p = 0.63$). We also confirmed

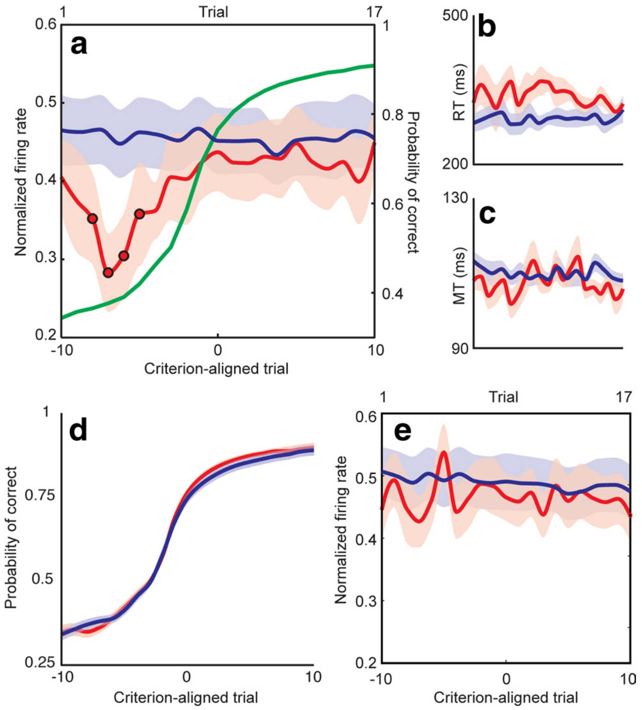

Figure 6. GPi firing encodes a facilitation window. ***a***, Normalized firing rate for novel (red) object trials as a function of criterion-aligned trial number for the subset of learning-related neurons. Familiar object trials (blue) are shown as a function of ordinal trial number (top *x*-axis), as alignment to criterion was not applicable. The learning curve (right *y*-axis) is shown in green. There was a significant ($p < 0.05$, circles) decrease in firing rate during novel object trials early in learning. Because pallidothalamic projections are inhibitory, this decrease encodes a facilitation window that can promote particular downstream motor programs. Reaction times (RT; ***b***) and movement times (MT; ***c***) for novel and familiar object trials showed no difference between precriterion and postcriterion values. *x*-Axis values for ***b*** and ***c*** are identical to ***a***. ***d***, Learning curves were identical between sessions during which learning-related neurons (red) and learning-unrelated neurons (blue) were recorded. ***e***, Population activity of the 52 learning-unrelated neurons did not exhibit a similar decrease in GPi firing. As in ***a***, novel object trials are depicted in red, and familiar object trials in blue. Shading indicates SEM.

that the direction of the impending movement did not influence the exploration-related changes by first calculating peri-go cue directional preferences for each cell. We then examined the changes in neuronal firing over learning after separating each trial based on whether the impending movement was in or against the cell's preferred direction. At no point was there a significant difference between these curves ($p > 0.05$), suggesting that the direction of movement did not influence the exploration-related changes.

To determine whether this decrease in GPi firing was specifically related to a particular phase of learning, we aligned the trials to the first presentation of the novel object, rather than to the criterion trial. If this effect was simply a function of stimulus novelty, ordinal trial alignment would make it even more prominent. Aligning to the first presentation rather than criterion trial, however, made the effect disappear (Fig. 7*a*). The presence of the decrease in firing depended upon correction for the pace of learning, suggesting that its timing was related to a particular early process in learning (occurring five to eight trials before criterion), rather than other nonspecific aspects of the task occurring soon after a block change.

Early in the learning block, the fraction of correct trials was relatively low. It is therefore possible that the observed decrease in firing rate was simply a reflection of the sparser reward frequency early in the block. To rule out that possibility, we performed a control task in 17 of the 73 total neurons, similar in design to the main learning task, except that responses in each trial were indicated by a change in the color of the target, such that no learning was required (see Materials and Methods, above). If the effect were simply a function of the changing amount of reward encountered over the course of a block, it would be identical between the learning and the adjacent control blocks. Removing the requirement to dynamically learn associations, however, eliminated the transient decrease in firing rate (Fig. 7*b*). This period of decreased GPi activity is therefore not simply a general appetitive effect of reward schedule or reinforcement.

Relationship between neuronal and behavioral data

To understand the functional significance of the observed brief decrease in GPi activity, we sought to identify an explicit relationship between peri-go cue GPi firing and behavioral choice. We constructed ROCs to evaluate the statistical relationships between changes in GPi firing and the animal's choice of action. The ROC analysis approximates the likelihood that an ideal observer would be able to predict the behavioral outcome in an individual trial from the neuronal activity (Britten et al., 1996). We tested two alternative hypotheses: (1) the firing rate in an individual trial predicts a choice different from the previous choice (exploration model), and (2) the firing rate in an individual trial predicts an impending correct choice (exploitation model). The exploration model describes a behavioral paradigm in which the animal is actively exploring the parameter space, intentionally choosing a response different from the last, despite the fact that the previous answer may have been correct. The exploitation model describes a strategy optimized to consistently choose the response most recently identified as correct, thereby maximizing the chance of obtaining reward.

Discrimination values (area under the ROC curve) for both models were calculated in sliding increments relative to the go cue for the 21 learning-related neurons. Example discrimination values for two neurons are shown in Figure 8*a*,*b* for both exploratory (blue) and exploitive (red) hypotheses. Significance was estimated using a bootstrap analysis (thick lines). Discrimination values tended to decrease significantly below chance (0.5) during exploration trials, and increase above chance during exploitation trials.

Population ROC discrimination values for the learning-related neurons are shown in Figure 8*c*. Before the go cue, discrimination values for both models remained near chance and overlapped in distribution. Nearly contemporaneous with the go cue (50 ms prior), however, the models began to diverge significantly ($p < 0.05$, *t* test) from each other and chance. The lower values for the exploration model indicate that lower firing rates predicted exploratory behavior, and the higher values for the exploitation model indicate that higher firing rates predicted exploitive behavior.

To relate the ROC analysis results back to the learning process, ROC discrimination values were calculated as a function of trial number. This analysis was performed in a time window starting at the point at which the two ROC curves diverged significantly (50 ms before go cue) and ending at the mean reaction time for novel objects (350 ms after go cue), thereby including only peri-go cue firing. The average firing rate within this window was paired with the animal's choice on that trial and submitted to the same two-hypothesis population ROC to generate discrimination values as a function of learning. Significance of the ROC discrimination values was again determined by a bootstrap analysis. Thus, at every point in learning, we arrived at a discrimina-

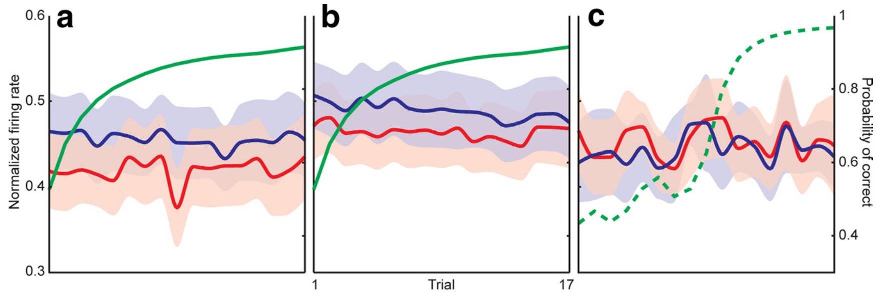

Figure 7. The facilitation window is not an effect of stimulus novelty or reward schedule. ***a***, The same firing rates of learning-related neurons for novel (red) and familiar (blue) object trials as depicted in Figure 6*a*, aligned to ordinal trial number, starting at the first correct response, rather than to criterion. The learning curve is shown in green. Alignment to ordinal trial number removed the decrease in GPi firing, suggesting that this effect was specific to learning, rather than simply to stimulus novelty. ***b***, Alignment to first correct response of the learning-unrelated neurons (depicted in Fig. 6*e*), again demonstrating no facilitation window. ***c***, GPi firing during a control task in which responses were guided by a color change in the target, such that the movements and reward schedules matched those of interleaved blocks of the normal learning task (see Materials and Methods). The dashed green line indicates the learning curve from the neighboring block of the regular task, as the control block itself had guided cues precluding learning. Removal of the necessity to actively learn the associations again eliminated the decrease in firing. Shaded regions indicate SEM. All axes share the same labels and ranges.

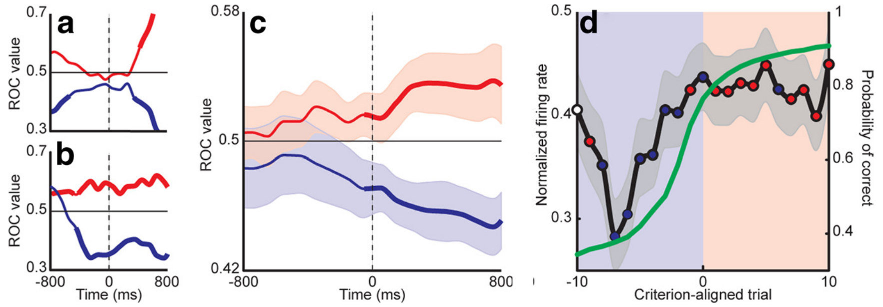

Figure 8. GPi firing predicts a behavioral shift from exploration to exploitation. ROC discrimination values were calculated for exploration (blue) and exploitation (red) hypotheses in a sliding window 400 ms wide stepped in 100 ms increments, centered on the go cue. Discrimination values for two example neurons are shown in ***a*** and ***b***. Thick lines represent significant differences from chance (0.5). Exploration tended to be associated with lower firing rates and exploitation with higher firing rates. ***c***, Population ROC discrimination values for the subpopulation of learning-related neurons. The values for the exploration (blue) and exploitation (red) hypotheses diverged near the time of the go cue. Significant differences between the two are denoted by a thick line. Shaded regions indicate SEM. Lower values for the exploration model indicate that lower firing rates predicted exploratory behavior, whereas higher values for the exploitation model indicate that higher firing rates predicted exploitive behavior. ***d***, To relate these ROC findings to the learning process, discrimination values were calculated as a function of trial number and compared with the actual average firing rate (Fig. 6*a*). For each trial, blue circles indicate that the firing rate predicted exploratory behavior and red circles indicate that the firing rate predicted exploitive behavior. White circles indicate trials in which the ROCs were not significantly different from chance. Starting eight trials before criterion, precriterion trials were characterized by exploratory behavior (blue shaded region). This pattern shifted around the time of criterion, such that the majority of postcriterion trials demonstrated exploitive behavior (red shaded region).

tion value for both exploration and exploitation hypotheses. By comparing the average firing rates for the learning-related neurons (Fig. 6*a*) to the discrimination values, we could thereby determine which of the two behavioral strategies is favored at various stages of learning. Because both the normalized firing rates and ROC values were constrained between 0 and 1, a low firing rate (as occurred five to eight trials before criterion) would predict an exploratory behavior on that trial if the peri-go cue discrimination value for exploration was significantly low, whereas that for exploitation was significantly high. This relationship was quantified by taking the absolute magnitude of the difference between significant discrimination values and normalized firing rates at each trial, and assigning that trial's behavioral preference to the behavior with the smaller difference.

These results are displayed in Figure 8*d*. Ten trials before criterion, neither ROC discrimination value was significantly different from chance (white circle). On the next trial, the firing rate predicted exploitive choices (red circle), possibly due to carry-over effects from the previous learning block. For the next seven trials, during the prominent decrease in GPi firing, the firing rate predicted exploratory choices (blue circles). In the vicinity of the criterion trial, the predictions alternated briefly, before settling on predictions of exploitive behavior for all but one of the postcriterion trials.

Discussion

In this study, we investigated the role of the GPi in associative motor learning in two nonhuman primates trained to pair a novel visual cue with a particular joystick movement. Specifically, we tested the hypothesis that the GPi is involved in generating the variability in action required in a reinforcement learning model. Our results demonstrate that GPi firing decreases transiently early in the learning process, before behavior indicates mastery of the association; on a trial-by-trial basis, lower peri-go cue firing predicts exploratory behavior, whereas higher firing predicts exploitive behavior; and GPi firing predicts the transition from an exploratory to exploitive behavior.

The GPi exerts its influence on motor control via its interconnections with the pedunculopontine complex and thalamus. It sends GABAergic inhibitory projections to the ventral anterior and ventral lateral nuclei of the thalamus, which in turn project to primary and supplementary motor cortex (Inase and Tanji, 1995). Thus, a decrease in GPi activity would release downstream thalamocortical circuits from inhibitory tone. In this context, the transient decrease in peri-go cue GPi firing that we observed would serve to facilitate various different motor programs. The timing of the firing decrease with regard to the learning process, together with the results of the ROC analysis, suggest that GPi firing encodes a window of facilitation that encourages exploratory behavior early in the learning process, before the maximally profitable response is identified. By the time the stimulus-response pairing is learned and behavior accordingly optimized, GPi firing increases, closing the facilitation window for exploration, and encouraging exploitation of the identified profitable behavior. These shifts between exploratory and exploitive behavioral phenotypes (Fig. 8*d*, shaded regions) therefore correlate with changes in GPi firing.

Exploring the possible parameter space of responses to a novel situation presents each response to evaluation, allowing profitable responses to be promoted and undesirable responses to be suppressed. Our data support the hypothesis that the early facilitation window in GPi firing is the mechanism by which this variability in action requisite for reinforcement learning is generated. Appropriately, however, this variability is present only until the association is learned, lest maladaptive persistence of this exploratory mechanism prevent the eventually necessary narrowing of the behavioral repertoire to the optimal choice. Theoretical models of BG function in the context of reinforcement learning have discussed the necessity of this variability in action, but lack an experimentally demonstrated neuronal substrate (Sridharan et al., 2006; Ponzi, 2008). In contrast, the other requirements of the model, such as a source for reward prediction error and a reward-contingent mechanism to modify future behavior, are consistently attributed to midbrain dopaminergic neurons and medium spiny striatal neurons, respectively.

Evidence across a range of species is accumulating for the role of the BG in facilitating early exploratory behavior. Rats learning to navigate a maze for food reward initially explore their environment, but then settle into a habitual response, which reemerges rapidly after the habit has been actively extinguished. The firing pattern of striatal projection neurons correlates with these back-and-forth shifts between exploratory and exploitive behavior (Barnes et al., 2005). As another example, the song of the young zebra finch is initially highly variable, but conforms over time to a stereotyped pattern via feedback from a tutor's song. The lateral magnocellular nucleus of the anterior nidopallium, the output nucleus of the songbird basal ganglia analog (Luo et al., 2001), is required for the vocal exploration characteristic of juvenile bird song, but not for production of stereotyped adult song (Kao et al., 2005; Olveczky et al., 2005; Kao and Brainard, 2006; Andalman and Fee, 2009). Our results further clarify the mechanism by which the primate basal ganglia promote early exploration in associative learning. GPi firing dynamically shifts between a facilitatory and inhibitory state, and this shift provides a physiological basis to explain the behavioral transition from exploration of a broad repertoire of responses to exploitation of the eventually identified maximally profitable response.

References

Alexander GE, Crutcher MD (1990) Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci 13:266 –271.

Andalman AS, Fee MS (2009) A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors. Proc Natl Acad Sci U S A 106:12518 –12523.

Barnes TD, Kubota Y, Hu D, Jin DZ, Graybiel AM (2005) Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. Nature 437:1158 –1161.

Brasted PJ, Wise SP (2004) Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum. Eur J Neurosci 19:721–740.

Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA (1996) A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis Neurosci 13:87–100.

DeLong MR (1971) Activity of pallidal neurons during movement. J Neurophysiol 34:414 – 427.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc Ser B 39:1–38.

Flaherty AW, Graybiel AM (1993) Two input systems for body representations in the primate striatal matrix: experimental evidence in the squirrel monkey. J Neurosci 13:1120 –1137.

Gale SD, Perkel DJ (2010) A basal ganglia pathway drives selective auditory responses in songbird dopaminergic neurons via disinhibition. J Neurosci 30:1027–1037.

Graybiel AM (2005) The basal ganglia: learning new tricks and loving it. Curr Opin Neurobiol 15:638 – 644.

Haber SN, Kim KS, Mailly P, Calzavara R (2006) Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. J Neurosci 26:8368 – 8376.

Inase M, Tanji J (1995) Thalamic distribution of projection neurons to the primary motor cortex relative to afferent terminal fields from the globus pallidus in the macaque monkey. J Comp Neurol 353:415– 426.

Inase M, Li BM, Takashima I, Iijima T (2001) Pallidal activity is involved in visuomotor association learning in monkeys. Eur J Neurosci 14:897–901.

Kao MH, Brainard MS (2006) Lesions of an avian basal ganglia circuit prevent context-dependent changes to song variability. J Neurophysiol 96:1441–1455.

Kao MH, Doupe AJ, Brainard MS (2005) Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. Nature 433:638 – 643.

Kawagoe R, Takikawa Y, Hikosaka O (1998) Expectation of reward modulates cognitive signals in the basal ganglia. Nat Neurosci 1:411– 416.

Kawagoe R, Takikawa Y, Hikosaka O (2004) Reward-predicting activity of dopamine and caudate neurons: a possible mechanism of motivational control of saccadic eye movement. J Neurophysiol 91:1013–1024.

Lauwereyns J, Watanabe K, Coe B, Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. Nature 418:413– 417.

Luo M, Ding L, Perkel DJ (2001) An avian basal ganglia pathway essential for vocal learning forms a closed topographic loop. J Neurosci 21:6836 – 6845.

Matsumoto N, Minamimoto T, Graybiel AM, Kimura M (2001) Neurons in the thalamic CM-Pf complex supply striatal neurons with information about behaviorally significant sensory events. J Neurophysiol 85:960 –976.

McFarland NR, Haber SN (2000) Convergent inputs from thalamic motor nuclei and frontal cortical areas to the dorsal striatum in the primate. J Neurosci 20:3798 –3813.

Olveczky BP, Andalman AS, Fee MS (2005) Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. PLoS Biol 3:e153.

Parent A, Hazrati LN (1995) Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. Brain Res Brain Res Rev 20:91–127.

Pasupathy A, Miller EK (2005) Different time courses of learning-related activity in the prefrontal cortex and striatum. Nature 433:873– 876.

Ponzi A (2008) Dynamical model of salience gated working memory, action selection and reinforcement based on basal ganglia and dopamine feedback. Neural Netw 21:322–330.

Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310:1337–1340.

Schultz W, Tremblay L, Hollerman JR (2003) Changes in behavior-related neuronal activity in the striatum during learning. Trends Neurosci 26:321–328.

Smith AC, Frank LM, Wirth S, Yanike M, Hu D, Kubota Y, Graybiel AM, Suzuki WA, Brown EN (2004) Dynamic analysis of learning in behavioral experiments. J Neurosci 24:447– 461.

Sridharan D, Prashanth PS, Chakravarthy VS (2006) The role of the basal ganglia in exploration in a neural model based on reinforcement learning. Int J Neural Syst 16:111–124.

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, Massachusetts: MIT.

Williams ZM, Eskandar EN (2006) Selective enhancement of associative learning by microstimulation of the anterior caudate. Nat Neurosci 9:562–568.